Lecture Notes on Statistics and Information Theory

John Duchi

December 6, 2023

Contents

1 Introduction and setting **8**
- 1.1 Information theory . . . 8
- 1.2 Moving to statistics . . . 9
- 1.3 A remark about measure theory . . . 10
- 1.4 Outline and chapter discussion . . . 10

2 An information theory review **12**
- 2.1 Basics of Information Theory . . . 12
  - 2.1.1 Definitions . . . 12
  - 2.1.2 Chain rules and related properties . . . 17
  - 2.1.3 Data processing inequalities: . . . 19
- 2.2 General divergence measures and definitions . . . 20
  - 2.2.1 Partitions, algebras, and quantizers . . . 20
  - 2.2.2 KL-divergence . . . 21
  - 2.2.3 f-divergences . . . 22
  - 2.2.4 Inequalities and relationships between divergences . . . 25
  - 2.2.5 Convexity and data processing for divergence measures . . . 29
- 2.3 First steps into optimal procedures: testing inequalities . . . 30
  - 2.3.1 Le Cam's inequality and binary hypothesis testing . . . 30
  - 2.3.2 Fano's inequality and multiple hypothesis testing . . . 31
- 2.4 A first operational result: entropy and source coding . . . 33
  - 2.4.1 The source coding problem . . . 34
  - 2.4.2 The Kraft-McMillan inequalities . . . 35
  - 2.4.3 Entropy rates and longer codes . . . 37
- 2.5 Bibliography . . . 39
- 2.6 Exercises . . . 39

3 Exponential families and statistical modeling **45**
- 3.1 Exponential family models . . . 45
- 3.2 Why exponential families? . . . 47
  - 3.2.1 Fitting an exponential family model . . . 50
- 3.3 Divergence measures and information for exponential families . . . 51
- 3.4 Generalized linear models and regression . . . 52
  - 3.4.1 Fitting a generalized linear model from a sample . . . 55
- 3.5 Lower bounds on testing a parameter's value . . . 56
- 3.6 Deferred proofs . . . 58

3.6.1 Proof of Proposition 3.2.2 . . . 58
3.7 Bibliography . . . 60
3.8 Exercises . . . 60

I Concentration, information, stability, and generalization 61

4 Concentration Inequalities 62
4.1 Basic tail inequalities . . . 62
4.1.1 Sub-Gaussian random variables . . . 64
4.1.2 Sub-exponential random variables . . . 68
4.1.3 Orlicz norms . . . 72
4.1.4 First applications of concentration: random projections . . . 73
4.1.5 A second application of concentration: codebook generation . . . 75
4.2 Martingale methods . . . 76
4.2.1 Sub-Gaussian martingales and Azuma-Hoeffding inequalities . . . 78
4.2.2 Examples and bounded differences . . . 79
4.3 Uniformity and metric entropy . . . 81
4.3.1 Symmetrization and uniform laws . . . 82
4.3.2 Metric entropy, coverings, and packings . . . 86
4.4 Generalization bounds . . . 90
4.4.1 Finite and countable classes of functions . . . 91
4.4.2 Large classes . . . 93
4.4.3 Structural risk minimization and adaptivity . . . 95
4.5 Technical proofs . . . 97
4.5.1 Proof of Theorem 4.1.11 . . . 97
4.5.2 Proof of Theorem 4.1.15 . . . 98
4.5.3 Proof of Theorem 4.3.6 . . . 99
4.6 Bibliography . . . 99
4.7 Exercises . . . 99

5 Generalization and stability 104
5.1 The variational representation of Kullback-Leibler divergence . . . 105
5.2 PAC-Bayes bounds . . . 106
5.2.1 Relative bounds . . . 108
5.2.2 A large-margin guarantee . . . 111
5.2.3 A mutual information bound . . . 113
5.3 Interactive data analysis . . . 114
5.3.1 The interactive setting . . . 115
5.3.2 Second moment errors and mutual information . . . 116
5.3.3 Limiting interaction in interactive analyses . . . 117
5.3.4 Error bounds for a simple noise addition scheme . . . 122
5.4 Bibliography and further reading . . . 123
5.5 Exercises . . . 124

6 Advanced techniques in concentration inequalities 128
6.1 Entropy and concentration inequalities . . . 128
6.1.1 The Herbst argument . . . 129
6.1.2 Tensorizing the entropy . . . 130
6.1.3 Concentration of convex functions . . . 134

7 Privacy and disclosure limitation 138
7.1 Disclosure limitation, privacy, and definitions . . . 138
7.1.1 Basic mechanisms . . . 140
7.1.2 Resilience to side information, Bayesian perspectives, and data processing . . . 144
7.2 Weakenings of differential privacy . . . 146
7.2.1 Basic mechanisms . . . 147
7.2.2 Connections between privacy measures . . . 149
7.2.3 Side information protections under weakened notions of privacy . . . 152
7.3 Composition and privacy based on divergence . . . 155
7.3.1 Composition of Rényi-private channels . . . 155
7.3.2 Privacy games and composition . . . 156
7.4 Additional mechanisms and privacy-preserving algorithms . . . 158
7.4.1 The exponential mechanism . . . 158
7.4.2 Local sensitivities and the inverse sensitivity mechanism . . . 161
7.5 Deferred proofs . . . 166
7.5.1 Proof of Lemma 7.2.10 . . . 166
7.6 Bibliography . . . 169
7.7 Exercises . . . 169

II Fundamental limits and optimality 176

8 Minimax lower bounds: the Le Cam, Fano, and Assouad methods 178
8.1 Basic framework and minimax risk . . . 178
8.2 Preliminaries on methods for lower bounds . . . 180
8.2.1 From estimation to testing . . . 181
8.2.2 Inequalities between divergences and product distributions . . . 182
8.2.3 Metric entropy and packing numbers . . . 184
8.3 Le Cam's method . . . 185
8.4 Fano's method . . . 187
8.4.1 The classical (local) Fano method . . . 187
8.4.2 A distance-based Fano method . . . 192
8.5 Assouad's method . . . 195
8.5.1 Well-separated problems . . . 195
8.5.2 From estimation to multiple binary tests . . . 195
8.5.3 Example applications of Assouad's method . . . 197
8.6 Nonparametric regression: minimax upper and lower bounds . . . 199
8.6.1 Kernel estimates of the function . . . 199
8.6.2 Minimax lower bounds on estimation with Assouad's method . . . 203
8.7 Global Fano Method . . . 206
8.7.1 A mutual information bound based on metric entropy . . . 206

8.7.2 Minimax bounds using global packings . . . 208
8.7.3 Example: non-parametric regression . . . 208
8.8 Deferred proofs . . . 210
8.8.1 Proof of Proposition 8.4.6 . . . 210
8.8.2 Proof of Corollary 8.4.7 . . . 210
8.8.3 Proof of Lemma 8.5.2 . . . 211
8.9 Bibliography . . . 211
8.10 Exercises . . . 211

9 Constrained risk inequalities **220**
9.1 Strong data processing inequalities . . . 220
9.2 Local privacy . . . 223
9.3 Communication complexity . . . 227
9.3.1 Classical communication complexity problems . . . 227
9.3.2 Deterministic communication: lower bounds and structure . . . 230
9.3.3 Randomization, information complexity, and direct sums . . . 232
9.3.4 The structure of randomized communication and communication complexity of primitives . . . 236
9.4 Communication complexity in estimation . . . 239
9.4.1 Direct sum communication bounds . . . 240
9.4.2 Communication data processing . . . 241
9.4.3 Applications: communication and privacy lower bounds . . . 243
9.5 Proof of Theorem 9.4.4 . . . 247
9.5.1 Proof of Lemma 9.5.3 . . . 251
9.6 Bibliography . . . 252
9.7 Exercises . . . 252

10 Testing and functional estimation **257**
10.1 Le Cam's convex hull method . . . 257
10.1.1 The χ^2-mixture bound . . . 259
10.1.2 Estimating errors and the norm of a Gaussian vector . . . 261
10.2 Minimax hypothesis testing . . . 263
10.2.1 Detecting a difference in populations . . . 264
10.2.2 Signal detection and testing a Gaussian mean . . . 265
10.2.3 Goodness of fit and two-sample tests for multinomials . . . 267
10.3 Geometrizing rates of convergence . . . 271
10.4 Best possible lower bounds and super-efficiency . . . 271
10.5 Bibliography . . . 271
10.6 A useful divergence calculation . . . 272
10.7 Exercises . . . 274

III Entropy, predictions, divergences, and information **277**

11 Predictions, loss functions, and entropies **278**
11.1 Proper losses, scoring rules, and generalized entropies . . . 279
11.1.1 A convexity primer . . . 280

11.1.2 From a proper loss to an entropy . . . 282
11.1.3 The information in an experiment . . . 283
11.2 Characterizing proper losses and Bregman divergences . . . 285
11.2.1 Characterizing proper losses for Y taking finitely many vales . . . 285
11.2.2 General proper losses . . . 288
11.2.3 Proper losses and vector-valued Y . . . 291
11.3 From entropies to convex losses, arbitrary predictions, and link functions . . . 294
11.3.1 Convex conjugate linkages . . . 294
11.3.2 Convex conjugate linkages with affine constraints . . . 298
11.4 Exponential families, maximum entropy, and log loss . . . 301
11.4.1 Maximizing entropy . . . 303
11.4.2 I-projections and maximum likelihood . . . 307
11.5 Technical and deferred proofs . . . 308
11.5.1 Finalizing the proof of Theorem 11.2.14 . . . 308
11.5.2 Proof of Proposition 11.4.1 . . . 309
11.5.3 Proof of Proposition 11.4.3 . . . 310
11.6 Exercises . . . 311

12 Calibration and Proper Losses **315**
12.1 Proper losses and calibration error . . . 316
12.2 Measuring calibration . . . 319
12.2.1 The impossibility of measuring calibration . . . 319
12.2.2 Alternative calibration measures . . . 322
12.3 Auditing and improving calibration at the population level . . . 325
12.3.1 The post-processing gap and calibration audits for squared error . . . 325
12.3.2 Calibration audits for losses based on conjugate linkages . . . 327
12.3.3 A population-level algorithm for calibration . . . 329
12.4 Calibeating: improving squared error by calibration . . . 330
12.4.1 Proof of Theorem 12.4.1 . . . 333
12.5 Continuous and equivalent calibration measures . . . 336
12.5.1 Calibration measures . . . 337
12.5.2 Equivalent calibration measures . . . 339
12.6 Deferred technical proofs . . . 345
12.6.1 Proof of Lemma 12.2.1 . . . 345
12.6.2 Proof of Proposition 12.5.2 . . . 346
12.6.3 Proof of Lemma 12.5.4 . . . 347
12.6.4 Proof of Theorem 12.5.6 . . . 348
12.7 Bibliography . . . 350
12.8 Exercises . . . 351

13 Surrogate Risk Consistency: the Classification Case **352**
13.1 General results . . . 358
13.2 Proofs of convex analytic results . . . 358
13.2.1 Proof of Lemma 13.0.4 . . . 358
13.2.2 Proof of Lemma 13.0.4 . . . 359
13.2.3 Proof of Lemma 13.0.6 . . . 359
13.3 Exercises . . . 359

14 Divergences, classification, and risk **362**
14.1 Generalized entropies . . . 368
14.2 From entropy to losses . . . 368
14.2.1 Classification case . . . 368
14.2.2 Structured prediction case . . . 368
14.3 Predictions, calibration, and scoring rules . . . 369
14.4 Surrogate risk consistency . . . 369
14.4.1 Uniformly convex case . . . 369
14.4.2 Structured prediction (discrete) case . . . 369
14.4.3 Proof of Theorem 14.4.1 . . . 370
14.5 Loss equivalence . . . 371
14.6 Proof of Theorem 14.5.1 . . . 373
14.6.1 Proof of Lemma 14.6.2 . . . 375
14.6.2 Proof of Lemma 14.6.4 . . . 376
14.7 Bibliography . . . 376
14.8 Exercises . . . 376

15 Fisher Information **378**
15.1 Fisher information: definitions and examples . . . 378
15.2 Estimation and Fisher information: elementary considerations . . . 380
15.3 Connections between Fisher information and divergence measures . . . 381

IV Online game playing and compression **384**

16 Universal prediction and coding **385**
16.1 Basics of minimax game playing with log loss . . . 385
16.2 Universal and sequential prediction . . . 387
16.3 Minimax strategies for regret . . . 389
16.4 Mixture (Bayesian) strategies and redundancy . . . 391
16.4.1 Bayesian redundancy and objective, reference, and Jeffreys priors . . . 394
16.4.2 Redundancy capacity duality . . . 396
16.5 Asymptotic normality and Theorem 16.4.1 . . . 396
16.5.1 Heuristic justification of asymptotic normality . . . 397
16.5.2 Heuristic calculations of posterior distributions and redundancy . . . 397
16.6 Proof of Theorem 16.4.5 . . . 398
16.7 Exercises . . . 400

17 Universal prediction with other losses **403**
17.1 Redudancy and expected regret . . . 403
17.1.1 Universal prediction via the log loss . . . 404
17.1.2 Examples . . . 406
17.2 Individual sequence prediction and regret . . . 408

18 Online convex optimization **413**
18.1 The problem of online convex optimization . . . 413
18.2 Online gradient and non-Euclidean gradient (mirror) descent . . . 415
18.2.1 Proof of Theorem 18.2.5 . . . 419
18.3 Online to batch conversions . . . 421
18.4 More refined convergence guarantees . . . 421
18.4.1 Proof of Proposition 18.4.1 . . . 422

19 Exploration, exploitation, and bandit problems **424**
19.1 Confidence-based algorithms . . . 425
19.2 Bayesian approaches to bandits . . . 429
19.2.1 Posterior (Thompson) sampling . . . 430
19.2.2 An information-theoretic analysis . . . 433
19.2.3 Information and exploration . . . 433
19.3 Online gradient descent approaches . . . 433
19.4 Further notes and references . . . 435
19.5 Technical proofs . . . 435
19.5.1 Proof of Claim (19.1.1) . . . 435

V Appendices **437**

A Miscellaneous mathematical results **438**
A.1 The roots of a polynomial . . . 438
A.2 Measure-theoretic development of divergence measures . . . 438

B Convex Analysis **439**
B.1 Convex sets . . . 439
B.1.1 Operations preserving convexity . . . 441
B.1.2 Representation and separation of convex sets . . . 443
B.2 Sublinear and support functions . . . 446
B.3 Convex functions . . . 449
B.3.1 Equivalent definitions of convex functions . . . 450
B.3.2 Continuity properties of convex functions . . . 452
B.3.3 Operations preserving convexity . . . 458
B.3.4 Smoothness properties, first-order developments for convex functions, and subdifferentiability . . . 460
B.3.5 Calculus rules of subgradients . . . 465

C Optimality, stability, and duality **468**
C.1 Optimality conditions and stability properties . . . 469
C.1.1 Subgradient characterizations for optimality . . . 469
C.1.2 Stability properties of minimizers . . . 471
C.2 Conjugacy and duality properties . . . 475
C.2.1 Gradient dualities and the Fenchel-Young inequality . . . 476
C.2.2 Smoothness and strict convexity of conjugates . . . 478
C.3 Exercises . . . 483

Chapter 1

Introduction and setting

This set of lecture notes explores some of the (many) connections relating information theory, statistics, computation, and learning. Signal processing, machine learning, and statistics all revolve around extracting useful information from signals and data. In signal processing and information theory, a central question is how to best *design* signals—and the channels over which they are transmitted—to maximally communicate and store information, and to allow the most effective decoding. In machine learning and statistics, by contrast, it is often the case that there is a fixed data distribution that nature provides, and it is the learner's or statistician's goal to recover information about this (unknown) distribution.

A central aspect of information theory is the discovery of *fundamental* results: results that demonstrate that certain procedures are optimal. That is, information theoretic tools allow a characterization of the attainable results in a variety of communication and statistical settings. As we explore in these notes in the context of statistical, inferential, and machine learning tasks, this allows us to develop procedures whose optimality we can certify—no better procedure is possible. Such results are useful for a myriad of reasons; we would like to avoid making bad decisions or false inferences, we may realize a task is impossible, and we can explicitly calculate the amount of data necessary for solving different statistical problems.

1.1 Information theory

Information theory is a broad field, but focuses on several main questions: what is information, how much information content do various signals and data hold, and how much information can be reliably transmitted over a channel. We will vastly oversimplify information theory into two main questions with corresponding chains of tasks.

1. How much information does a signal contain?
2. How much information can a noisy channel reliably transmit?

In this context, we provide two main high-level examples, one for each of these tasks.

Example 1.1.1 (Source coding)**:** The source coding, or data compression problem, is to take information from a source, compress it, decompress it, and recover the original message. Graphically, we have

$$\mathsf{Source} \rightarrow \mathsf{Compressor} \rightarrow \mathsf{Decompressor} \rightarrow \mathsf{Receiver}$$

The question, then, is how to design a compressor (encoder) and decompressor (decoder) that uses the fewest number of bits to describe a source (or a message) while preserving all the information, in the sense that the receiver receives the correct message with high probability. This fewest number of bits is then the information content of the source (signal). $\diamond$

Example 1.1.2: The channel coding, or data transmission problem, is the same as the source coding problem of Example 1.1.1, except that between the compressor and decompressor is a source of noise, a *channel*. In this case, the graphical representation is

$$\text{Source} \rightarrow \text{Compressor} \rightarrow \text{Channel} \rightarrow \text{Decompressor} \rightarrow \text{Receiver}$$

Here the question is the maximum number of bits that may be sent per each channel use in the sense that the receiver may reconstruct the desired message with low probability of error. Because the channel introduces noise, we require some redundancy, and information theory studies the exact amount of redundancy and number of bits that must be sent to allow such reconstruction. $\diamond$

1.2 Moving to statistics

Statistics and machine learning can—broadly—be studied with the same views in mind. Broadly, statistics and machine learning can be thought of as (perhaps shoehorned into) source coding and a channel coding problems.

In the analogy with source coding, we observe a sequence of data points $X_1, \ldots, X_n$ drawn from some (unknown) distribution P on a space $\mathcal{X}$. For example, we might be observing species that biologists collect. Then the analogue of source coding is to construct a model (often a generative model) that encodes the data using relatively few bits: that is,

$$\text{Source } (P) \overset{X_1,\ldots,X_n}{\longrightarrow} \text{Compressor} \overset{\widehat{P}}{\rightarrow} \text{Decompressor} \rightarrow \text{Receiver}.$$

Here, we estimate $\widehat{P}$—an empirical version of the distribution P that is easier to describe than the original signal $X_1, \ldots, X_n$, with the hope that we learn information about the generating distribution P, or at least describe it efficiently.

In our analogy with channel coding, we make a connection with estimation and inference. Roughly, the major problem in statistics we consider is as follows: there exists some unknown function f on a space $\mathcal{X}$ that we wish to estimate, and we are able to observe a noisy version of $f(X_i)$ for a series of X_i drawn from a distribution P. Recalling the graphical description of Example 1.1.2, we now have a channel $P(Y \mid f(X))$ that gives us noisy observations of $f(X)$ for each X_i, but we may (generally) now longer choose the encoder/compressor. That is, we have

$$\text{Source } (P) \overset{X_1,\ldots,X_n}{\longrightarrow} \text{Compressor} \overset{f(X_1),\ldots,f(X_n)}{\longrightarrow} \text{Channel } P(Y \mid f(X)) \overset{Y_1,\ldots,Y_n}{\longrightarrow} \text{Decompressor}.$$

The estimation—decompression—problem is to either estimate f, or, in some cases, to estimate other aspects of the source probability distribution P. In general, in statistics, we do not have any choice in the design of the compressor f that transforms the original signal $X_1, \ldots, X_n$, which makes it somewhat different from traditional ideas in information theory. In some cases that we explore later—such as experimental design, randomized controlled trials, reinforcement learning and bandits (and associated exploration/exploitation tradeoffs)—we are also able to influence the compression part of the above scheme.

Example 1.2.1: A classical example of the statistical paradigm in this lens is the usual linear regression problem. Here the data X_i belong to $\mathbb{R}^d$, and the compression function $f(x) = \theta^\top x$ for some vector $\theta \in \mathbb{R}^d$. Then the channel is often of the form

$$Y_i = \underbrace{\theta^\top X_i}_{\text{signal}} + \underbrace{\varepsilon_i}_{\text{noise}},$$

where $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$ are independent mean zero normal perturbations. The goal is, given a sequence of pairs (X_i, Y_i), to recover the true θ in the linear model.
In *active learning* or *active sensing* scenarios, also known as (sequential) experimental design, we may choose the sequence X_i so as to better explore properties of θ. Later in the course we will investigate whether it is possible to improve estimation by these strategies. As one concrete idea, if we allow infinite *power*, which in this context corresponds to letting $\|X_i\| \to \infty$—choosing very "large" vectors x_i—then the signal of $\theta^\top X_i$ should swamp any noise and make estimation easier. $\diamond$

For the remainder of the class, we explore these ideas in substantially more detail.

1.3 A remark about measure theory

As this book focuses on a number of fundamental questions in statistics, machine learning, and information theory, fully general statements of the results often require measure theory. Thus, formulae such as $\int f(x) dP(x)$ or $\int f(x) d\mu(x)$ appear. While knowledge of measure theory is certainly useful and may help appreciate the results, it is completely inessential to developing the intuition and, I hope, understanding the proofs and main results. Indeed, the best strategy (for a reader unfamiliar with measure theory) is to simply replace every instance of a formula such as $d\mu(x)$ with dx. The most frequent cases we encounter will be the following: we wish to compute the expectation of a function f of random variable X following distribution P, that is, $\mathbb{E}_P[f(X)]$. Normally, we would write $\mathbb{E}_P[f(X)] = \int f(x) dP(x)$, or sometimes $\mathbb{E}_P[f(X)] = \int f(x) p(x) d\mu(x)$, saying that "$P$ has density p with respect to the underlying measure μ." Instead, one may simply (and intuitively) assume that x really has density p over the reals, and instead of computing the integral

$$\mathbb{E}_P[f(X)] = \int f(x) dP(x) \quad \text{or} \quad \mathbb{E}_P[f(X)] = \int f(x) p(x) d\mu(x),$$

assume we may write

$$\mathbb{E}_P[f(X)] = \int f(x) p(x) dx.$$

Nothing will be lost.

1.4 Outline and chapter discussion

We divide the lecture notes into four distinct parts, each of course interacting with the others, but it is possible to read each as a reasonably self-contained unit. The lecture notes begin with a revew (Chapter 2) that introduces the basic information-theoretic quantities that we discuss: mutual information, entropy, and divergence measures. It is required reading for all the chapters that follow.

Part I of the notes covers what I term "stability" based results. At a high level, this means that we ask what can be gained by considering situations where individual observations in a sequence of random variables $X_1, \ldots, X_n$ have little effect on various functions of the sequence. We begin in Chapter 4 with basic concentration inequalities, discussing how sums and related quantities can converge quickly; while this material is essential for the remainder of the lectures, it does not depend on particular information-theoretic techniques. We discuss some heuristic applications to problems in statistical learning—empirical risk minimization—in this section of the notes. We provide a treatment of more advanced ideas in Chapter 6, including some approaches to concentration via entropy methods. We then turn in Chapter 5 carefully investigate generalization and convergence guarantees—arguing that functions of a sample $X_1, \ldots, X_n$ are representative of the full population P from which the sample is drawn—based on controlling different information-theoretic quantities. In this context, we develop PAC-Bayesian bounds, and we also use the same framework to present tools to control generalization and convergence in *interactive* data analyses. These types of analyses reflect modern statistics, where one performs some type of data exploration before committing to a fuller analysis, but which breaks classical statistical approaches, because the analysis now depends on the sample. Finally, we provide a chapter (Chapter 7) on disclosure limitation and privacy techniques, all of which repose on different notions of stability in distribution.

Part II studies fundamental limits, using information-theoretic techniques to derive *lower bounds* on the possible rates of convergence for various estimation, learning, and other statistical problems.

Part III revisits all of our information theoretic notions from Chapter 2, but instead of simply giving definitions and a few consequences, provides operational interpretations of the different information-theoretic quantities, such as entropy. Of course this includes Shannon's original results on the relationship between coding and entropy (Chapter 2.4.1), but we also provide an interpretation of entropy and information as measures of uncertainty in statistical experiments and statistical learning, which is a perspective typically missing from information-theoretic treatments of entropy (Chapters **TBD**). We also relate these ideas to game-playing and maximum likelihood estimation. Finally, we relate generic divergence measures to questions of optimality and consistency in statistical and machine learning problems, which allows us to delineate when (at least in asymptotic senses) it is possible to computationally efficiently learn good predictors and design good experiments.

Chapter 2

An information theory review

In this first introductory chapter, we discuss and review many of the basic concepts of information theory in effort to introduce them to readers unfamiliar with the tools. Our presentation is relatively brisk, as our main goal is to get to the meat of the chapters on applications of the inequalities and tools we develop, but these provide the starting point for everything in the sequel. One of the main uses of information theory is to prove what, in an information theorist's lexicon, are known as *converse results*: fundamental limits that guarantee no procedure can improve over a particular benchmark or baseline. We will give the first of these here to preview more of what is to come, as these fundamental limits form one of the core connections between statistics and information theory. The tools of information theory, in addition to their mathematical elegance, also come with strong operational interpretations: they give quite precise answers and explanations for a variety of real engineering and statistical phenomena. We will touch on one of these here (the connection between source coding, or lossless compression, and the Shannon entropy), and much of the remainder of the book will explore more.

2.1 Basics of Information Theory

In this section, we review the basic definitions in information theory, including (Shannon) entropy, KL-divergence, mutual information, and their conditional versions. Before beginning, I must make an apology to any information theorist reading these notes: any time we use a log, it will always be base-e. This is more convenient for our analyses, and it also (later) makes taking derivatives much nicer.

In this first section, we will assume that all distributions are discrete; this makes the quantities somewhat easier to manipulate and allows us to completely avoid any complicated measure-theoretic quantities. In Section 2.2 of this note, we show how to extend the important definitions (for our purposes)—those of KL-divergence and mutual information—to general distributions, where basic ideas such as entropy no longer make sense. However, even in this general setting, we will see we essentially lose no generality by assuming all variables are discrete.

2.1.1 Definitions

Here, we provide the basic definitions of entropy, information, and divergence, assuming the random variables of interest are discrete or have densities with respect to Lebesgue measure.

Entropy: We begin with a central concept in information theory: the entropy. Let P be a distribution on a finite (or countable) set $\mathcal{X}$, and let p denote the probability mass function associated with P. That is, if X is a random variable distributed according to P, then $P(X = x) = p(x)$. The *entropy of* X (or of P) is defined as

$$H(X) := -\sum_x p(x) \log p(x).$$

Because $p(x) \le 1$ for all x, it is clear that this quantity is positive. We will show later that if $\mathcal{X}$ is finite, the maximum entropy distribution on $\mathcal{X}$ is the uniform distribution, setting $p(x) = 1/|\mathcal{X}|$ for all x, which has entropy $\log(|\mathcal{X}|)$.

Later in the class, we provide a number of operational interpretations of the entropy. The most common interpretation—which forms the beginning of Shannon's classical information theory [158]—is via the source-coding theorem. We present Shannon's source coding theorem in Section 2.4.1, where we show that if we wish to encode a random variable X, distributed according to P, with a k-ary string (i.e. each entry of the string takes on one of k values), then the minimal expected length of the encoding is given by $H(X) = -\sum_x p(x) \log_k p(x)$. Moreover, this is achievable (to within a length of at most 1 symbol) by using Huffman codes (among many other types of codes). As an example of this interpretation, we may consider encoding a random variable X with equi-probable distribution on m items, which has $H(X) = \log(m)$. In base-2, this makes sense: we simply assign an integer to each item and encode each integer with the natural (binary) integer encoding of length $\lceil \log m \rceil$.

We can also define the *conditional entropy*, which is the amount of information left in a random variable after observing another. In particular, we define

$$H(X \mid Y = y) = -\sum_x p(x \mid y) \log p(x \mid y) \quad \text{and} \quad H(X \mid Y) = \sum_y p(y) H(X \mid Y = y),$$

where $p(x \mid y)$ is the p.m.f. of X given that $Y = y$.

Let us now provide a few examples of the entropy of various discrete random variables

Example 2.1.1 (Uniform random variables)**:** As we noted earlier, if a random variable X is uniform on a set of size m, then $H(X) = \log m$. $\diamond$

Example 2.1.2 (Bernoulli random variables)**:** Let $h_2(p) = -p \log p - (1-p) \log(1-p)$ denote the binary entropy, which is the entropy of a $\mathsf{Bernoulli}(p)$ random variable. $\diamond$

Example 2.1.3 (Geometric random variables)**:** A random variable X is $\mathsf{Geometric}(p)$, for some $p \in [0, 1]$, if it is supported on $\{1, 2, \ldots\}$, and $P(X = k) = (1-p)^{k-1} p$; this is the probability distribution of the number X of $\mathsf{Bernoulli}(p)$ trials until a single success. The entropy of such a random variable is

$$H(X) = -\sum_{k=1}^{\infty} (1-p)^{k-1} p \left[(k-1) \log(1-p) + \log p\right] = -\sum_{k=0}^{\infty} (1-p)^k p \left[k \log(1-p) + \log p\right].$$

As $\sum_{k=0}^{\infty} \alpha^k = \frac{1}{1-\alpha}$ and $\frac{d}{d\alpha} \frac{1}{1-\alpha} = \frac{1}{(1-\alpha)^2} = \sum_{k=1}^{\infty} k \alpha^{k-1}$, we have

$$H(X) = -p \log(1-p) \cdot \sum_{k=1}^{\infty} k(1-p)^k - p \log p \cdot \sum_{k=1}^{\infty} (1-p)^k = -\frac{1-p}{p} \log(1-p) - (1-p) \log p.$$

As $p \downarrow 0$, we see that $H(X) \uparrow \infty$. $\diamond$

Example 2.1.4 (A random variable with infinite entropy)**:** While most "reasonable" discrete random variables have finite entropy, it is possible to construct distributions with infinite entropy. Indeed, let X have p.m.f. on $\{2, 3, \ldots\}$ defined by

$$p(k) = \frac{A}{k \log^2 k} \quad \text{where} \quad A^{-1} = \sum_{k=2}^{\infty} \frac{1}{k \log^2 k} < \infty,$$

the last sum finite as $\int_2^\infty \frac{1}{x \log^\alpha x} dx < \infty$ if and only if $\alpha > 1$: for $\alpha = 1$, we have $\int_e^x \frac{1}{t \log t} = \log \log x$, while for $\alpha > 1$, we have

$$\frac{d}{dx} (\log x)^{1-\alpha} = (1 - \alpha) \frac{1}{x \log^\alpha x}$$

so that $\int_e^\infty \frac{1}{t \log^\alpha t} dt = \frac{1}{e(1-\alpha)}$. To see that the entropy is infinite, note that

$$H(X) = A \sum_{k \geq 2} \frac{\log A + \log k + 2 \log \log k}{k \log^2 k} \geq A \sum_{k \geq 2} \frac{\log k}{k \log^2 k} - C = \infty,$$

where C is a numerical constant. $\diamond$

KL-divergence: Now we define two additional quantities, which are actually *much more* fundamental than entropy: they can always be defined for any distributions and any random variables, as they measure distance between distributions. Entropy simply makes no sense for non-discrete random variables, let alone random variables with continuous and discrete components, though it proves useful for some of our arguments and interpretations.

Before defining these quantities, we recall the definition of a convex function $f : \mathbb{R}^k \to \mathbb{R}$ as any bowl-shaped function, that is, one satisfying

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y) \tag{2.1.1}$$

for all $\lambda \in [0, 1]$, all x, y. The function f is *strictly* convex if the convexity inequality (2.1.1) is strict for $\lambda \in (0, 1)$ and $x \neq y$. We recall a standard result:

Proposition 2.1.5 (Jensen's inequality)**.** *Let f be convex. Then for any random variable X,*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Moreover, if f is strictly convex*, then $f(\mathbb{E}[X]) < \mathbb{E}[f(X)]$ unless X is constant.*

Now we may define and provide a few properties of the KL-divergence. Let P and Q be distributions defined on a discrete set $\mathcal{X}$. The *KL-divergence* between them is

$$D_{\rm kl}\left(P \| Q\right) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

We observe immediately that $D_{\rm kl}\left(P \| Q\right) \geq 0$. To see this, we apply Jensen's inequality (Proposition 2.1.5) to the function $-\log$ and the random variable $q(X)/p(X)$, where X is distributed according to P:

$$\begin{aligned} D_{\rm kl}\left(P \| Q\right) &= -\mathbb{E}\left[\log \frac{q(X)}{p(X)}\right] \geq -\log \mathbb{E}\left[\frac{q(X)}{p(X)}\right] \\ &= -\log\left(\sum_x p(x) \frac{q(x)}{p(x)}\right) = -\log(1) = 0. \end{aligned}$$

Moreover, as log is strictly convex, we have $D_{\rm kl}\left(P\|Q\right) > 0$ unless $P = Q$. Another consequence of the positivity of the KL-divergence is that whenever the set $\mathcal{X}$ is finite with cardinality $|\mathcal{X}| < \infty$, for any random variable X supported on $\mathcal{X}$ we have $H(X) \le \log|\mathcal{X}|$. Indeed, letting $m = |\mathcal{X}|$, Q be the uniform distribution on $\mathcal{X}$ so that $q(x) = \frac{1}{m}$, and X have distribution P on $\mathcal{X}$, we have

$$0 \le D_{\rm kl}\left(P\|Q\right) = \sum_x p(x) \log \frac{p(x)}{q(x)} = -H(X) - \sum_x p(x) \log q(x) = -H(X) + \log m, \tag{2.1.2}$$

so that $H(X) \le \log m$. Thus, the uniform distribution has the highest entropy over all distributions on the set $\mathcal{X}$.

Mutual information: Having defined KL-divergence, we may now describe the information content between two random variables X and Y. The *mutual information* $I(X;Y)$ between X and Y is the KL-divergence between their joint distribution and their products (marginal) distributions. More mathematically,

$$I(X;Y) := \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{2.1.3}$$

We can rewrite this in several ways. First, using Bayes' rule, we have $p(x,y)/p(y) = p(x \mid y)$, so

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(y)p(x \mid y) \log \frac{p(x \mid y)}{p(x)} \\ &= -\sum_x \sum_y p(y)p(x \mid y) \log p(x) + \sum_y p(y) \sum_x p(x \mid y) \log p(x \mid y) \\ &= H(X) - H(X \mid Y). \end{aligned}$$

Similarly, we have $I(X;Y) = H(Y) - H(Y \mid X)$, so mutual information can be thought of as the amount of entropy removed (on average) in X by observing Y. We may also think of mutual information as measuring the similarity between the joint distribution of X and Y and their distribution when they are treated as independent.

Comparing the definition (2.1.3) to that for KL-divergence, we see that if P_{XY} is the joint distribution of X and Y, while P_X and P_Y are their marginal distributions (distributions when X and Y are treated independently), then

$$I(X;Y) = D_{\rm kl}\left(P_{XY}\|P_X \times P_Y\right) \ge 0.$$

Moreover, we have $I(X;Y) > 0$ unless X and Y are independent.

As with entropy, we may also define the *conditional information between X and Y given Z*, which is the mutual information between X and Y when Z is observed (on average). That is,

$$I(X;Y \mid Z) := \sum_z I(X;Y \mid Z = z)p(z) = H(X \mid Z) - H(X \mid Y,Z) = H(Y \mid Z) - H(Y \mid X,Z).$$

Entropies of continuous random variables For continuous random variables, we may define an analogue of the entropy known as *differential entropy*, which for a random variable X with density p is defined by

$$h(X) := -\int p(x) \log p(x) dx. \tag{2.1.4}$$

Note that the differential entropy may be negative—it is no longer directly a measure of the number of bits required to describe a random variable X (on average), as was the case for the entropy. We can similarly define the conditional entropy

$$h(X \mid Y) = -\int p(y) \int p(x \mid y) \log p(x \mid y) dx dy.$$

We remark that the conditional differential entropy of X given Y for Y with arbitrary distribution—so long as X has a density—is

$$h(X \mid Y) = \mathbb{E}\left[-\int p(x \mid Y) \log p(x \mid Y) dx\right],$$

where $p(x \mid y)$ denotes the conditional density of X when $Y = y$. The KL divergence between distributions P and Q with densities p and q becomes

$$D_{\rm kl}\left(P \| Q\right) = \int p(x) \log \frac{p(x)}{q(x)} dx,$$

and similarly, we have the analogues of mutual information as

$$I(X; Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy = h(X) - h(X \mid Y) = h(Y) - h(Y \mid X).$$

As we show in the next subsection, we can define the KL-divergence between arbitrary distributions (and mutual information between arbitrary random variables) more generally without requiring discrete or continuous distributions. Before investigating these issues, however, we present a few examples. We also see immediately that for X uniform on a set $[a, b]$, we have $h(X) = \log(b - a)$.

Example 2.1.6 (Entropy of normal random variables)**:** The differential entropy (2.1.4) of a normal random variable is straightforward to compute. Indeed, for $X \sim \mathsf{N}(\mu, \sigma^2)$ we have $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x - \mu)^2)$, so that

$$h(X) = -\int p(x) \left[\frac{1}{2} \log \frac{1}{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x - \mu)^2\right] = \frac{1}{2} \log(2\pi\sigma^2) + \frac{\mathbb{E}[(X - \mu)^2]}{2\sigma^2} = \frac{1}{2} \log(2\pi e \sigma^2).$$

For a general multivariate Gaussian, where $X \sim \mathsf{N}(\mu, \Sigma)$ for a vector $\mu \in \mathbb{R}^n$ and $\Sigma \succ 0$ with density $p(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu))$, we similarly have

$$\begin{aligned} h(X) &= \frac{1}{2}\mathbb{E}\left[n \log(2\pi) + \log \det(\Sigma) + (X - \mu)^\top \Sigma^{-1}(X - \mu)\right] \\ &= \frac{n}{2} \log(2\pi) + \frac{1}{2} \log \det(\Sigma) + \frac{1}{2} \operatorname{tr}(\Sigma\Sigma^{-1}) = \frac{n}{2} \log(2\pi e) + \frac{1}{2} \log \det(e\Sigma). \end{aligned}$$

$\diamond$

Continuing our examples with normal distributions, we may compute the divergence between two multivariate Gaussian distributions:

Example 2.1.7 (Divergence between Gaussian distributions)**:** Let P be the multivariate normal $\mathsf{N}(\mu_1, \Sigma)$, and Q be the multivariate normal distribution with mean μ_2 and identical covariance $\Sigma \succ 0$. Then we have that

$$D_{\rm kl}\left(P \| Q\right) = \frac{1}{2}(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2). \tag{2.1.5}$$

We leave the computation of the identity (2.1.5) to the reader. $\diamond$

An interesting consequence of Example 2.1.7 is that if a random vector X has a given covariance $\Sigma \in \mathbb{R}^{n \times n}$, then the multivariate Gaussian with identical covariance has larger differential entropy. Put another way, differential entropy for random variables with second moments is always maximized by the Gaussian distribution.

Proposition 2.1.8. *Let X be a random vector on $\mathbb{R}^n$ with a density, and assume that $\mathrm{Cov}(X) = \Sigma$. Then for $Z \sim \mathsf{N}(0, \Sigma)$, we have*

$$h(X) \le h(Z).$$

Proof Without loss of generality, we assume that X has mean 0. Let P be the distribution of X with density p, and let Q be multivariate normal with mean 0 and covariance Σ; let Z be this random variable. Then

$$\begin{aligned} D_{\rm kl}\left(P \| Q\right) = \int p(x) \log \frac{p(x)}{q(x)} dx &= -h(X) + \int p(x) \left[\frac{n}{2} \log(2\pi) - \frac{1}{2} x^\top \Sigma^{-1} x\right] dx \\ &= -h(X) + h(Z), \end{aligned}$$

because Z has the same covariance as X. As $0 \le D_{\rm kl}\left(P \| Q\right)$, we have $h(Z) \ge h(X)$ as desired. □

We remark in passing that the fact that Gaussian random variables have the largest entropy has been used to prove stronger variants of the central limit theorem; see the original results of Barron [16], as well as later quantitative results on the increase of entropy of normalized sums by Artstein et al. [9] and Madiman and Barron [134].

2.1.2 Chain rules and related properties

We now illustrate several of the properties of entropy, KL divergence, and mutual information; these allow easier calculations and analysis.

Chain rules: We begin by describing relationships between collections of random variables $X_1, \ldots, X_n$ and individual members of the collection. (Throughout, we use the notation $X_i^j = (X_i, X_{i+1}, \ldots, X_j)$ to denote the sequence of random variables from indices i through j.)

For the entropy, we have the simplest chain rule:

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1^{n-1}).$$

This follows from the standard decomposition of a probability distribution $p(x, y) = p(x)p(y \mid x)$. to see the chain rule, then, note that

$$\begin{aligned} H(X, Y) &= -\sum_{x,y} p(x)p(y \mid x) \log p(x)p(y \mid x) \\ &= -\sum_x p(x) \sum_y p(y \mid x) \log p(x) - \sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) = H(X) + H(Y \mid X). \end{aligned}$$

Now set $X = X_1^{n-1}$, $Y = X_n$, and simply induct.

A related corollary of the definitions of mutual information is the well-known result that *conditioning reduces entropy*:

$$H(X \mid Y) \le H(X) \quad \text{because} \quad I(X; Y) = H(X) - H(X \mid Y) \ge 0.$$

So on average, knowing about a variable Y can only decrease your uncertainty about X. That conditioning reduces entropy for continuous random variables is also immediate, as for X continuous we have $I(X; Y) = h(X) - h(X \mid Y) \ge 0$, so that $h(X) \ge h(X \mid Y)$.

Chain rules for information and divergence: As another immediate corollary to the chain rule for entropy, we see that mutual information also obeys a chain rule:

$$I(X; Y_1^n) = \sum_{i=1}^n I(X; Y_i \mid Y_1^{i-1}).$$

Indeed, we have

$$I(X; Y_1^n) = H(Y_1^n) - H(Y_1^n \mid X) = \sum_{i=1}^n \left[H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid X, Y_1^{i-1})\right] = \sum_{i=1}^n I(X; Y_i \mid Y_1^{i-1}).$$

The KL-divergence obeys similar chain rules, making mutual information and KL-divergence measures useful tools for evaluation of distances and relationships between groups of random variables.

As a second example, suppose that the distribution $P = P_1 \times P_2 \times \cdots \times P_n$, and $Q = Q_1 \times \cdots \times Q_n$, that is, that P and Q are product distributions over independent random variables $X_i \sim P_i$ or $X_i \sim Q_i$. Then we immediately have the tensorization identity

$$D_{\rm kl}\left(P \| Q\right) = D_{\rm kl}\left(P_1 \times \cdots \times P_n \| Q_1 \times \cdots \times Q_n\right) = \sum_{i=1}^n D_{\rm kl}\left(P_i \| Q_i\right).$$

We remark in passing that these two identities hold for arbitrary distributions P_i and Q_i or random variables X, Y. As a final tensorization identiy, we consider a more general chain rule for KL-divergences, which will frequently be useful. We abuse notation temporarily, and for random variables X and Y with distributions P and Q, respectively, we denote

$$D_{\rm kl}\left(X \| Y\right) := D_{\rm kl}\left(P \| Q\right).$$

In analogy to the entropy, we can also define the *conditional KL divergence*. Let X and Y have distributions $P_{X|z}$ and $P_{Y|z}$ conditioned on $Z = z$, respectively. Then we define

$$D_{\rm kl}\left(X \| Y \mid Z\right) = \mathbb{E}_Z[D_{\rm kl}\left(P_{X|Z} \| P_{Y|Z}\right)],$$

so that if Z is discrete we have $D_{\rm kl}\left(X \| Y \mid Z\right) = \sum_z p(z) D_{\rm kl}\left(P_{X|z} \| P_{Y|z}\right)$. With this notation, we have the chain rule

$$D_{\rm kl}\left(X_1, \ldots, X_n \| Y_1, \ldots, Y_n\right) = \sum_{i=1}^n D_{\rm kl}\left(X_i \| Y_i \mid X_1^{i-1}\right), \tag{2.1.6}$$

because (in the discrete case, which—as we discuss presently—is fully general for this purpose) for distributions P_{XY} and Q_{XY} we have

$$\begin{aligned} D_{\rm kl}\left(P_{XY} \| Q_{XY}\right) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} = \sum_{x,y} p(x) p(y \mid x) \left[\log \frac{p(y \mid x)}{q(y \mid x)} + \log \frac{p(x)}{q(x)}\right] \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \sum_y p(y \mid x) \log \frac{p(y \mid x)}{q(y \mid x)}, \end{aligned}$$

where the final equality uses that $\sum_y p(y \mid x) = 1$ for all x.

Expanding upon this, we give several *tensorization* identities, showing how to transform questions about the joint distribution of many random variables to simpler questions about their

marginals. As a first example, we see that as a consequence of the fact that conditioning decreases entropy, we see that for any sequence of (discrete or continuous, as appropriate) random variables, we have

$$H(X_1, \ldots, X_n) \le H(X_1) + \cdots + H(X_n) \text{ and } h(X_1, \ldots, X_n) \le h(X_1) + \ldots + h(X_n).$$

Both equalities hold with equality if and only if $X_1, \ldots, X_n$ are mutually independent. (The only if follows because $I(X;Y) > 0$ whenever X and Y are not independent, by Jensen's inequality and the fact that $D_{\rm kl}(P\|Q) > 0$ unless $P = Q$.)

We return to information and divergence now. Suppose that random variables Y_i are independent conditional on X, meaning that

$$P(Y_1 = y_1, \ldots, Y_n = y_n \mid X = x) = P(Y_1 = y_1 \mid X = x) \cdots P(Y_n = y_n \mid X = x).$$

Such scenarios are common—as we shall see—when we make multiple observations from a fixed distribution parameterized by some X. Then we have the inequality

$$\begin{aligned} I(X; Y_1, \ldots, Y_n) &= \sum_{i=1}^n [H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid X, Y_1^{i-1})] \\ &= \sum_{i=1}^n [H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid X)] \le \sum_{i=1}^n [H(Y_i) - H(Y_i \mid X)] = \sum_{i=1}^n I(X; Y_i), \end{aligned} \tag{2.1.7}$$

where the inequality follows because conditioning reduces entropy.

2.1.3 Data processing inequalities:

A standard problem in information theory (and statistical inference) is to understand the degradation of a signal after it is passed through some noisy channel (or observation process). The simplest of such results, which we will use frequently, is that we can only lose information by adding noise. In particular, assume we have the Markov chain

$$X \to Y \to Z.$$

Then we obtain the classical *data processing inequality*.

Proposition 2.1.9. *With the above Markov chain, we have $I(X;Z) \le I(X;Y)$.*

Proof We expand the mutual information $I(X; Y, Z)$ in two ways:

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y \mid Z) \\ &= I(X; Y) + \underbrace{I(X; Z \mid Y)}_{=0}, \end{aligned}$$

where we note that the final equality follows because X is independent of Z given Y:

$$I(X; Z \mid Y) = H(X \mid Y) - H(X \mid Y, Z) = H(X \mid Y) - H(X \mid Y) = 0.$$

Since $I(X; Y \mid Z) \ge 0$, this gives the result. □

There are related data processing inequalities for the KL-divergence—which we generalize in the next section—as well. In this case, we may consider a simple Markov chain $X \to Z$. If we let P_1 and P_2 be distributions on X and Q_1 and Q_2 be the induced distributions on Z, that is, $Q_i(A) = \int \mathbb{P}(Z \in A \mid x) dP_i(x)$, then we have

$$D_{\rm kl}\left(Q_1 \| Q_2\right) \le D_{\rm kl}\left(P_1 \| P_2\right),$$

the basic KL-divergence data processing inequality. A consequence of this is that, for any function f and random variables X and Y on the same space, we have

$$D_{\rm kl}\left(f(X) \| f(Y)\right) \le D_{\rm kl}\left(X \| Y\right).$$

We explore these data processing inequalities more when we generalize KL-divergences in the next section and in the exercises.

2.2 General divergence measures and definitions

Having given our basic definitions of mutual information and divergence, we now show how the definitions of KL-divergence and mutual information extend to arbitrary distributions P and Q and arbitrary sets $\mathcal{X}$. This requires a bit of setup, including defining set algebras (which, we will see, simply correspond to quantization of the set $\mathcal{X}$), but allows us to define divergences in full generality.

2.2.1 Partitions, algebras, and quantizers

Let $\mathcal{X}$ be an arbitrary space. A *quantizer* on $\mathcal{X}$ is any function that maps $\mathcal{X}$ to a finite collection of integers. That is, fixing $m < \infty$, a quantizer is any function $\mathsf{q} : \mathcal{X} \to \{1, \ldots, m\}$. In particular, a quantizer q partitions the space $\mathcal{X}$ into the subsets of $x \in \mathcal{X}$ for which $\mathsf{q}(x) = i$. A related notion—we will see the precise relationship presently—is that of an algebra of sets on $\mathcal{X}$. We say that a collection of sets $\mathcal{A}$ is an *algebra* on $\mathcal{X}$ if the following are true:

1. The set $\mathcal{X} \in \mathcal{A}$.
2. The collection of sets $\mathcal{A}$ is closed under finite set operations: union, intersection, and complementation. That is, $A, B \in \mathcal{A}$ implies that $A^c \in \mathcal{A}$, $A \cap B \in \mathcal{A}$, and $A \cup B \in \mathcal{A}$.

There is a 1-to-1 correspondence between quantizers—and their associated partitions of the set $\mathcal{X}$—and finite algebras on a set $\mathcal{X}$, which we discuss briefly.[1] It should be clear that there is a one-to-one correspondence between finite *partitions* of the set $\mathcal{X}$ and quantizers q, so we must argue that finite partitions of $\mathcal{X}$ are in one-to-one correspondence with finite algebras defined over $\mathcal{X}$.

In one direction, we may consider a quantizer $\mathsf{q} : \mathcal{X} \to \{1, \ldots, m\}$. Let the sets $A_1, \ldots, A_m$ be the partition associated with q, that is, for $x \in A_i$ we have $\mathsf{q}(x) = i$, or $A_i = \mathsf{q}^{-1}(\{i\})$. Then we may define an algebra $\mathcal{A}_{\mathsf{q}}$ as the collection of all finite set operations performed on $A_1, \ldots, A_m$ (note that this is a finite collection, as finite set operations performed on the partition $A_1, \ldots, A_m$ induce only a finite collection of sets).

For the other direction, consider a finite algebra $\mathcal{A}$ over the set $\mathcal{X}$. We can then construct a quantizer $\mathsf{q}_{\mathcal{A}}$ that corresponds to this algebra. To do so, we define an *atom* of $\mathcal{A}$ as any non-empty set $A \in \mathcal{A}$ such that if $B \subset A$ and $B \in \mathcal{A}$, then $B = A$ or $B = \emptyset$. That is, the atoms of $\mathcal{A}$ are the "smallest" sets in $\mathcal{A}$. We claim there is a unique partition of $\mathcal{X}$ with atomic sets from $\mathcal{A}$; we prove this inductively.

[1] Pedantically, this one-to-one correspondence holds up to permutations of the partition induced by the quantizer.

Base case: There is at least 1 atomic set, as $\mathcal{A}$ is finite; call it A_1.

Induction step: Assume we have atomic sets $A_1, \ldots, A_k \in \mathcal{A}$. Let $B = (A_1 \cup \cdots \cup A_k)^c$ be their complement, which we assume is non-empty (otherwise we have a partition of $\mathcal{X}$ into atomic sets). The complement B is either atomic, in which case the sets $\{A_1, A_2, \ldots, A_k, B\}$ are a partition of $\mathcal{X}$ consisting of atoms of $\mathcal{A}$, or B is not atomic. If B is not atomic, consider all the sets of the form $A \cap B$ for $A \in \mathcal{A}$. Each of these belongs to $\mathcal{A}$, and at least one of them is atomic, as there is a finite number of them. This means there is a non-empty set $A_{k+1} \subset B$ such that A_{k+1} is atomic.

By repeating this induction, which must stop at some finite index m as $\mathcal{A}$ is finite, we construct a collection $A_1, \ldots, A_m$ of disjoint atomic sets in $\mathcal{A}$ for which and $\cup_i A_i = \mathcal{X}$. (The uniqueness is an exercise for the reader.) Thus we may define the quantizer $\mathsf{q}_{\mathcal{A}}$ via

$$\mathsf{q}_{\mathcal{A}}(x) = i \quad \text{when } x \in A_i.$$

2.2.2 KL-divergence

In this section, we present the general definition of a KL-divergence, which holds for *any* pair of distributions. Let P and Q be distributions on a space $\mathcal{X}$. Now, let $\mathcal{A}$ be a finite algebra on $\mathcal{X}$ (as in the previous section, this is equivalent to picking a partition of $\mathcal{X}$ and then constructing the associated algebra), and assume that its atoms are $\mathsf{atoms}(\mathcal{A})$. The KL-divergence between P and Q *conditioned on* $\mathcal{A}$ is

$$D_{\rm kl}\left(P \| Q \mid \mathcal{A}\right) := \sum_{A \in \mathsf{atoms}(\mathcal{A})} P(A) \log \frac{P(A)}{Q(A)}.$$

That is, we simply sum over the partition of $\mathcal{X}$. Another way to write this is as follows. Let $\mathsf{q} : \mathcal{X} \to \{1, \ldots, m\}$ be a quantizer, and define the sets $A_i = \mathsf{q}^{-1}(\{i\})$ to be the pre-images of each i (i.e. the different quantization regions, or the partition of $\mathcal{X}$ that q induces). Then the *quantized* KL-divergence between P and Q is

$$D_{\rm kl}\left(P \| Q \mid \mathsf{q}\right) := \sum_{i=1}^{m} P(A_i) \log \frac{P(A_i)}{Q(A_i)}.$$

We may now give the fully general definition of KL-divergence: the KL-divergence between P and Q is defined as

$$\begin{aligned} D_{\rm kl}\left(P \| Q\right) &:= \sup\left\{D_{\rm kl}\left(P \| Q \mid \mathcal{A}\right) \quad \text{such that } \mathcal{A} \text{ is a finite algebra on } \mathcal{X}\right\} \\ &= \sup\left\{D_{\rm kl}\left(P \| Q \mid \mathsf{q}\right) \quad \text{such that } \mathsf{q} \text{ quantizes } \mathcal{X}\right\}. \end{aligned} \tag{2.2.1}$$

This also gives a rigorous definition of mutual information. Indeed, if X and Y are random variables with joint distribution P_{XY} and marginal distributions P_X and P_Y, we simply define

$$I(X; Y) = D_{\rm kl}\left(P_{XY} \| P_X \times P_Y\right).$$

When P and Q have densities p and q, the definition (2.2.1) reduces to

$$D_{\rm kl}\left(P \| Q\right) = \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx,$$

while if P and Q both have probability mass functions p and q, then—as we see in Exercise 2.6—the definition (2.2.1) is equivalent to

$$D_{\rm kl}\left(P\|Q\right) = \sum_x p(x) \log \frac{p(x)}{q(x)},$$

precisely as in the discrete case.

We remark in passing that if the set $\mathcal{X}$ is a product space, meaning that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ for some $n < \infty$ (this is the case for mutual information, for example), then we may assume our quantizer *always* quantizes sets of the form $A = A_1 \times A_2 \times \cdots \times A_n$, that is, Cartesian products. Written differently, when we consider algebras on $\mathcal{X}$, the atoms of the algebra may be assumed to be Cartesian products of sets, and our partitions of $\mathcal{X}$ can always be taken as Cartesian products. (See Gray [94, Chapter 5].) Written slightly differently, if P and Q are distributions on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and q^i is a quantizer for the set $\mathcal{X}_i$ (inducing the partition $A_1^i, \ldots, A_{m_i}^i$ of $\mathcal{X}_i$) we may define

$$D_{\rm kl}\left(P\|Q \mid \mathsf{q}^1, \ldots, \mathsf{q}^n\right) = \sum_{j_1, \ldots, j_n} P(A_{j_1}^1 \times A_{j_2}^2 \times \cdots \times A_{j_n}^n) \log \frac{P(A_{j_1}^1 \times A_{j_2}^2 \times \cdots \times A_{j_n}^n)}{Q(A_{j_1}^1 \times A_{j_2}^2 \times \cdots \times A_{j_n}^n)}.$$

Then the general definition (2.2.1) of KL-divergence specializes to

$$D_{\rm kl}\left(P\|Q\right) = \sup\left\{D_{\rm kl}\left(P\|Q \mid \mathsf{q}^1, \ldots, \mathsf{q}^n\right) \quad \text{such that } \mathsf{q}^i \text{ quantizes } \mathcal{X}_i\right\}.$$

So we only need consider "rectangular" sets in the definitions of KL-divergence.

Measure-theoretic definition of KL-divergence If you have never seen measure theory before, skim this section; while the notation may be somewhat intimidating, it is fine to always consider only continuous or fully discrete distributions. We will describe an interpretation that will mean for our purposes that one never needs to really think about measure theoretic issues.

The general definition (2.2.1) of KL-divergence is equivalent to the following. Let μ be a measure on $\mathcal{X}$, and assume that P and Q are absolutely continuous with respect to μ, with densities p and q, respectively. (For example, take $\mu = P + Q$.) Then

$$D_{\rm kl}\left(P\|Q\right) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x). \tag{2.2.2}$$

The proof of this fact is somewhat involved, requiring the technology of Lebesgue integration. (See Gray [94, Chapter 5].)

For those who have not seen measure theory, the interpretation of the equality (2.2.2) should be as follows. When integrating a function $f(x)$, replace $\int f(x) d\mu(x)$ with one of two pairs of symbols: one may simply think of $d\mu(x)$ as dx, so that we are performing standard integration $\int f(x) dx$, or one should think of the integral operation $\int f(x) d\mu(x)$ as summing the argument of the integral, so $d\mu(x) = 1$ and $\int f(x) d\mu(x) = \sum_x f(x)$. (This corresponds to μ being "counting measure" on $\mathcal{X}$.)

2.2.3 f-divergences

A more general notion of divergence is the so-called f-divergence, or Ali-Silvey divergence [4, 54] (see also the alternate interpretations in the article by Liese and Vajda [131]). Here, the definition is as follows. Let P and Q be probability distributions on the set $\mathcal{X}$, and let $f : \mathbb{R}_+ \to \mathbb{R}$ be a

convex function satisfying $f(1) = 0$. If $\mathcal{X}$ is a discrete set, then the *f-divergence* between P and Q is

$$D_f\left(P\|Q\right) := \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right).$$

More generally, for any set $\mathcal{X}$ and a quantizer $\mathsf{q} : \mathcal{X} \to \{1, \ldots, m\}$, letting $A_i = \mathsf{q}^{-1}(\{i\}) = \{x \in \mathcal{X} \mid \mathsf{q}(x) = i\}$ be the partition the quantizer induces, we can define the quantized divergence

$$D_f\left(P\|Q \mid \mathsf{q}\right) = \sum_{i=1}^m Q(A_i) f\left(\frac{P(A_i)}{Q(A_i)}\right),$$

and the general definition of an f divergence is (in analogy with the definition (2.2.1) of general KL divergences)

$$D_f\left(P\|Q\right) := \sup\left\{D_f\left(P\|Q \mid \mathsf{q}\right) \quad \text{such that } \mathsf{q} \text{ quantizes } \mathcal{X}\right\}. \tag{2.2.3}$$

The definition (2.2.3) shows that, any time we have computations involving f-divergences—such as KL-divergence or mutual information—it is no loss of generality, when performing the computations, to assume that all distributions have finite discrete support. There is a measure-theoretic version of the definition (2.2.3) which is frequently easier to use. Assume w.l.o.g. that P and Q are absolutely continuous with respect to the base measure μ. The f divergence between P and Q is then

$$D_f\left(P\|Q\right) := \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) d\mu(x). \tag{2.2.4}$$

This definition, it turns out, is not *quite* as general as we would like—in particular, it is unclear how we should define the integral for points x such that $q(x) = 0$. With that in mind, we recall that the perspective transform (see Appendices B.1.1 and B.3.3) of a function $f : \mathbb{R} \to \mathbb{R}$ is defined by $\operatorname{pers}(f)(t, u) = uf(t/u)$ if $u > 0$ and by $+\infty$ if $u \le 0$. This function is convex in its arguments (Proposition B.3.12). In fact, this is not quite enough for the fully correct definition. The *closure* of a convex function f is $\operatorname{cl} f(x) = \sup\{\ell(x) \mid \ell \le f, \ \ell \text{ linear}\}$, the supremum over all linear functions that globally lower bound f. Then [104, Proposition IV.2.2.2] the closer of $\operatorname{pers}(f)$ is defined, for any $t' \in \operatorname{int}\operatorname{dom} f$, by

$$\operatorname{cl}\operatorname{pers}(f)(t, u) = \begin{cases} uf(t/u) & \text{if } u > 0 \\ \lim_{\alpha \downarrow 0} \alpha f(t' - t + t/\alpha) & \text{if } u = 0 \\ +\infty & \text{if } u < 0. \end{cases}$$

(The choice of t' does not affect the definition.) Then the fully general formula expressing the f-divergence is

$$D_f\left(P\|Q\right) = \int_{\mathcal{X}} \operatorname{cl}\operatorname{pers}(f)(p(x), q(x)) d\mu(x). \tag{2.2.5}$$

This is what we mean by equation (2.2.4), which we use without comment.

In the exercises, we explore several properties of f-divergences, including the quantized representation (2.2.3), showing different data processing inequalities and orderings of quantizers based on the fineness of their induced partitions. Broadly, f-divergences satisfy essentially the same properties as KL-divergence, such as data-processing inequalities, and they provide a generalization of mutual information. We explore f-divergences from additional perspectives later—they are important both for optimality in estimation and related to consistency and prediction problems, as we discuss in Chapter 14.

Examples We give several examples of f-divergences here; in Section 8.2.2 we provide a few examples of their uses as well as providing a few natural inequalities between them.

Example 2.2.1 (KL-divergence)**:** By taking $f(t) = t \log t$, which is convex and satisfies $f(1) = 0$, we obtain $D_f(P\|Q) = D_{\rm kl}(P\|Q)$. $\diamond$

Example 2.2.2 (KL-divergence, reversed)**:** By taking $f(t) = -\log t$, we obtain $D_f(P\|Q) = D_{\rm kl}(Q\|P)$. $\diamond$

Example 2.2.3 (Total variation distance)**:** The total variation distance between probability distributions P and Q defined on a set $\mathcal{X}$ is the maximum difference between probabilities they assign on subsets of $\mathcal{X}$:

$$\|P - Q\|_{\rm TV} := \sup_{A \subset \mathcal{X}} |P(A) - Q(A)| = \sup_{A \subset \mathcal{X}} (P(A) - Q(A)), \tag{2.2.6}$$

where the second equality follows by considering compliments $P(A^c) = 1 - P(A)$. The total variation distance, as we shall see later, is important for verifying the optimality of different tests, and appears in the measurement of difficulty of solving hypothesis testing problems. The choice $f(t) = \frac{1}{2}|t - 1|$, we obtain the total variation distance, that is, $\|P - Q\|_{\rm TV} = D_f(P\|Q)$. There are several alternative characterizations, which we provide as Lemma 2.2.4 next; it will be useful in the sequel when we develop inequalities relating the divergences. $\diamond$

Lemma 2.2.4. *Let P, Q be probability measures with densities p, q with respect to a base measure μ and $f(t) = \frac{1}{2}|t - 1|$. Then*

$$\begin{aligned}
\|P - Q\|_{\rm TV} = D_f(P\|Q) &= \frac{1}{2} \int |p(x) - q(x)| d\mu(x) \\
&= \int [p(x) - q(x)]_+ \, d\mu(x) = \int [q(x) - p(x)]_+ \, d\mu(x) \\
&= P(dP/dQ > 1) - Q(dP/dQ > 1) = Q(dQ/dP > 1) - P(dQ/dP > 1).
\end{aligned}$$

In particular, the set $A = \{x \mid p(x)/q(x) \geq 1\}$ maximizes $P(B) - Q(B)$ over $B \subset \mathcal{X}$ and so achieves $\|P - Q\|_{\rm TV} = P(A) - Q(A)$.

Proof Eliding the measure-theoretic details,[2] we immediately have

$$\begin{aligned}
D_f(P\|Q) &= \frac{1}{2} \int \left| \frac{p(x)}{q(x)} - 1 \right| q(x) d\mu(x) = \frac{1}{2} \int |p(x) - q(x)| d\mu(x) \\
&= \frac{1}{2} \int_{x: p(x) > q(x)} [p(x) - q(x)] \, d\mu(x) + \frac{1}{2} \int_{x: q(x) > p(x)} [q(x) - p(x)] \, d\mu(x) \\
&= \frac{1}{2} \int [p(x) - q(x)]_+ \, d\mu(x) + \frac{1}{2} \int [q(x) - p(x)]_+ \, d\mu(x).
\end{aligned}$$

Considering the last inegral $\int [q(x) - p(x)]_+ \, d\mu(x)$, we see that the set $A = \{x : q(x) > p(x)\}$ satisfies

$$Q(A) - P(A) = \int_A (q(x) - p(x)) d\mu(x) \geq \int_B (q(x) - p(x)) d\mu(x) = Q(B) - P(B)$$

[2] To make this fully rigorous, we would use the Hahn decomposition of the signed measure $P - Q$ to recognize that $\int f(dP - dQ) = \int f [dP - dQ]_+ - \int f [dQ - dP]_+$ for any integrable f.

for any set B, as any $x \in B \setminus A$ clearly satisfies $q(x) - p(x) \le 0$. $\square$

Example 2.2.5 (Hellinger distance)**:** The *Hellinger distance* between probability distributions P and Q defined on a set $\mathcal{X}$ is generated by the function $f(t) = (\sqrt{t} - 1)^2 = t - 2\sqrt{t} + 1$. The Hellinger distance is then

$$d_{\rm hel}(P, Q)^2 := \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x). \tag{2.2.7}$$

The non-squared version $d_{\rm hel}(P, Q)$ is indeed a distance between probability measures P and Q. It is sometimes convenient to rewrite the Hellinger distance in terms of the *affinity* between P and Q, as

$$d_{\rm hel}(P, Q)^2 = \frac{1}{2} \int (p(x) + q(x) - 2\sqrt{p(x)q(x)}) d\mu(x) = 1 - \int \sqrt{p(x)q(x)} d\mu(x), \tag{2.2.8}$$

which makes clear that $d_{\rm hel}(P, Q) \in [0, 1]$ is on roughly the same scale as the variation distance; we will say more later. $\diamond$

Example 2.2.6 (χ^2 divergence)**:** The *χ^2-divergence* is generated by taking $f(t) = (t - 1)^2$, so that

$$D_{\chi^2}(P \| Q) := \int \left(\frac{p(x)}{q(x)} - 1 \right)^2 q(x) d\mu(x) = \int \frac{p(x)^2}{q(x)} d\mu(x) - 1, \tag{2.2.9}$$

where the equality is immediate because $\int p d\mu = \int q d\mu = 1$. $\diamond$

2.2.4 Inequalities and relationships between divergences

Important to our development will come will be different families of inequalities relating the different divergence measures. These inequalities will be particularly important because, in some cases, different distributions admit easy calculations with some divergences, such as KL or χ^2 divergence, but it can be challenging to work with others that may be more "natural" for a particular problem. Most importantly, replacing a variation distance by bounding it with an alternative divergence is often convenient for analyzing the properties of product distributions (as will become apparent in Chapter 8). We record several of these results here, making a passing connection to mutual information as well.

The first inequality shows that the Hellinger distance and variation distance roughly generate the same topology on collections of distributions, as they upper and lower bound the other (if we tolerate polynomial losses).

Proposition 2.2.7. *The total variation distance and Hellinger distance satisfy*

$$d_{\rm hel}^2(P, Q) \le \|P - Q\|_{\rm TV} \le d_{\rm hel}(P, Q) \sqrt{2 - d_{\rm hel}^2(P, Q)}.$$

Proof We begin with the upper bound. We have by Hölder's inequality that

$$\begin{aligned}
\frac{1}{2} \int |p(x) - q(x)| d\mu(x) &= \int |\sqrt{p(x)} - \sqrt{q(x)}| \cdot |\sqrt{p(x)} + \sqrt{q(x)}| d\mu(x) \\
&\le \left(\frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x) \right)^{\frac{1}{2}} \left(\frac{1}{2} \int (\sqrt{p(x)} + \sqrt{q(x)})^2 d\mu(x) \right)^{\frac{1}{2}} \\
&= d_{\rm hel}(P, Q) \left(1 + \int \sqrt{p(x)q(x)} d\mu(x) \right)^{\frac{1}{2}}.
\end{aligned}$$

As in Example 2.2.5, we have $\int \sqrt{p(x)q(x)}d\mu(x) = 1 - d_{\rm hel}(P,Q)^2$, so this (along with the representation Lemma 2.2.4 for variation distance) implies

$$\|P - Q\|_{\rm TV} = \frac{1}{2}\int |p(x) - q(x)|d\mu(x) \le d_{\rm hel}(P,Q)(2 - d_{\rm hel}^2(P,Q))^{\frac{1}{2}}.$$

For the lower bound on total variation, note that for any $a, b \in \mathbb{R}_+$, we have $a + b - 2\sqrt{ab} \le |a - b|$ (check the cases $a > b$ and $a < b$ separately); thus

$$d_{\rm hel}^2(P,Q) = \frac{1}{2}\int \left[p(x) + q(x) - 2\sqrt{p(x)q(x)}\right] d\mu(x) \le \frac{1}{2}\int |p(x) - q(x)|d\mu(x),$$

as desired. □

Several important inequalitites relate the variation distance to the KL-divergence. We state two important inequalities in the next proposition, both of which are important enough to justify their own names.

Proposition 2.2.8. *The total variation distance satisfies the following relationships.*

(a) Pinsker's inequality*: for any distributions P and Q,*

$$\|P - Q\|_{\rm TV}^2 \le \frac{1}{2} D_{\rm kl}\left(P\|Q\right). \tag{2.2.10}$$

(b) The Bretagnolle-Huber *inequality: for any distributions P and Q,*

$$\|P - Q\|_{\rm TV} \le \sqrt{1 - \exp(-D_{\rm kl}\left(P\|Q\right))} \le 1 - \frac{1}{2}\exp(-D_{\rm kl}\left(P\|Q\right)).$$

Proof Exercise 2.19 outlines one proof of Pinsker's inequality using the data processing inequality (Proposition 2.2.13). We present an alternative via the Cauchy-Schwarz inequality. Using the definition (2.2.1) of the KL-divergence, we may assume without loss of generality that P and Q are finitely supported, say with p.m.f.s $p_1, \ldots, p_m$ and $q_1, \ldots, q_m$. Define the negative entropy function $h(p) = \sum_{i=1}^m p_i \log p_i$. Then showing that $D_{\rm kl}\left(P\|Q\right) \ge 2\|P - Q\|_{\rm TV}^2 = \frac{1}{2}\|p - q\|_1^2$ is equivalent to showing that

$$h(p) \ge h(q) + \langle \nabla h(q), p - q\rangle + \frac{1}{2}\|p - q\|_1^2, \tag{2.2.11}$$

because by inspection $h(p) - h(q) - \langle \nabla h(q), p - q\rangle = \sum_i p_i \log \frac{p_i}{q_i}$. We do this via a Taylor expansion: we have

$$\nabla h(p) = [\log p_i + 1]_{i=1}^m \quad \text{and} \quad \nabla^2 h(p) = \operatorname{diag}([1/p_i]_{i=1}^m).$$

By Taylor's theorem, there is some $\tilde{p} = (1 - t)p + tq$, where $t \in [0, 1]$, such that

$$h(p) = h(q) + \langle \nabla h(q), p - q\rangle + \frac{1}{2}\langle p - q, \nabla^2 h(\tilde{p})(p - q)\rangle.$$

But looking at the final quadratic, we have for any vector v and any $p \ge 0$ satisfying $\sum_i p_i = 1$,

$$\langle v, \nabla^2 h(\tilde{p})v\rangle = \sum_{i=1}^m \frac{v_i^2}{p_i} = \|p\|_1 \sum_{i=1}^m \frac{v_i^2}{p_i} \ge \left(\sum_{i=1}^m \sqrt{p_i}\frac{|v_i|}{\sqrt{p_i}}\right)^2 = \|v\|_1^2,$$

where the inequality follows from Cauchy-Schwarz applied to the vectors $[\sqrt{p_i}]_i$ and $[|v_i|/\sqrt{p_i}]_i$. Thus inequality (2.2.11) holds.

For the claim (b), we use Proposition 2.2.7. Let $a = \int \sqrt{p(x)q(x)}d\mu(x)$ be a shorthand for the affinity, so that $d_{\rm hel}^2(P,Q) = 1 - a$. Then Proposition 2.2.7 gives $\|P - Q\|_{\rm TV} \le \sqrt{1-a}\sqrt{1+a} = \sqrt{1-a^2}$. Now apply Jensen's inequality to the exponential: we have

$$\int \sqrt{p(x)q(x)}d\mu(x) = \int \sqrt{\frac{q(x)}{p(x)}}p(x)d\mu(x) = \int \exp\left(\frac{1}{2}\log\frac{q(x)}{p(x)}\right)p(x)d\mu(x)$$
$$\ge \exp\left(\frac{1}{2}\int p(x)\log\frac{q(x)}{p(x)}d\mu(x)\right) = \exp\left(-\frac{1}{2}D_{\rm kl}\left(P\|Q\right)\right).$$

In particular, $\sqrt{1-a^2} \le \sqrt{1 - \exp(-\frac{1}{2}D_{\rm kl}\left(P\|Q\right))^2}$, which is the first claim of part (b). For the second, note that $\sqrt{1-c} \le 1 - \frac{1}{2}c$ for $c \in [0,1]$ by concavity of the square root. □

We also have the following bounds on the KL-divergence in terms of the χ^2-divergence.

Proposition 2.2.9. *For any distributions P, Q,*

$$D_{\rm kl}\left(P\|Q\right) \le \log(1 + D_{\chi^2}\left(P\|Q\right)) \le D_{\chi^2}\left(P\|Q\right).$$

Proof By Jensen's inequality, we have

$$D_{\rm kl}\left(P\|Q\right) \le \log\int\frac{dP^2}{dQ} = \log(1 + D_{\chi^2}\left(P\|Q\right)).$$

The second inequality is immediate as $\log(1+t) \le t$ for all $t > -1$. □

It is also possible to relate mutual information between distributions to f-divergences, and even to bound the mutual information above and below by the Hellinger distance for certain problems. In this case, we consider the following situation: let $V \in \{0,1\}$ uniformly at random, and conditional on $V = v$, draw $X \sim P_v$ for some distribution P_v on a space $\mathcal{X}$. Then we have that

$$I(X;V) = \frac{1}{2}D_{\rm kl}\left(P_0\|\overline{P}\right) + \frac{1}{2}D_{\rm kl}\left(P_1\|\overline{P}\right)$$

where $\overline{P} = \frac{1}{2}P_0 + \frac{1}{2}P_1$. The divergence measure on the right side of the preceding identity is a special case of the *Jenson-Shannon divergence*, defined for $\lambda \in [0,1]$ by

$$D_{{\rm js},\lambda}\left(P\|Q\right) := \lambda D_{\rm kl}\left(P\|\lambda P + (1-\lambda)Q\right) + D_{\rm kl}\left(Q\|\lambda P + (1-\lambda)Q\right), \tag{2.2.12}$$

which is a symmetrized and bounded variant of the typical KL-divergence (we use the shorthand $D_{\rm js}\left(P\|Q\right) := D_{{\rm js},\frac{1}{2}}\left(P\|Q\right)$ for the symmetric case). As a consequence, we also have

$$I(X;V) = \frac{1}{2}D_f\left(P_0\|P_1\right) + \frac{1}{2}D_f\left(P_1\|P_0\right),$$

where $f(t) = -t\log(\frac{1}{2t} + \frac{1}{2}) = t\log\frac{2t}{t+1}$, so that the mutual information is a particular f-divergence. This form—as we see in the later chapters—is frequently convenient because it gives an object with similar tensorization properties to KL-divergence while enjoying the boundedness properties of Hellinger and variation distances. The following proposition captures the latter properties.

Proposition 2.2.10. *Let (X, V) be distributed as above. Then*

$$\log 2 \cdot d_{\rm hel}^2(P_0, P_1) \le I(X; V) = D_{\rm js}\left(P_0 \| P_1\right) \le \min \left\{ \begin{array}{c} \log 2 \cdot \|P_0 - P_1\|_{\rm TV}, \\ 2 \cdot d_{\rm hel}^2(P_0, P_1) \end{array} \right\}.$$

Proof The lower bound and upper bound involving the variation distance both follow from analytic bounds on the binary entropy functional $h_2(p) = -p \log p - (1-p) \log(1-p)$. By expanding the mutual information and letting p_0 and p_1 be densities of P_0 and P_1 with respect to some base measure μ, we have

$$\begin{aligned} 2I(X; V) = 2D_{\rm js}\left(P_0 \| P_1\right) &= \int p_0 \log \frac{2p_0}{p_0 + p_1} d\mu + \int p_1 \log \frac{2p_1}{p_0 + p_1} d\mu \\ &= 2 \log 2 + \int (p_0 + p_1) \left[\frac{p_0}{p_1 + p_1} \log \frac{p_0}{p_0 + p_1} + \frac{p_1}{p_1 + p_1} \log \frac{p_1}{p_0 + p_1} \right] d\mu \\ &= 2 \log 2 - \int (p_0 + p_1) h_2 \left(\frac{p_0}{p_1 + p_0} \right) d\mu. \end{aligned}$$

We claim that

$$2 \log 2 \cdot \min\{p, 1 - p\} \le h_2(p) \le 2 \log 2 \cdot \sqrt{p(1-p)}$$

for all $p \in [0, 1]$ (see Exercises 2.17 and 2.18). Then the upper and lower bounds on the information become nearly immediate.

For the variation-based upper bound on $I(X; V)$, we use the lower bound $h_2(p) \ge 2 \log 2 \cdot \min\{p, 1 - p\}$ to write

$$\begin{aligned} \frac{2}{\log 2} I(X; V) &\le 2 - \int (p_0(x) + p_1(x)) \min \left\{ \frac{p_0(x)}{p_0(x) + p_1(x)}, \frac{p_1(x)}{p_0(x) + p_1(x)} \right\} d\mu(x) \\ &= 2 - 2 \int \min\{p_0(x), p_1(x)\} d\mu(x) \\ &= 2 \int (p_1(x) - \min\{p_0(x), p_1(x)\}) d\mu(x) = 2 \int_{p_1 > p_0} (p_1(x) - p_0(x)) d\mu(x). \end{aligned}$$

But of course the final integral is $\|P_1 - P_0\|_{\rm TV}$, giving $I(X; V) \le \log 2 \, \|P_0 - P_1\|_{\rm TV}$. Conversely, for the lower bound on $D_{\rm js}\left(P_0 \| P_1\right)$, we use the upper bound $h_2(p) \le 2 \log 2 \cdot \sqrt{p(1-p)}$ to obtain

$$\begin{aligned} \frac{1}{\log 2} I(X; V) &\ge 1 - \int (p_0 + p_1) \sqrt{\frac{p_0}{p_1 + p_0} \left(1 - \frac{p_0}{p_1 + p_0}\right)} d\mu \\ &= 1 - \int \sqrt{p_0 p_1} d\mu = \frac{1}{2} \int (\sqrt{p_0} - \sqrt{p_1})^2 d\mu = d_{\rm hel}^2(P_0, P_1) \end{aligned}$$

as desired.

The Hellinger-based upper bound is simpler: by Proposition 2.2.9, we have

$$\begin{aligned} D_{\rm js}\left(P_0 \| P_1\right) &= \frac{1}{2} D_{\rm kl}\left(P_0 \| (P_0 + P_1)/2\right) + \frac{1}{2} D_{\rm kl}\left(P_1 \| (P_0 + P_1)/2\right) \\ &\le \frac{1}{2} D_{\chi^2}\left(P_0 \| (P_0 + P_1)/2\right) + \frac{1}{2} D_{\chi^2}\left(P_1 \| (P_0 + P_1)/2\right) \\ &= \frac{1}{2} \int \frac{(p_0 - p_1)^2}{p_0 + p_1} d\mu = \frac{1}{2} \int \frac{(\sqrt{p_0} - \sqrt{p_1})^2 (\sqrt{p_0} + \sqrt{p_1})^2}{p_0 + p_1} d\mu. \end{aligned}$$

Now note that $(a + b)^2 \le 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, and so $(\sqrt{p_0} + \sqrt{p_1})^2 \le 2(p_0 + p_1)$, and thus the final integral has bound $\int (\sqrt{p_0} - \sqrt{p_1})^2 d\mu = 2 d_{\rm hel}^2(P_0, P_1)$. $\square$

2.2.5 Convexity and data processing for divergence measures

f-divergences satisfy a number of very useful properties, which we use repeatedly throughout the lectures. As the KL-divergence is an f-divergence, it of course satisfies these conditions; however, we state them in fuller generality, treating the KL-divergence results as special cases and corollaries.

We begin by exhibiting the general data processing properties and convexity properties of f-divergences, each of which specializes to KL divergence. We leave the proof of each of these as exercises. First, we show that f-divergences are jointly convex in their arguments.

Proposition 2.2.11. *Let P_1, P_2, Q_1, Q_2 be distributions on a set $\mathcal{X}$ and $f : \mathbb{R}_+ \to \mathbb{R}$ be convex. Then for any $\lambda \in [0, 1]$,*

$$D_f\left(\lambda P_1 + (1-\lambda)P_2 \| \lambda Q_1 + (1-\lambda)Q_2\right) \le \lambda D_f\left(P_1 \| Q_1\right) + (1-\lambda) D_f\left(P_2 \| Q_2\right).$$

The proof of this proposition we leave as Exercise 2.11, which we treat as a consequence of the more general "log-sum" like inequalities of Exercise 2.8. It is, however, an immediate consequence of the fully specified definition (2.2.5) of an f-divergence, because $\mathrm{pers}(f)$ is jointly convex. As an immediate corollary, we see that the same result is true for KL-divergence as well.

Corollary 2.2.12. *The KL-divergence $D_{\rm kl}\left(P\|Q\right)$ is jointly convex in its arguments P and Q.*

We can also provide more general data processing inequalities for f-divergences, paralleling those for the KL-divergence. In this case, we consider random variables X and Z on spaces $\mathcal{X}$ and $\mathcal{Z}$, respectively, and a Markov transition kernel K giving the Markov chain $X \to Z$. That is, $K(\cdot \mid x)$ is a probability distribution on $\mathcal{Z}$ for each $x \in \mathcal{X}$, and conditioned on $X = x$, Z has distribution $K(\cdot \mid x)$ so that $K(A \mid x) = \mathbb{P}(Z \in A \mid X = x)$. Certainly, this includes the situation when $Z = \phi(X)$ for some function ϕ, and more generally when $Z = \phi(X, U)$ for a function ϕ and some additional randomness U. For a distribution P on X, we then define the marginals

$$K_P(A) := \int_{\mathcal{X}} K(A, x) dP(x).$$

We then have the following proposition.

Proposition 2.2.13. *Let P and Q be distributions on X and let K be any Markov kernel. Then*

$$D_f\left(K_P \| K_Q\right) \le D_f\left(P\|Q\right).$$

See Exercise 2.10 for a proof.

As a corollary, we obtain the following data processing inequality for KL-divergences, where we abuse notation to write $D_{\rm kl}\left(X\|Y\right) = D_{\rm kl}\left(P\|Q\right)$ for random variables $X \sim P$ and $Y \sim Q$.

Corollary 2.2.14. *Let $X, Y \in \mathcal{X}$ be random variables, let $U \in \mathcal{U}$ be independent of X and Y, and let $\phi : \mathcal{X} \times \mathcal{U} \to \mathcal{Z}$ for some spaces $\mathcal{X}$,$\mathcal{U}$, $\mathcal{Z}$. Then*

$$D_{\rm kl}\left(\phi(X, U) \| \phi(Y, U)\right) \le D_{\rm kl}\left(X\|Y\right).$$

Thus, further processing of random variables can only bring them "closer" in the space of distributions; downstream processing of signals cannot make them further apart as distributions.

2.3 First steps into optimal procedures: testing inequalities

As noted in the introduction, a central benefit of the information theoretic tools we explore is that they allow us to certify the optimality of procedures—that no other procedure could (substantially) improve upon the one at hand. The main tools for these certifications are often inequalities governing the best possible behavior of a variety of statistical tests. Roughly, we put ourselves in the following scenario: nature chooses one of a possible set of (say) k worlds, indexed by probability distributions $P_1, P_2, \ldots, P_k$, and conditional on nature's choice of the world—the distribution $P^\star \in \{P_1, \ldots, P_k\}$ chosen—we observe data X drawn from $P^\star$. Intuitively, it will be difficult to decide which distribution P_i is the true $P^\star$ if all the distributions are similar—the divergence between the P_i is small, or the information between X and $P^\star$ is negligible—and easy if the distances between the distributions P_i are large. With this outline in mind, we present two inequalities, and first examples of their application, to make concrete these connections to the notions of information and divergence defined in this section.

2.3.1 Le Cam's inequality and binary hypothesis testing

The simplest instantiation of the above setting is the case when there are only two possible distributions, P_1 and P_2, and our goal is to make a decision on whether P_1 or P_2 is the distribution generating data we observe. Concretely, suppose that nature chooses one of the distributions P_1 or P_2 at random, and let $V \in \{1, 2\}$ index this choice. Conditional on $V = v$, we then observe a sample X drawn from P_v. Denoting by $\mathbb{P}$ the joint distribution of V and X, we have for any test $\Psi : \mathcal{X} \to \{1, 2\}$ that the probability of error is then

$$\mathbb{P}(\Psi(X) \neq V) = \frac{1}{2} P_1(\Psi(X) \neq 1) + \frac{1}{2} P_2(\Psi(X) \neq 2).$$

We can give an exact expression for the minimal possible error in the above hypothesis test. Indeed, a standard result of Le Cam (see [127, 177, Lemma 1]) is the following variational representation of the total variation distance (2.2.6), which is the f-divergence associated with $f(t) = \frac{1}{2}|t-1|$, as a function of testing error.

Proposition 2.3.1. *Let $\mathcal{X}$ be an arbitrary set. For any distributions P_1 and P_2 on $\mathcal{X}$, we have*

$$\inf_\Psi \{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\} = 1 - \|P_1 - P_2\|_{\rm TV},$$

where the infimum is taken over all tests $\Psi : \mathcal{X} \to \{1, 2\}$.

Proof Any test $\Psi : \mathcal{X} \to \{1, 2\}$ has an acceptance region, call it $A \subset \mathcal{X}$, where it outputs 1 and a region A^c where it outputs 2.

$$P_1(\Psi \neq 1) + P_2(\Psi \neq 2) = P_1(A^c) + P_2(A) = 1 - P_1(A) + P_2(A).$$

Taking an infimum over such acceptance regions, we have

$$\inf_\Psi \{P_1(\Psi \neq 1) + P_2(\Psi \neq 2)\} = \inf_{A \subset \mathcal{X}} \{1 - (P_1(A) - P_2(A))\} = 1 - \sup_{A \subset \mathcal{X}} (P_1(A) - P_2(A)),$$

which yields the total variation distance as desired. □

In the two-hypothesis case, we also know that the optimal test, by the Neyman-Pearson lemma, is a likelihood ratio test. That is, assuming that P_1 and P_2 have densities p_1 and p_2, the optimal test is of the form

$$\Psi(X) = \begin{cases} 1 & \text{if } \frac{p_1(X)}{p_2(X)} \ge t \\ 2 & \text{if } \frac{p_1(X)}{p_2(X)} < t \end{cases}$$

for some threshold $t \ge 0$. In the case that the prior probabilities on P_1 and P_2 are each $\frac{1}{2}$, then $t = 1$ is optimal.

We give one example application of Proposition 2.3.1 to the problem of testing a normal mean.

Example 2.3.2 (Testing a normal mean)**:** Suppose we observe $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P$ for $P = P_1$ or $P = P_2$, where P_v is the normal distribution $\mathsf{N}(\mu_v, \sigma^2)$, where $\mu_1 \neq \mu_2$. We would like to understand the sample size n necessary to guarantee that no test can have small error, that is, say, that

$$\inf_{\Psi} \left\{P_1(\Psi(X_1, \ldots, X_n) \neq 1) + P_2(\Psi(X_1, \ldots, X_n) \neq 2)\right\} \ge \frac{1}{2}.$$

By Proposition 2.3.1, we have that

$$\inf_{\Psi} \left\{P_1(\Psi(X_1, \ldots, X_n) \neq 1) + P_2(\Psi(X_1, \ldots, X_n) \neq 2)\right\} \ge 1 - \|P_1^n - P_2^n\|_{\text{TV}},$$

where P_v^n denotes the n-fold product of P_v, that is, the distribution of $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P_v$. The interaction between total variation distance and product distributions is somewhat subtle, so it is often advisable to use a divergence measure more attuned to the i.i.d. nature of the sampling scheme. Two such measures are the KL-divergence and Hellinger distance, both of which we explore in the coming chapters. With that in mind, we apply Pinsker's inequality (2.2.10) to see that $\|P_1^n - P_2^n\|_{\text{TV}}^2 \le \frac{1}{2} D_{\text{kl}}\left(P_1^n \| P_2^n\right) = \frac{n}{2} D_{\text{kl}}\left(P_1 \| P_2\right)$, which implies that

$$1 - \|P_1^n - P_2^n\|_{\text{TV}} \ge 1 - \sqrt{\frac{n}{2} D_{\text{kl}}\left(P_1 \| P_2\right)^{\frac{1}{2}}} = 1 - \sqrt{\frac{n}{2}\left(\frac{1}{2\sigma^2}(\mu_1 - \mu_2)^2\right)^{\frac{1}{2}}} = 1 - \frac{\sqrt{n}}{2}\frac{|\mu_1 - \mu_2|}{\sigma}.$$

In particular, if $n \le \frac{\sigma^2}{(\mu_1 - \mu_2)^2}$, then we have our desired lower bound of $\frac{1}{2}$.

Conversely, a calculation yields that $n \ge \frac{C\sigma^2}{(\mu_1 - \mu_2)^2}$, for some numerical constant $C \ge 1$, implies small probability of error. We leave this calculation to the reader. $\diamond$

2.3.2 Fano's inequality and multiple hypothesis testing

There are of course situations in which we do not wish to simply test two hypotheses, but have multiple hypotheses present. In such situations, Fano's inequality, which we present shortly, is the most common tool for proving fundamental limits, lower bounds on probability of error, and converses (to results on achievability of some performance level) in information theroy. We write this section in terms of general random variables, ignoring the precise setting of selecting an index in a family of distributions, though that is implicit in what we do.

Let X be a random variable taking values in a finite set $\mathcal{X}$, and assume that we observe a (different) random variable Y, and then must estimate or guess the true value of X. That is, we have the Markov chain

$$X \to Y \to \widehat{X},$$

and we wish to provide lower bounds on the probability of error—that is, that $\widehat{X} \neq X$. If we let the function $h_2(p) = -p \log p - (1-p) \log(1-p)$ denote the binary entropy (entropy of a Bernoulli random variable with parameter p), Fano's inequality takes the following form [e.g. 53, Chapter 2]:

Proposition 2.3.3 (Fano inequality)**.** *For any Markov chain $X \to Y \to \widehat{X}$, we have*

$$h_2(\mathbb{P}(\widehat{X} \neq X)) + \mathbb{P}(\widehat{X} \neq X) \log(|\mathcal{X}| - 1) \geq H(X \mid \widehat{X}). \tag{2.3.1}$$

Proof This proof follows by expanding an entropy functional in two different ways. Let E be the indicator for the event that $\widehat{X} \neq X$, that is, $E = 1$ if $\widehat{X} \neq X$ and is 0 otherwise. Then we have

$$\begin{aligned} H(X, E \mid \widehat{X}) &= H(X \mid E, \widehat{X}) + H(E \mid \widehat{X}) \\ &= \mathbb{P}(E = 1) H(X \mid E = 1, \widehat{X}) + \mathbb{P}(E = 0) \underbrace{H(X \mid E = 0, \widehat{X})}_{=0} + H(E \mid \widehat{X}), \end{aligned}$$

where the zero follows because given there is no error, X has no variability given $\widehat{X}$. Expanding the entropy by the chain rule in a different order, we have

$$H(X, E \mid \widehat{X}) = H(X \mid \widehat{X}) + \underbrace{H(E \mid \widehat{X}, X)}_{=0},$$

because E is perfectly predicted by $\widehat{X}$ and X. Combining these equalities, we have

$$H(X \mid \widehat{X}) = H(X, E \mid \widehat{X}) = \mathbb{P}(E = 1) H(X \mid E = 1, \widehat{X}) + H(E \mid X).$$

Noting that $H(E \mid X) \leq H(E) = h_2(\mathbb{P}(E = 1))$, as conditioning reduces entropy, and that $H(X \mid E = 1, \widehat{X}) \leq \log(|\mathcal{X}| - 1)$, as X can take on at most $|\mathcal{X}| - 1$ values when there is an error, completes the proof. □

We can rewrite Proposition 2.3.3 in a convenient way when X is uniform in $\mathcal{X}$. Indeed, by definition of the mutual information, we have $I(X; \widehat{X}) = H(X) - H(X \mid \widehat{X})$, so Proposition 8.4.1 implies that in the canonical hypothesis testing problem from Section 8.2.1, we have

Corollary 2.3.4. *Assume that X is uniform on $\mathcal{X}$. For any Markov chain $X \to Y \to \widehat{X}$,*

$$\mathbb{P}(\widehat{X} \neq X) \geq 1 - \frac{I(X; Y) + \log 2}{\log(|\mathcal{X}|)}. \tag{2.3.2}$$

Proof Let $P_{\mathsf{error}} = \mathbb{P}(X \neq \widehat{X})$ denote the probability of error. Noting that $h_2(p) \leq \log 2$ for any $p \in [0, 1]$ (recall inequality (2.1.2), that is, that uniform random variables maximize entropy), then using Proposition 8.4.1, we have

$$\log 2 + P_{\mathsf{error}} \log(|\mathcal{X}|) \geq h_2(P_{\mathsf{error}}) + P_{\mathsf{error}} \log(|\mathcal{X}| - 1) \overset{(i)}{\geq} H(X \mid \widehat{X}) \overset{(ii)}{=} H(X) - I(X; \widehat{X}).$$

Here step (i) uses Proposition 2.3.3 and step (ii) uses the definition of mutual information, that $I(X; \widehat{X}) = H(X) - H(X \mid \widehat{X})$. The data processing inequality implies that $I(X; \widehat{X}) \leq I(X; Y)$, and using $H(X) = \log(|\mathcal{X}|)$ completes the proof. □

In particular, Corollary 2.3.4 shows that when X is chosen uniformly at random and we observe Y, we have

$$\inf_{\Psi} \mathbb{P}(\Psi(Y) \neq X) \geq 1 - \frac{I(X;Y) + \log 2}{\log |\mathcal{X}|},$$

where the infimum is taken over all testing procedures Ψ. Some interpretation of this quantity is helpful. If we think roughly of the number of bits it takes to describe a variable X uniformly chosen from $\mathcal{X}$, then we expect that $\log_2 |\mathcal{X}|$ bits are necessary (and sufficient). Thus, until we collect enough information that $I(X;Y) \approx \log |\mathcal{X}|$, so that $I(X;Y)/\log |\mathcal{X}| \approx 1$, we are unlikely to be unable to identify the variable X with any substantial probability. So we must collect enough bits to actually discover X.

Example 2.3.5 (20 questions game)**:** In the 20 questions game—a standard children's game—there are two players, the "chooser" and the "guesser," and an agreed upon universe $\mathcal{X}$. The chooser picks an element $x \in \mathcal{X}$, and the guesser's goal is to find x by using a series of yes/no questions about x. We consider optimal strategies for each player in this game, assuming that $\mathcal{X}$ is finite and letting $m = |\mathcal{X}|$ be the universe size for shorthand.

For the guesser, it is clear that at most $\lceil \log_2 m \rceil$ questions are necessary to guess the item X that the chooser has picked—at each round of the game, the guesser asks a question that eliminates half of the remaining possible items. Indeed, let us assume that $m = 2^l$ for some $l \in \mathbb{N}$; if not, the guesser can always make her task more difficult by increasing the size of $\mathcal{X}$ until it is a power of 2. Thus, after k rounds, there are $m2^{-k}$ items left, and we have

$$m \left(\frac{1}{2}\right)^k \leq 1 \text{ if and only if } k \geq \log_2 m.$$

For the converse—the chooser's strategy—let $Y_1, Y_2, \ldots, Y_k$ be the sequence of yes/no answers given to the guesser. Assume that the chooser picks X uniformly at random in $\mathcal{X}$. Then Fano's inequality (2.3.2) implies that for the guess $\widehat{X}$ the guesser makes,

$$\mathbb{P}(\widehat{X} \neq X) \geq 1 - \frac{I(X; Y_1, \ldots, Y_k) + \log 2}{\log m}.$$

By the chain rule for mutual information, we have

$$I(X; Y_1, \ldots, Y_k) = \sum_{i=1}^k I(X; Y_i \mid Y_{1:i-1}) = \sum_{i=1}^k H(Y_i \mid Y_{1:i-1}) - H(Y_i \mid Y_{1:i-1}, X) \leq \sum_{i=1}^k H(Y_i).$$

As the answers Y_i are yes/no, we have $H(Y_i) \leq \log 2$, so that $I(X; Y_{1:k}) \leq k \log 2$. Thus we find

$$\mathbb{P}(\widehat{X} \neq X) \geq 1 - \frac{(k+1)\log 2}{\log m} = \frac{\log_2 m - 1}{\log_2 m} - \frac{k}{\log_2 m},$$

so that we the guesser must have $k \geq \log_2(m/2)$ to be guaranteed that she will make no mistakes. $\diamond$

2.4 A first operational result: entropy and source coding

The final section of this chapter explores the basic results in source coding. Source coding—in its simplest form—tells us precisely the number of bits (or some other form of information storage) are necessary to perfectly encode a seqeunce of random variables $X_1, X_2, \ldots$ drawn according to a known distribution P.

2.4.1 The source coding problem

Assume we receive data consisting of a sequence of symbols $X_1, X_2, \ldots$, drawn from a known distribution P on a finite or countable space $\mathcal{X}$. We wish to choose an encoding, represented by a *d-ary code function* C that maps $\mathcal{X}$ to finite strings consisting of the symbols $\{0, 1, \ldots, d-1\}$. We denote this by $\mathsf{C} : \mathcal{X} \to \{0, 1, \ldots, d-1\}^*$, where the superscript * denotes the length may change from input to input, and use $\ell_{\mathsf{C}}(x)$ to denote the length of the string $\mathsf{C}(x)$.

In general, we will consider a variety of types of codes; we define each in order of complexity of their decoding.

Definition 2.1. *A d-ary code* $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$ *is* non-singular *if for each* $x, x' \in \mathcal{X}$ *we have*

$$\mathsf{C}(x) \neq \mathsf{C}(x') \quad \text{if } x \neq x'.$$

While Definition 2.1 is natural, generally speaking, we wish to transmit or encode a variety of codewords simultaneously, that is, we wish to encode a sequence $X_1, X_2, \ldots$ using the natural *extension* of the code C as the string $\mathsf{C}(X_1)\mathsf{C}(X_2)\mathsf{C}(X_3)\cdots$, where $\mathsf{C}(x_1)\mathsf{C}(x_2)$ denotes the concatenation of the strings $\mathsf{C}(x_1)$ and $\mathsf{C}(x_2)$. In this case, we require that the code be uniquely decodable:

Definition 2.2. *A d-ary code* $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$ *is* uniquely decodable *if for all sequences* $x_1, \ldots, x_n \in \mathcal{X}$ *and* $x'_1, \ldots, x'_n \in \mathcal{X}$ *we have*

$$\mathsf{C}(x_1)\mathsf{C}(x_2)\cdots\mathsf{C}(x_n) = \mathsf{C}(x'_1)\mathsf{C}(x'_2)\cdots\mathsf{C}(x'_n) \quad \text{if and only if } x_1 = x'_1, \ldots, x_n = x'_n.$$

That is, the extension of the code C *to sequences is non-singular.*

While more useful (generally) than simply non-singular codes, uniquely decodable codes may require inspection of an entire string before recovering the first element. With that in mind, we now consider the easiest to use codes, which can always be decoded instantaneously.

Definition 2.3. *A d-ary code* $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$ *is* uniquely decodable *or* instantaneous *if no codeword is the prefix to another codeword.*

As is hopefully apparent from the definitions, all prefix/instantaneous codes are uniquely decodable, which are in turn non-singular. The converse is not true, though we will see a sense in which—as long as we care only about encoding sequences—using prefix instead of uniquely decodable codes has negligible consequences.

For example, written English, with periods (.) and spaces ( ) included at the ends of words (among other punctuation) is an instantaneous encoding of English into the symbols of the alphabet and punctuation, as punctuation symbols enforce that no "codeword" is a prefix of any other. A few more concrete examples may make things more clear.

Example 2.4.1 (Encoding strategies)**:** Consider the encoding schemes below, which encode the letters a, b, c, and d.

Symbol	$\mathsf{C}_1(x)$	$\mathsf{C}_2(x)$	$\mathsf{C}_3(x)$
a	0	00	0
b	00	10	10
c	000	11	110
d	0000	110	111

By inspection, it is clear that C_1 is non-singular but certainly not uniquely decodable (does the sequence 0000 correspond to aaaa, bb, aab, aba, baa, ca, ac, or d?), while C_3 is a prefix code. We leave showing that C_2 is uniquely decodable as an exercise. $\diamond$

2.4.2 The Kraft-McMillan inequalities

We now turn to a few results on the connections between source-coding and entropy. Our first result, the *Kraft-McMillan inequality*, is an essential result that—as we shall see–essentially says that there is no difference in code-lengths attainable by prefix codes and uniquely decodable codes.

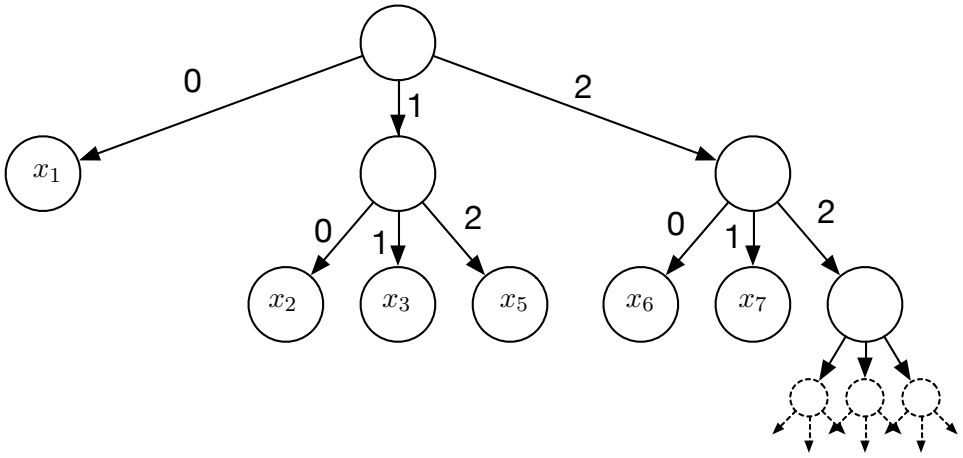

Figure 2.1. Prefix-tree encoding of a set of symbols. The encoding for x_1 is `0`, for x_2 is `10`, for x_3 is `11`, for x_4 is `12`, for x_5 is `20`, for x_6 is `21`, and nothing is encoded as `1`, `2`, or `22`.

Theorem 2.4.2. *Let $\mathcal{X}$ be a finite or countable set, and let $\ell : \mathcal{X} \to \mathbb{N}$ be a function. If $\ell(x)$ is the length of the encoding of the symbol x in a uniquely decodable d-ary code, then*

$$\sum_{x \in \mathcal{X}} d^{-\ell(x)} \le 1. \tag{2.4.1}$$

Conversely, given any function $\ell : \mathcal{X} \to \mathbb{N}$ satisfying inequality (2.4.1), there is a prefix code whose codewords have length $\ell(x)$ for each $x \in \mathcal{X}$.

Proof We prove the first statement of the theorem first by a counting and asymptotic argument.

We begin by assuming that $\mathcal{X}$ is finite; we eliminate this assumption subsequently. As a consequence, there is some maximum length $\ell_{\max}$ such that $\ell(x) \le \ell_{\max}$ for all $x \in \mathcal{X}$. For a sequence $x_1, \ldots, x_n \in \mathcal{X}$, we have by the definition of our encoding strategy that $\ell(x_1, \ldots, x_n) = \sum_{i=1}^n \ell(x_i)$. In addition, for each m we let

$$E_n(m) := \{x_{1:n} \in \mathcal{X}^n \text{ such that } \ell(x_{1:n}) = m\}$$

denote the symbols x encoded with codewords of length m in our code, then as the code is uniquely decodable we certainly have $\mathrm{card}(E_n(m)) \le d^m$ for all n and m. Moreover, for all $x_{1:n} \in \mathcal{X}^n$ we have $\ell(x_{1:n}) \le n\ell_{\max}$. We thus re-index the sum $\sum_x d^{-\ell(x)}$ and compute

$$\begin{aligned}\sum_{x_1, \ldots, x_n \in \mathcal{X}^n} d^{-\ell(x_1, \ldots, x_n)} &= \sum_{m=1}^{n\ell_{\max}} \mathrm{card}(E_n(m)) d^{-m} \\ &\le \sum_{m=1}^{n\ell_{\max}} d^{m-m} = n\ell_{\max}.\end{aligned}$$

The preceding relation is true for all $n \in \mathbb{N}$, so that

$$\left(\sum_{x_{1:n} \in \mathcal{X}^n} d^{-\ell(x_{1:n})} \right)^{1/n} \le n^{1/n} \ell_{\max}^{1/n} \to 1$$

as $n \to \infty$. In particular, using that

$$\sum_{x_{1:n} \in \mathcal{X}^n} d^{-\ell(x_{1:n})} = \sum_{x_1, \ldots, x_n \in \mathcal{X}^n} d^{-\ell(x_1)} \cdots d^{-\ell(x_n)} = \left(\sum_{x \in \mathcal{X}} d^{-\ell(x)} \right)^n,$$

we obtain $\sum_{x \in \mathcal{X}} d^{-\ell(x)} \le 1$.

Returning to the case that $\mathrm{card}(\mathcal{X}) = \infty$, by defining the sequence

$$D_k := \sum_{x \in \mathcal{X}, \ell(x) \le k} d^{-\ell(x)},$$

as each subset $\{x \in \mathcal{X} : \ell(x) \le k\}$ is uniquely decodable, we have $D_k \le 1$ for all k. Then $1 \ge \lim_{k \to \infty} D_k = \sum_{x \in \mathcal{X}} d^{-\ell(x)}$.

The achievability of such a code is straightforward by a pictorial argument (recall Figure 2.1), so we sketch the result non-rigorously. Indeed, let $\mathcal{T}_d$ be an (infinite) d-ary tree. Then, at each level m of the tree, assign one of the nodes at that level to each symbol $x \in \mathcal{X}$ such that $\ell(x) = m$. Eliminate the subtree below that node, and repeat with the remaining symbols. The codeword corresponding to symbol x is then the path to the symbol in the tree.

JCD Comment: Fill out this proof, potentially deferring it.

□

With the Kraft-McMillan theorem in place, we we may directly relate the entropy of a random variable to the length of possible encodings for the variable; in particular, we show that the entropy is essentially *the best* possible code length of a uniquely decodable source code. In this theorem, we use the shorthand

$$H_d(X) := -\sum_{x \in \mathcal{X}} p(x) \log_d p(x).$$

Theorem 2.4.3. *Let $X \in \mathcal{X}$ be a discrete random variable distributed according to P and let ℓ_C be the length function associated with a d-ary encoding $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$. In addition, let $\mathcal{C}$ be the set of all uniquely decodable d-ary codes for $\mathcal{X}$. Then*

$$H_d(X) \le \inf \left\{ \mathbb{E}_P[\ell_{\mathsf{C}}(X)] \ : \ \mathsf{C} \in \mathcal{C} \right\} \le H_d(X) + 1.$$

Proof The lower bound is an argument by convex optimization, while for the upper bound we give an explicit length function and (implicit) prefix code attaining the bound. For the lower bound, we assume for simplicity that $\mathcal{X}$ is finite, and we identify $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$ (let $m = |\mathcal{X}|$ for shorthand). Then as $\mathcal{C}$ consists of *uniquely decodable* codebooks, all the associated length functions must satisfy the Kraft-McMillan inequality (2.4.1). Letting $\ell_i = \ell(i)$, the minimal encoding length is at least

$$\inf_{\ell \in \mathbb{R}^m} \left\{ \sum_{i=1}^m p_i \ell_i : \sum_{i=1}^m d^{-\ell_i} \le 1 \right\}.$$

By introducing the Lagrange multiplier $\lambda \geq 0$ for the inequality constraint, we may write the Lagrangian for the preceding minimization problem as

$$\mathcal{L}(\ell, \lambda) = p^\top \ell + \lambda \left(\sum_{i=1}^n d^{-\ell_i} - 1 \right) \quad \text{with} \quad \nabla_\ell \mathcal{L}(\ell, \lambda) = p - \lambda \left[d^{-\ell_i} \log d \right]_{i=1}^m .$$

In particular, the optimal ℓ satisfies $\ell_i = \log_d \frac{\theta}{p_i}$ for some constant θ, and solving $\sum_{i=1}^m d^{-\log_d \frac{\theta}{p_i}} = 1$ gives $\theta = 1$ and $\ell(i) = \log_d \frac{1}{p_i}$.

To attain the result, simply set our encoding to be $\ell(x) = \left\lceil \log_d \frac{1}{P(X=x)} \right\rceil$, which satisfies the Kraft-McMillan inequality and thus yields a valid prefix code with

$$\mathbb{E}_P[\ell(X)] = \sum_{x \in \mathcal{X}} p(x) \left\lceil \log_d \frac{1}{p(x)} \right\rceil \leq - \sum_{x \in \mathcal{X}} p(x) \log_d p(x) + 1 = H_d(X) + 1$$

as desired. □

Theorem 2.4.3 thus shows that, at least to within an additive constant of 1, the entropy both upper and lower bounds the expected length of a uniquely decodable code for the random variable X. This is the first of our promised "operational interpretations" of the entropy.

2.4.3 Entropy rates and longer codes

Theorem 2.4.3 is a bit unsatisfying in that the additive constant 1 may be quite large relative to the entropy. By allowing encoding longer sequences, we can (asymptotically) eliminate this error factor. To that end, we here show that it is possible, at least for appropriate distributions on random variables X_i, to achieve a per-symbol encoding length that approaches a limiting version of the Shannon entropy of a random variable. We give two definitions capturing the limiting entropy properties of sequences of random variables.

Definition 2.4. *The* entropy rate *of a sequence $X_1, X_2, \ldots$ of random variables is*

$$H(\{X_i\}) := \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n) \tag{2.4.2}$$

whenever the limit exists.

In some situations, the limit (2.4.2) may not exist. However, there are a variety of situations in which it does, and we focus generally on a specific but common instance in which the limit does exist. First, we recall the definition of a stationary sequence of random variables.

Definition 2.5. *We say a sequence $X_1, X_2, \ldots$ of random variable is stationary if for all n and all $k \in \mathbb{N}$ and all measurable sets $A_1, \ldots, A_k \subset \mathcal{X}$ we have*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_k \in A_k) = \mathbb{P}(X_{n+1} \in A_1, \ldots, X_{n+k} \in A_k).$$

With this definition, we have the following result.

Proposition 2.4.4. *Let the sequence of random variables $\{X_i\}$, taking values in the discrete space $\mathcal{X}$, be stationary. Then*

$$H(\{X_i\}) = \lim_{n\to\infty} H(X_n \mid X_1, \ldots, X_{n-1})$$

and the limits (2.4.2) and above exist.

Proof We begin by making the following standard observation of Cesàro means: if $c_n = \frac{1}{n}\sum_{i=1}^n a_i$ and $a_i \to a$, then $c_n \to a$.[3] Now, we note that for a stationary sequence, we have that

$$H(X_n \mid X_{1:n-1}) = H(X_{n+1} \mid X_{2:n}),$$

and using that conditioning decreases entropy, we have

$$H(X_{n+1} \mid X_{1:n}) \le H(X_n \mid X_{1:n-1}).$$

Thus the sequence $a_n := H(X_n \mid X_{1:n-1})$ is non-increasing and bounded below by 0, so that it has some limit $\lim_{n\to\infty} H(X_n \mid X_{1:n-1})$. As $H(X_1, \ldots, X_n) = \sum_{i=1}^n H(X_i \mid X_{1:i-1})$ by the chain rule for entropy, we achieve the result of the proposition. □

Finally, we present a result showing that it is possible to achieve average code length of at most the entropy rate, which for stationary sequences is smaller than the entropy of any single random variable X_i. To do so, we require the use of a block code, which (while it may be prefix code) treats sets of random variables $(X_1, \ldots, X_m) \in \mathcal{X}^m$ as a single symbol to be jointly encoded.

Proposition 2.4.5. *Let the sequence of random variables $X_1, X_2, \ldots$ be stationary. Then for any $\epsilon > 0$, there exists an $m \in \mathbb{N}$ and a d-ary (prefix) block encoder $\mathsf{C} : \mathcal{X}^m \to \{0, \ldots, d-1\}^*$ such that*

$$\lim_n \frac{1}{n}\mathbb{E}_P[\ell_{\mathsf{C}}(X_{1:n})] \le H(\{X_i\}) + \epsilon = \lim_n H(X_n \mid X_1, \ldots, X_{n-1}) + \epsilon.$$

Proof Let $\mathsf{C} : \mathcal{X}^m \to \{0, 1, \ldots, d-1\}^*$ be any prefix code with

$$\ell_{\mathsf{C}}(x_{1:m}) \le \left\lceil \log \frac{1}{P(X_{1:m} = x_{1:m})} \right\rceil.$$

Then whenever n/m is an integer, we have

$$\begin{aligned}\mathbb{E}_P\left[\ell_{\mathsf{C}}(X_{1:n})\right] = \sum_{i=1}^{n/m} \mathbb{E}_P\left[\ell_{\mathsf{C}}(X_{mi+1}, \ldots, X_{m(i+1)})\right] &\le \sum_{i=1}^{n/m} \left[H(X_{mi+1}, \ldots, X_{m(i+1)}) + 1\right] \\ &= \frac{n}{m} + \frac{n}{m} H(X_1, \ldots, X_m).\end{aligned}$$

Dividing by n gives the result by taking m suitably large that $\frac{1}{m} + \frac{1}{m}H(X_1, \ldots, X_m) \le \epsilon + H(\{X_i\})$.

[3] Indeed, let $\epsilon > 0$ and take N such that $n \ge N$ implies that $|a_i - a| < \epsilon$. Then for $n \ge N$, we have

$$c_n - a = \frac{1}{n}\sum_{i=1}^n (a_i - a) = \frac{N(c_N - a)}{n} + \frac{1}{n}\sum_{i=N+1}^n (a_i - a) \in \frac{N(c_N - a)}{n} \pm \epsilon.$$

Taking $n \to \infty$ yields that the term $N(c_N - a)/n \to 0$, which gives that $c_n - a \in [-\epsilon, \epsilon]$ eventually for any $\epsilon > 0$, which is our desired result.

Note that if the m does not divide n, we may also encode the length of the sequence of encoded words in each block of length m; in particular, if the block begins with a 0, it encodes m symbols, while if it begins with a 1, then the next $\lceil \log_d m \rceil$ bits encode the length of the block. This would yields an increase in the expected length of the code to

$$\mathbb{E}_P[\ell_{\mathsf{C}}(X_{1:n})] \le \frac{2n + \lceil \log_2 m \rceil}{m} + \frac{n}{m} H(X_1, \ldots, X_m).$$

Dividing by n and letting $n \to \infty$ gives the result, as we can always choose m large. □

2.5 Bibliography

The material in this chapter is classical in information theory. For all of our treatment of mutual information, entropy, and KL-divergence in the discrete case, Cover and Thomas provide an essentially complete treatment in Chapter 2 of their book [53]. Gray [94] provides a more advanced (measure-theoretic) version of these results, with Chapter 5 covering most of our results (or Chapter 7 in the newer addition of the same book). Csiszár and Körner [55] is the classic reference for coding theorems and results on communication, including stronger converse results.

The f-divergence was independently discovered by Ali and Silvey [4] and Csiszár [54], and is consequently sometimes called an Ali-Silvey divergence or Csiszár divergence. Liese and Vajda [131] provide a survey of f-divergences and their relationships with different statistical concepts (taking a Bayesian point of view), and various authors have extended the pairwise divergence measures to divergence measures between multiple distributions [98], making connections to experimental design and classification [89, 70], which we investigate later in book. The inequalities relating divergences in Section 2.2.4 are now classical, and standard references present them [127, 167]. For a proof that equality (2.2.4) is equivalent to the definition (2.2.3) with the appropriate closure operations, see the paper [70, Proposition 1]. We borrow the proof of the upper bound in Proposition 2.2.10 from the paper [132].

2.6 Exercises

Our first few questions investigate properties of a divergence between distributions that is weaker than the KL-divergence, but is intimately related to optimal testing. Let P_1 and P_2 be arbitrary distributions on a space $\mathcal{X}$. The *total variation distance* between P_1 and P_2 is defined as

$$\|P_1 - P_2\|_{\mathrm{TV}} := \sup_{A \subset \mathcal{X}} |P_1(A) - P_2(A)| .$$

Exercise 2.1: Prove the following identities about total variation. Throughout, let P_1 and P_2 have densities p_1 and p_2 on a (common) set $\mathcal{X}$.

(a) $2 \|P_1 - P_2\|_{\mathrm{TV}} = \int |p_1(x) - p_2(x)| dx$.

(b) For functions $f : \mathcal{X} \to \mathbb{R}$, define the supremum norm $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. Show that $2 \|P_1 - P_2\|_{\mathrm{TV}} = \sup_{\|f\|_\infty \le 1} \int_{\mathcal{X}} f(x)(p_1(x) - p_2(x)) dx$.

(c) $\|P_1 - P_2\|_{\mathrm{TV}} = \int \max\{p_1(x), p_2(x)\} dx - 1$.

(d) $\|P_1 - P_2\|_{\rm TV} = 1 - \int \min\{p_1(x), p_2(x)\}dx$.

(e) For functions $f, g : \mathcal{X} \to \mathbb{R}$,

$$\inf \left\{ \int f(x)p_1(x)dx + \int g(x)p_2(x)dx : f + g \geq 1, f \geq 0, g \geq 0 \right\} = 1 - \|P_1 - P_2\|_{\rm TV}.$$

Exercise 2.2 (Divergence between multivariate normal distributions)**:** Let P_1 be $\mathsf{N}(\theta_1, \Sigma)$ and P_2 be $\mathsf{N}(\theta_2, \Sigma)$, where $\Sigma \succ 0$ is a positive definite matrix. What is $D_{\rm kl}\left(P_1 \| P_2\right)$?

Exercise 2.3 (The optimal test between distributions)**:** Prove Le-Cam's inequality: for any function ψ with $\operatorname{dom} \psi \supset \mathcal{X}$ and any distributions P_1, P_2,

$$P_1(\psi(X) \neq 1) + P_2(\psi(X) \neq 2) \geq 1 - \|P_1 - P_2\|_{\rm TV}.$$

Thus, the sum of the probabilities of error in a hypothesis testing problem, where based on a sample X we must decide whether P_1 or P_2 is more likely, has value at least $1 - \|P_1 - P_2\|_{\rm TV}$. Given P_1 and P_2 is this risk attainable?

Exercise 2.4: A random variable X has $\mathsf{Laplace}(\lambda, \mu)$ distribution if it has density $p(x) = \frac{\lambda}{2}\exp(-\lambda|x-\mu|)$. Consider the hypothesis test of P_1 versus P_2, where X has distribution $\mathsf{Laplace}(\lambda, \mu_1)$ under P_1 and distribution $\mathsf{Laplace}(\lambda, \mu_2)$ under P_2, where $\mu_1 < \mu_2$. Show that the minimal value over all tests ψ of P_1 versus P_2 is

$$\inf_\psi \left\{P_1(\psi(X) \neq 1) + P_2(\psi(X) \neq 2)\right\} = \exp\left(-\frac{\lambda}{2}|\mu_1 - \mu_2|\right).$$

Exercise 2.5 (Log-sum inequality)**:** Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be non-negative reals. Show that

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i\right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}.$$

(Hint: use the convexity of the function $x \mapsto -\log(x)$.)

Exercise 2.6: Given quantizers g_1 and g_2, we say that g_1 is a *finer* quantizer than g_2 under the following condition: assume that g_1 induces the partition $A_1, \ldots, A_n$ and g_2 induces the partition $B_1, \ldots, B_m$; then for any of the sets B_i, there are exists some k and sets $A_{i_1}, \ldots, A_{i_k}$ such that $B_i = \cup_{j=1}^k A_{i_j}$. We let $g_1 \prec g_2$ denote that g_1 is a finer quantizer than g_2. Prove

(a) Finer partitions increase the KL divergence: if $g_1 \prec g_2$,

$$D_{\rm kl}\left(P \| Q \mid g_2\right) \leq D_{\rm kl}\left(P \| Q \mid g_1\right).$$

(b) If $\mathcal{X}$ is discrete (so P and Q have p.m.f.s p and q) then

$$D_{\rm kl}\left(P \| Q\right) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

Exercise 2.7 (f-divergences generalize standard divergences)**:** Show the following properties of f-divergences:

(a) If $f(t) = |t-1|$, then $D_f(P\|Q) = 2\|P-Q\|_{\rm TV}$.

(b) If $f(t) = t\log t$, then $D_f(P\|Q) = D_{\rm kl}(P\|Q)$.

(c) If $f(t) = t\log t - \log t$, then $D_f(P\|Q) = D_{\rm kl}(P\|Q) + D_{\rm kl}(Q\|P)$.

(d) For any convex f satisfying $f(1) = 0$, $D_f(P\|Q) \geq 0$. (Hint: use Jensen's inequality.)

Exercise 2.8 (Generalized "log-sum" inequalities)**:** Let $f : \mathbb{R}_+ \to \mathbb{R}$ be an arbitrary convex function.

(a) Let $a_i, b_i, i = 1, \ldots, n$ be non-negative reals. Prove that

$$\left(\sum_{i=1}^n a_i\right) f\left(\frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n a_i}\right) \leq \sum_{i=1}^n a_i f\left(\frac{b_i}{a_i}\right).$$

(b) Generalizing the preceding result, let $a : \mathcal{X} \to \mathbb{R}_+$ and $b : \mathcal{X} \to \mathbb{R}_+$, and let μ be a finite measure on $\mathcal{X}$ with respect to which a is integrable. Show that

$$\int a(x)d\mu(x) f\left(\frac{\int b(x)d\mu(x)}{\int a(x)d\mu(x)}\right) \leq \int a(x) f\left(\frac{b(x)}{a(x)}\right) d\mu(x).$$

If you are unfamiliar with measure theory, prove the following essentially equivalent result: let $u : \mathcal{X} \to \mathbb{R}_+$ satisfy $\int u(x)dx < \infty$. Show that

$$\int a(x)u(x)dx f\left(\frac{\int b(x)u(x)dx}{\int a(x)u(x)dx}\right) \leq \int a(x) f\left(\frac{b(x)}{a(x)}\right) u(x)dx$$

whenever $\int a(x)u(x)dx < \infty$. (It is possible to demonstrate this remains true under appropriate limits even when $\int a(x)u(x)dx = +\infty$, but it is a mess.)

(Hint: use the fact that the perspective of a function f, defined by $h(x,t) = tf(x/t)$ for $t > 0$, is jointly convex in x and t (see Proposition B.3.12).

Exercise 2.9 (Data processing and f-divergences I)**:** As with the KL-divergence, given a quantizer g of the set $\mathcal{X}$, where g induces a partition $A_1, \ldots, A_m$ of $\mathcal{X}$, we define the f-divergence between P and Q conditioned on g as

$$D_f(P\|Q \mid g) := \sum_{i=1}^m Q(A_i) f\left(\frac{P(A_i)}{Q(A_i)}\right) = \sum_{i=1}^m Q(g^{-1}(\{i\})) f\left(\frac{P(g^{-1}(\{i\}))}{Q(g^{-1}(\{i\}))}\right).$$

Given quantizers g_1 and g_2, we say that g_1 is a *finer* quantizer than g_2 under the following condition: assume that g_1 induces the partition $A_1, \ldots, A_n$ and g_2 induces the partition $B_1, \ldots, B_m$; then for any of the sets B_i, there are exists some k and sets $A_{i_1}, \ldots, A_{i_k}$ such that $B_i = \cup_{j=1}^k A_{i_j}$. We let $g_1 \prec g_2$ denote that g_1 is a finer quantizer than g_2.

(a) Let g_1 and g_2 be quantizers of the set $\mathcal{X}$, and let $g_1 \prec g_2$, meaning that g_1 is a finer quantization than g_2. Prove that

$$D_f(P\|Q \mid g_2) \leq D_f(P\|Q \mid g_1).$$

Equivalently, show that whenever $\mathcal{A}$ and $\mathcal{B}$ are collections of sets partitioning $\mathcal{X}$, but $\mathcal{A}$ is a finer partition of $\mathcal{X}$ than $\mathcal{B}$, that

$$\sum_{B\in\mathcal{B}} Q(B) f\left(\frac{P(B)}{Q(B)}\right) \le \sum_{A\in\mathcal{A}} Q(A) f\left(\frac{P(A)}{Q(A)}\right).$$

(*Hint:* Use the result of Question 2.8(a)).

(b) Suppose that $\mathcal{X}$ is countable (or finite) so that P and Q have p.m.f.s p and q. Show that

$$D_f\left(P\|Q\right) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right),$$

where on the left we are using the partition definition (2.2.3); you should show that the partition into discrete parts of $\mathcal{X}$ achieves the supremum. You may assume that $\mathcal{X}$ is finite. (Though feel free to prove the result in the case that $\mathcal{X}$ is infinite.)

Exercise 2.10 (General data processing inequalities)**:** Let f be a convex function satisfying $f(1) = 0$. Let K be a Markov transition kernel from $\mathcal{X}$ to $\mathcal{Z}$, that is, $K(\cdot, x)$ is a probability distribution on $\mathcal{Z}$ for each $x \in \mathcal{X}$. (Written differently, we have $X \to Z$, and conditioned on $X = x$, Z has distribution $K(\cdot, x)$, so that $K(A, x)$ is the probability that $Z \in A$ given $X = x$.)

(a) Define the marginals $K_P(A) = \int K(A, x)p(x)dx$ and $K_Q(A) = \int K(A, x)q(x)dx$. Show that

$$D_f\left(K_P\|K_Q\right) \le D_f\left(P\|Q\right).$$

Hint: by equation (2.2.3), w.l.o.g. we may assume that $\mathcal{Z}$ is finite and $\mathcal{Z} = \{1, \ldots, m\}$; also recall Question 2.8.

(b) Let X and Y be random variables with joint distribution P_{XY} and marginals P_X and P_Y. Define the f-information between X and Y as

$$I_f(X; Y) := D_f\left(P_{XY}\|P_X \times P_Y\right).$$

Use part (a) to show the following general data processing inequality: if we have the Markov chain $X \to Y \to Z$, then

$$I_f(X; Z) \le I_f(X; Y).$$

Exercise 2.11 (Convexity of f-divergences)**:** Prove Proposition 2.2.11. *Hint:* Use Question 2.8.

Exercise 2.12 (Variational forms of KL divergence)**:** Let P and Q be arbitrary distributions on a common space $\mathcal{X}$. Prove the following variational representation, known as the Donsker-Varadhan theorem, of the KL divergence:

$$D_{\rm kl}\left(P\|Q\right) = \sup_{f:\mathbb{E}_Q[e^{f(X)}]<\infty} \left\{\mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp(f(X))]\right\}.$$

You may assume that P and Q have densities.

Exercise 2.13: Let P and Q have densities p and q with respect to the base measure μ over the set $\mathcal{X}$. (Recall that this is no loss of generality, as we may take $\mu = P + Q$.) Define the support $\operatorname{supp} P := \{x \in \mathcal{X} : p(x) > 0\}$. Show that

$$D_{\rm kl}\left(P\|Q\right) \geq \log \frac{1}{Q(\operatorname{supp} P)}.$$

Exercise 2.14: Let P_1 be $\mathsf{N}(\theta_1, \Sigma_1)$ and P_2 be $\mathsf{N}(\theta_2, \Sigma_2)$, where $\Sigma_i \succ 0$ are positive definite matrices. Give $D_{\rm kl}\left(P_1\|P_2\right)$.

Exercise 2.15: Let $\{P_v\}_{v \in \mathcal{V}}$ be an arbitrary collection of distributions on a space $\mathcal{X}$ and μ be a probability measure on $\mathcal{V}$. Show that if $V \sim \mu$ and conditional on $V = v$, we draw $X \sim P_v$, then

(a) $I(X;V) = \int D_{\rm kl}\left(P_v\|\overline{P}\right) d\mu(v)$, where $\overline{P} = \int P_v d\mu(v)$ is the (weighted) average of the P_v. You may assume that $\mathcal{V}$ is discrete if you like.

(b) For any distribution Q on $\mathcal{X}$, $I(X;V) = \int D_{\rm kl}\left(P_v\|Q\right) d\mu(v) - D_{\rm kl}\left(\overline{P}\|Q\right)$. Conclude that $I(X;V) \leq \int D_{\rm kl}\left(P_v\|Q\right) d\mu(v)$, or, equivalently, $\overline{P}$ minimizes $\int D_{\rm kl}\left(P_v\|Q\right) d\mu(v)$ over all probabilities Q.

Exercise 2.16 (The triangle inequality for variation distance)**:** Let P and Q be distributions on $X_1^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$, and let $P_i(\cdot \mid x_1^{i-1})$ be the conditional distribution of X_i given $X_1^{i-1} = x_1^{i-1}$ (and similarly for Q_i). Show that

$$\|P - Q\|_{\rm TV} \leq \sum_{i=1}^n \mathbb{E}_P\left[\left\|P_i(\cdot \mid X_1^{i-1}) - Q_i(\cdot \mid X_1^{i-1})\right\|_{\rm TV}\right],$$

where the expectation is taken over X_1^{i-1} distributed according to P.

Exercise 2.17: Let $h(p) = -p\log p - (1-p)\log(1-p)$. Show that $h(p) \geq 2\log 2 \cdot \min\{p, 1-p\}$.

Exercise 2.18 (Lin [132], Theorem 8)**:** Let $h(p) = -p\log p - (1-p)\log(1-p)$. Show that $h(p) \leq 2\log 2 \cdot \sqrt{p(1-p)}$.

Exercise 2.19 (Proving Pinsker's inequality via data processing)**:** We work through a proof of Proposition 2.2.8.(a) using the data processing inequality for f-divergences (Proposition 2.2.13).

(a) Define $D_{\rm kl}\left(p\|q\right) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$. Argue that to prove Pinsker's inequality (2.2.10), it is enough to show that $(p-q)^2 \leq \frac{1}{2}D_{\rm kl}\left(p\|q\right)$.

(b) Define the negative binary entropy $h(p) = p\log p + (1-p)\log(1-p)$. Show that

$$h(p) \geq h(q) + h'(q)(p-q) + 2(p-q)^2$$

for any $p, q \in [0,1]$.

(c) Conclude Pinsker's inequality (2.2.10).

JCD Comment: Below are a few potential questions

Exercise 2.20: Use the paper "A New Metric for Probability Distributions" by Dominik Endres and Johannes Schindelin to prove that if $V \sim \mathsf{Uniform}\{0, 1\}$ and $X \mid V = v \sim P_v$, then $\sqrt{I(X; V)}$ is a metric on distributions. (Said differently, $D_{\rm js}(P\|Q)^{1/2}$ is a metric on distributions, and it generates the same topology as the TV-distance.)

Exercise 2.21: Relate the generalized Jensen-Shannon divergence between m distributions to redundancy in encoding.

Chapter 3

Exponential families and statistical modeling

Our second introductory chapter focuses on readers who may be less familiar with statistical modeling methodology and the how and why of fitting different statistical models. As in the preceding introductory chapter on information theory, this chapter will be a fairly terse blitz through the main ideas. Nonetheless, the ideas and distributions here should give us something on which to hang our hats, so to speak, as the distributions and models provide the basis for examples throughout the book. Exponential family models form the basis of much of statistics, as they are a natural step away from the most basic families of distributions—Gaussians—which admit exact computations but are brittle, to a more flexible set of models that retain enough analytical elegance to permit careful analyses while giving power in modeling. A key property is that fitting exponential family models reduces to the minimization of convex functions—convex optimization problems—an operation we treat as a technology akin to evaluating a function like sin or cos. This perspective (which is accurate enough) will arise throughout this book, and informs the philosophy we adopt that once we formulate a problem as convex, it is solved.

3.1 Exponential family models

We begin by defining exponential family distributions, giving several examples to illustrate a few of their properties. There are three key objects when defining a d-dimensional exponential family distribution on an underlying space $\mathcal{X}$: the *sufficient statistic* $\phi : \mathcal{X} \to \mathbb{R}^d$ representing what we model, a *canonical parameter* vector $\theta \in R^d$, and a *carrier* $h : \mathcal{X} \to \mathbb{R}_+$.

In the discrete case, where $\mathcal{X}$ is a discrete set, the exponential family associated with the sufficient statistic ϕ and carrier h has probability mass function

$$p_\theta(x) = h(x) \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right),$$

where A is the *log-partition-function*, sometimes called the *cumulant generating function*, with

$$A(\theta) := \log \sum_{x \in \mathcal{X}} h(x) \exp(\langle \theta, \phi(x) \rangle).$$

In the continuous case, p_θ is instead a density on $\mathcal{X} \subset \mathbb{R}^k$, and p_θ takes the identical form above but

$$A(\theta) = \log \int_{\mathcal{X}} h(x) \exp(\langle \theta, \phi(x) \rangle) dx.$$

We can abstract away from this distinction between discrete and continuous distributions by making the definition measure-theoretic, which we do here for completeness. (But recall the remarks in Section 1.3.)

With our notation, we have the following definition.

Definition 3.1. *The* exponential family *associated with the function ϕ and base measure μ is defined as the set of distributions with densities p_θ with respect to μ, where*

$$p_\theta(x) = \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right), \tag{3.1.1}$$

and the function A is the log-partition-function *(or cumulant function)*

$$A(\theta) := \log \int_{\mathcal{X}} \exp\left(\langle \theta, \phi(x) \rangle\right) d\mu(x) \tag{3.1.2}$$

whenever A is finite (and is $+\infty$ otherwise). The family is regular *if the domain*

$$\Theta := \{\theta \mid A(\theta) < \infty\}$$

is open.

In Definition 3.1, we have included the carrier h in the base measure μ, and frequently we will give ourselves the general notation

$$p_\theta(x) = h(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta)).$$

In some scenarios, it may be convient to re-parameterize the problem in terms of some function $\eta(\theta)$ instead of θ itself; we will not worry about such issues and simply use the formulae that are most convenient.

We now give a few examples of exponential family models.

Example 3.1.1 (Bernoulli distribution)**:** In this case, we have $X \in \{0, 1\}$ and $P(X = 1) = p$ for some $p \in [0, 1]$ in the classical version of a Bernoulli. Thus we take μ to be the counting measure on $\{0, 1\}$, and by setting $\theta = \log \frac{p}{1-p}$ to obtain a canonical representation, we have

$$\begin{aligned} P(X = x) = p(x) = p^x (1 - p)^{1-x} &= \exp(x \log p - x \log(1 - p)) \\ &= \exp\left(x \log \frac{p}{1-p} + \log(1 - p)\right) = \exp\left(x\theta - \log(1 + e^\theta)\right). \end{aligned}$$

The Bernoulli family thus has log-partition function $A(\theta) = \log(1 + e^\theta)$. $\diamond$

Example 3.1.2 (Poisson distribution)**:** The Poisson distribution (for count data) is usually parameterized by some $\lambda > 0$, and for $x \in \mathbb{N}$ has distribution $P_\lambda(X = x) = (1/x!)\lambda^x e^- \lambda$. Thus by taking μ to be counting (discrete) measure on $\{0, 1, \ldots\}$ and setting $\theta = \log \lambda$, we find the density (probability mass function in this case)

$$p(x) = \frac{1}{x!} \lambda^x e^{-\lambda} = \exp(x \log \lambda - \lambda) \frac{1}{x!} = \exp(x\theta - e^\theta) \frac{1}{x!}.$$

Notably, taking $h(x) = (x!)^{-1}$ and log-partition $A(\theta) = e^\theta$, we have probability mass function $p_\theta(x) = h(x) \exp(\theta x - A(\theta))$. $\diamond$

Example 3.1.3 (Normal distribution, mean parameterization)**:** For the d-dimensional normal distribution, we take μ to be Lebesgue measure on $\mathbb{R}^d$. If we fix the covariance and vary only the mean μ in the family $\mathsf{N}(\mu, \Sigma)$, then $X \sim \mathsf{N}(\mu, \Sigma)$ has density

$$p_\mu(x) = \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) - \frac{1}{2}\log\det(2\pi\Sigma)\right).$$

Setting $h(x) = -\frac{1}{2}x^\top\Sigma^{-1}x$ and reparameterizing $\theta = \Sigma^{-1}\mu$, we obtain

$$p_\theta(x) = \underbrace{\exp\left(-\frac{1}{2}x^\top\Sigma^{-1}x - \frac{1}{2}\log\det(2\pi\Sigma)\right)}_{=:h(x)} \exp\left(x^\top\theta - \frac{1}{2}\theta^\top\Sigma\theta\right).$$

In particular, we have carrier $h(x) = \exp(-\frac{1}{2}x^\top\Sigma^{-1}x)/((2\pi)^{d/2}\det(\Sigma))$, sufficient statistic $\phi(x) = x$, and log partition $A(\theta) = \frac{1}{2}\theta^\top\Sigma^{-1}\theta$. $\diamond$

Example 3.1.4 (Normal distribution)**:** Let $X \sim \mathsf{N}(\mu, \Sigma)$. We may re-parameterize this as as $\Theta = \Sigma^{-1}$ and $\theta = \Sigma^{-1}\mu$, and we have density

$$p_{\theta,\Theta}(x) \propto \exp\left(\langle\theta, x\rangle - \frac{1}{2}\langle xx^\top, \Theta\rangle\right),$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product. See Exercise 3.1. $\diamond$

In some cases, it is analytically convenient to include a few more conditions on the exponential family.

Definition 3.2. *Let $\{P_\theta\}_{\theta\in\Theta}$ be an exponential family as in Definition 3.1. The sufficient statistic ϕ is* minimal *if $\Theta = \operatorname{dom} A \subset \mathbb{R}^d$ is full-dimensional and there exists no vector u such that*

$$\langle u, \phi(x)\rangle \quad \textit{is constant } \mu\textit{-almost surely.}$$

Definition 3.2 is essentially equivalent to stating that $\phi(x) = (\phi_1(x), \ldots, \phi_d(x))$ has linearly independent components when viewed as vectors $[\phi_i(x)]_{x\in\mathcal{X}}$. While we do not prove this, via a suitable linear transformation—a variant of Gram-Schmidt orthonormalization—one may modify any non-minimal exponential family $\{P_\theta\}$ into an equivalent minimal exponential family $\{Q_\eta\}$, meaning that the two collections satisfy the equality $\{P_\theta\} = \{Q_\eta\}$ (see Brown [39, Chapter 1]).

3.2 Why exponential families?

There are many reasons for us to study exponential families. The first major reason is their analytical tractability: as the normal distribution does, they often admit relatively straightforward computation, therefore forming a natural basis for modeling decisions. Their analytic tractability has made them the objects of substantial study for nearly the past hundred years; Brown [39] provides a deep and elegant treatment. Moreover, as we see later, they arise as the solutions to several natural optimization problems on the space of probability distributions, and they also enjoy certain robustness properties related to optimal Bayes' procedures (there is, of course, more to come on this topic).

Here, we enumerate a few of their key analytical properties, focusing on the cumulant generating (or log partition) function $A(\theta) = \log \int e^{\langle \theta, \phi(x) \rangle} d\mu(x)$. We begin with a heuristic calculation, where we assume that we exchange differentiation and integration. Assuming that this is the case, we then obtain the important expectation and covariance relationships that

$$\begin{aligned}\nabla A(\theta) &= \frac{1}{\int e^{\langle \theta, \phi(x) \rangle} d\mu(x)} \int \nabla_\theta e^{\langle \theta, \phi(x) \rangle} d\mu(x) \\ &= e^{-A(\theta)} \int \nabla_\theta e^{\langle \theta, \phi(x) \rangle} d\mu(x) = \int \phi(x) e^{\langle \theta, \phi(x) \rangle - A(\theta)} d\mu(x) = \mathbb{E}_\theta[\phi(X)]\end{aligned}$$

because $e^{\langle \theta, \phi(x) \rangle - A(\theta)} = p_\theta(x)$. A completely similar (and still heuristic, at least at this point) calculation gives

$$\nabla^2 A(\theta) = \mathbb{E}_\theta[\phi(X)\phi(X)^\top] - \mathbb{E}_\theta[\phi(X)]\mathbb{E}_\theta[\phi(X)]^\top = \mathrm{Cov}_\theta(\phi(X)).$$

That these identities hold is no accident and is central to the appeal of exponential family models.

The first and, from our perspective, most important result about exponential family models is their convexity. While (assuming the differentiation relationships above hold) the differentiation identity that $\nabla^2 A(\theta) = \mathrm{Cov}_\theta(\phi(X)) \succeq 0$ makes convexity of A immediate, one can also provide a direct argument without appealing to differentiation.

Proposition 3.2.1. *The cumulant-generating function $\theta \mapsto A(\theta)$ is convex, and it is strictly convex if and only if $\mathrm{Cov}_\theta(\phi(X))$ is positive definite for all $\theta \in \mathrm{dom}\, A$.*

Proof Let $\theta_\lambda = \lambda\theta_1 + (1-\lambda)\theta_2$, where $\theta_1, \theta_2 \in \Theta$. Then $1/\lambda \geq 1$ and $1/(1-\lambda) \geq 1$, and Hölder's inequality implies

$$\begin{aligned}\log \int \exp(\langle \theta_\lambda, \phi(x) \rangle) d\mu(x) &= \log \int \exp(\langle \theta_1, \phi(x) \rangle)^\lambda \exp(\langle \theta_2, \phi(x) \rangle)^{1-\lambda} d\mu(x) \\ &\leq \log \left(\int \exp(\langle \theta_1, \phi(x) \rangle)^{\frac{\lambda}{\lambda}} d\mu(x) \right)^\lambda \left(\int \exp(\langle \theta_2, \phi(x) \rangle)^{\frac{1-\lambda}{1-\lambda}} d\mu(x) \right)^{1-\lambda} \\ &= \lambda \log \int \exp(\langle \theta_1, \phi(x) \rangle) d\mu(x) + (1-\lambda) \log \int \exp(\langle \theta_2, \phi(x) \rangle) d\mu(x),\end{aligned}$$

as desired. The strict convexity will be a consequence of Proposition 3.2.2 to come, as there we formally show that $\nabla^2 A(\theta) = \mathrm{Cov}_\theta(\phi(X))$. □

We now show that $A(\theta)$ is indeed infinitely differentiable and how it generates the moments of the sufficient statistics $\phi(x)$. To describe the properties, we provide a bit of notation related to tensor products: for a vector $x \in \mathbb{R}^d$, we let

$$x^{\otimes k} := \underbrace{x \otimes x \otimes \cdots \otimes x}_{k \text{ times}}$$

denote the kth order tensor, or multilinear operator, that for $v_1, \ldots, v_k \in \mathbb{R}^d$ satisfies

$$x^{\otimes k}(v_1, \ldots, v_k) := \langle x, v_1 \rangle \cdots \langle x, v_k \rangle = \prod_{i=1}^k \langle x, v_i \rangle.$$

When $k = 2$, this is the familiar outer product $x^{\otimes 2} = xx^\top$. (More generally, one may think of $x^{\otimes k}$ as a $d \times d \times \cdots \times d$ box, where the $(i_1, \ldots, i_k)$ entry is $[x^{\otimes k}]_{i_1,\ldots,i_k} = x_{i_1} \cdots x_{i_k}$.) With this notation, our first key result regards the differentiability of A, where we can compute (all) derivatives of $e^{A(\theta)}$ by interchanging integration and differentiation.

Proposition 3.2.2. *The cumulant-generating function $\theta \mapsto A(\theta)$ is infinitely differentiable on the interior of its domain $\Theta := \{\theta \in \mathbb{R}^d : A(\theta) < \infty\}$. The moment-generating function*

$$M(\theta) := \int \exp(\langle \theta, \phi(x) \rangle) d\mu(x)$$

is analytic on the set $\Theta_{\mathbb{C}} := \{z \in \mathbb{C}^d \mid \mathrm{Re}\, z \in \Theta\}$. Additionally, the derivatives of M are computed by passing through the integral, that is,

$$\begin{aligned} \nabla_\theta^k M(\theta) = \nabla_\theta^k \int e^{\langle \theta, \phi(x) \rangle} d\mu(x) &= \int \nabla_\theta^k e^{\langle \theta, \phi(x) \rangle} d\mu(x) \\ &= \int \phi(x)^{\otimes k} \exp(\langle \theta, \phi(x) \rangle) d\mu(x). \end{aligned}$$

The proof of the proposition is involved and requires complex analysis, so we defer it to Sec. 3.6.1.

As particular consequences of Proposition 3.2.2, we can rigorously demonstrate the expectation and covariance relationships that

$$\nabla A(\theta) = \frac{1}{\int e^{\langle \theta, \phi(x) \rangle} d\mu(x)} \int \nabla e^{\langle \theta, \phi(x) \rangle} d\mu(x) = \int \phi(x) p_\theta(x) d\mu(x) = \mathbb{E}_\theta[\phi(X)]$$

and

$$\begin{aligned} \nabla^2 A(\theta) &= \frac{1}{\int e^{\langle \theta, \phi(x) \rangle} d\mu(x)} \int \phi(x)^{\otimes 2} e^{\langle \theta, \phi(x) \rangle} d\mu(x) - \frac{(\int \phi(x) e^{\langle \theta, \phi(x) \rangle} d\mu(x))^{\otimes 2}}{(\int e^{\langle \theta, \phi(x) \rangle} d\mu(x))^2} \\ &= \mathbb{E}_\theta[\phi(X)\phi(X)^\top] - \mathbb{E}_\theta[\phi(X)]\mathbb{E}_\theta[\phi(X)]^\top \\ &= \mathrm{Cov}_\theta(\phi(X)). \end{aligned}$$

Minimal exponential families (Definition 3.2) also enjoy a few additional regularity properties. Recall that A is *strictly convex* if

$$A(\lambda\theta_0 + (1-\lambda)\theta_1) < \lambda A(\theta_0) + (1-\lambda) A(\theta_1)$$

whenever $\lambda \in (0, 1)$ and $\theta_0, \theta_1 \in \mathrm{dom}\, A$. We have the following proposition.

Proposition 3.2.3. *Let $\{P_\theta\}$ be a regular exponential family. The log partition function A is strictly convex if and only if $\{P_\theta\}$ is minimal.*

Proof If the family is minimal, then $\mathrm{Var}_\theta(u^\top \phi(X)) > 0$ for any vector u, while $\mathrm{Var}_\theta(u^\top \phi(X)) = u^\top \nabla^2 A(\theta) u$. This implies the strict positive definiteness $\nabla^2 A(\theta) \succ 0$, which is equivalent to strict convexity (see Corollary B.3.2 in Appendix B.3.1). Conversely, if $\nabla^2 A(\theta) \succ 0$ for all $\theta \in \Theta$, then $\mathrm{Var}_\theta(u^\top \phi(X)) > 0$ for all $u \neq 0$ and so $u^\top \phi(x)$ is non-constant in x. $\square$

3.2.1 Fitting an exponential family model

The convexity and differentiability properties make exponential family models especially attractive from a computational perspective. A major focus in statistics is the convergence of estimates of different properties of a population distribution P and whether these estimates are computable. We will develop tools to address the first of these questions, and attendant optimality guarantees, throughout this book. To set the stage for what follows, let us consider what this entails in the context of exponential family models.

Suppose we have a population P (where, for simplicity, we assume P has a density p), and for a given exponential family $\mathcal{P}$ with densities $\{p_\theta\}$, we wish to find the model closest to P. Then it is natural (if we take on faith that the information-theoretic measures we have developed are the "right" ones) find the distribution $P_\theta \in \mathcal{P}$ closest to P in KL-divergence, that is, to solve

$$\minimize_\theta \; D_{\rm kl}\left(P \| P_\theta\right) = \int p(x) \log \frac{p(x)}{p_\theta(x)} dx. \tag{3.2.1}$$

This is evidently equivalent to minimizing

$$-\int p(x) \log p_\theta(x) dx = \int p(x) \left[-\langle \theta, \phi(x) \rangle + A(\theta)\right] dx = -\langle \theta, \mathbb{E}_P[\phi(X)] \rangle + A(\theta).$$

This is always a convex optimization problem (see Appendices B and C for much more on this), as A is convex and the first term is linear, and so has no non-global optima. Here and throughout, as we mention in the introductory remarks to this chapter, we treat convex optimization as a technology: as long as the dimension of a problem is not too large and its objective can be evaluated, it is (essentially) computationally trivial.

Of course, we never have access to the population P fully; instead, we receive a sample $X_1, \ldots, X_n$ from P. In this case, a natural approach is to replace the expected (negative) log likelihood above with its empirical version and solve

$$\minimize_\theta \; -\sum_{i=1}^n \log p_\theta(X_i) = \sum_{i=1}^n [-\langle \theta, \phi(X_i) \rangle + A(\theta)], \tag{3.2.2}$$

which is still a convex optimization problem (as the objective is convex in θ). The maximum likelihood estimate is any vector $\widehat{\theta}_n$ minimizing the negative log likelihood (3.2.2), which by setting gradients to 0 is evidently any vector satisfying

$$\nabla A(\widehat{\theta}_n) = \mathbb{E}_{\widehat{\theta}_n}[\phi(X)] = \frac{1}{n} \sum_{i=1}^n \phi(X_i). \tag{3.2.3}$$

In particular, we need only find a parameter $\widehat{\theta}_n$ matching moments of the empirical distribution of the observed $X_i \sim P$. This $\widehat{\theta}_n$ is unique whenever $\mathrm{Cov}_\theta(\phi(X)) \succ 0$ for all θ, that is, when the covariance of ϕ is full rank in the exponential family model, because then the objective in the minimization problem (3.2.2) is strictly convex.

Let us proceed heuristically for a moment to develop a rough convergence guarantee for the estimator $\widehat{\theta}_n$; the next paragraph assumes a comfort with some of classical asymptotic statistics (and the central limit theorem) and is not essential for what comes later. Then we can see how minimizers of the problem (3.2.2) converge to their population counterparts. Assume that the data

X_i are i.i.d. from an exponential family model $P_{\theta^\star}$. Then we expect that the maximum likelihood estimate $\widehat{\theta}_n$ should converge to $\theta^\star$, and so

$$\frac{1}{n}\sum_{i=1}^n \phi(X_i) = \nabla A(\widehat{\theta}_n) = \nabla A(\theta^\star) + (\nabla^2 A(\theta^\star) + o(1))(\widehat{\theta}_n - \theta^\star).$$

But of course, $\nabla A(\theta^\star) = \mathbb{E}_{\theta^\star}[\phi(X)]$, and so the central limit theorem gives that

$$\frac{1}{n}\sum_{i=1}^n (\phi(X_i) - \nabla A(\theta^\star)) \stackrel{\cdot}{\sim} \mathsf{N}\left(0, n^{-1}\mathrm{Cov}_{\theta^\star}(\phi(X))\right) = \mathsf{N}\left(0, n^{-1}\nabla^2 A(\theta^\star)\right),$$

where $\stackrel{\cdot}{\sim}$ means "is approximately distributed as." Multiplying by $(\nabla^2 A(\theta^\star)+o(1))^{-1} \approx \nabla^2 A(\theta^\star)^{-1}$, we thus see (still working in our heuristic)

$$\begin{aligned}\widehat{\theta}_n - \theta^\star &= (\nabla^2 A(\theta^\star) + o(1))^{-1}\frac{1}{n}\sum_{i=1}^n (\phi(X_i) - \nabla A(\theta^\star)) \\ &\stackrel{\cdot}{\sim} \mathsf{N}\left(0, n^{-1} \cdot \nabla^2 A(\theta^\star)^{-1}\right), \end{aligned} \tag{3.2.4}$$

where we use that $BZ \sim \mathsf{N}(0, B\Sigma B^\top)$ if $Z \sim \mathsf{N}(0, \Sigma)$. (It is possible to make each of these steps fully rigorous.) Thus the cumulant generating function A governs the error we expect in $\widehat{\theta}_n - \theta^\star$.

Much of the rest of this book explores properties of these types of minimization problems: at what rates do we expect $\widehat{\theta}_n$ to converge to a global minimizer of problem (3.2.1)? Can we show that these rates are optimal? Is this the "right" strategy for choosing a parameter? Exponential families form a particular working example to motivate this development.

3.3 Divergence measures and information for exponential families

Their nice analytic properties mean that exponential family models also play nicely with the information theoretic tools we develop. Indeed, consider the KL-divergence between two exponential family distributions P_θ and $P_{\theta+\Delta}$, where $\Delta \in \mathbb{R}^d$. Then we have

$$\begin{aligned} D_{\rm kl}\left(P_\theta \| P_{\theta+\Delta}\right) &= \mathbb{E}_\theta\left[\langle\theta, \phi(X)\rangle - A(\theta) - \langle\theta+\Delta, \phi(X)\rangle + A(\theta+\Delta)\right] \\ &= A(\theta+\Delta) - A(\theta) - \mathbb{E}_\theta[\langle\Delta, \phi(X)\rangle] \\ &= A(\theta+\Delta) - A(\theta) - \nabla A(\theta)^\top\Delta. \end{aligned}$$

Similarly, we have

$$\begin{aligned} D_{\rm kl}\left(P_{\theta+\Delta} \| P_\theta\right) &= \mathbb{E}_{\theta+\Delta}\left[\langle\theta+\Delta, \phi(X)\rangle - A(\theta+\Delta) - \langle\theta, \phi(X)\rangle + A(\theta)\right] \\ &= A(\theta) - A(\theta+\Delta) + \mathbb{E}_{\theta+\Delta}[\langle\Delta, \phi(X)\rangle] \\ &= A(\theta) - A(\theta+\Delta) - \nabla A(\theta+\Delta)^\top(-\Delta). \end{aligned}$$

These identities give an immediate connection with convexity. Indeed, for a differentiable convex function h, the *first-order divergence* associated with h is

$$D_h(u, v) = h(u) - h(v) - \langle\nabla h(v), u - v\rangle, \tag{3.3.1}$$

which is always nonnegative, and is the gap between the linear approximation to the (convex) function h and its actual value. In much of the statistical and machine learning literature, the

divergence (3.3.1) is called a *Bregman divergence*, though we will use the more evocative first-order divergence. These will appear frequently throughout the book and, more generally, appear frequently in work on optimization and statistics.

JCD Comment: Put in a picture of a Bregman divergence

We catalog these results as the following proposition.

Proposition 3.3.1. *Let $\{P_\theta\}$ be an exponential family model with cumulant generating function $A(\theta)$. Then*

$$D_{\rm kl}\left(P_\theta \| P_{\theta+\Delta}\right) = D_A(\theta+\Delta, \theta) \quad \text{and} \quad D_{\rm kl}\left(P_{\theta+\Delta} \| P_\theta\right) = D_A(\theta, \theta+\Delta).$$

Additionally, there exists a $t \in [0,1]$ such that

$$D_{\rm kl}\left(P_\theta \| P_{\theta+\Delta}\right) = \frac{1}{2}\Delta^\top \nabla^2 A(\theta + t\Delta)\Delta,$$

and similarly, there exists a $t \in [0,1]$ such that

$$D_{\rm kl}\left(P_{\theta+\Delta} \| P_\theta\right) = \frac{1}{2}\Delta^\top \nabla^2 A(\theta + t\Delta)\Delta.$$

Proof We have already shown the first two statements; the second two are applications of Taylor's theorem. □

When the perturbation Δ is small, that A is infinitely differentiable then gives that

$$D_{\rm kl}\left(P_\theta \| P_{\theta+\Delta}\right) = \frac{1}{2}\Delta^\top \nabla^2 A(\theta)\Delta + O(\|\Delta\|^3),$$

so that the Hessian $\nabla^2 A(\theta)$ tells quite precisely how the KL divergence changes as θ varies (locally). As we saw already in Example 2.3.2 (and see the next section), when the KL-divergence between two distributions is small, it is hard to test between them, and in the sequel, we will show converses to this. The Hessian $\nabla^2 A(\theta^\star)$ also governs the error in the estimate $\widehat{\theta}_n - \theta^\star$ in our heuristic (3.2.4). When the Hessian $\nabla^2 A(\theta)$ is quite positive semidefinite, the KL divergence $D_{\rm kl}\left(P_\theta \| P_{\theta+\Delta}\right)$ is large, and the asymptotic covariance (3.2.4) is small. For this—and other reasons we address later—for exponential family models, we call

$$\nabla^2 A(\theta) = \mathrm{Cov}_\theta(\phi(X)) = \mathbb{E}_\theta[\nabla \log p_\theta(X) \nabla \log p_\theta(X)^\top] \tag{3.3.2}$$

the *Fisher information* of the parameter θ in the model $\{P_\theta\}$.

3.4 Generalized linear models and regression

We can specialize the general modeling strategies that exponential families provide to more directly address prediction problems, where we wish to predict a target $Y \in \mathcal{Y}$ given covariates $X \in \mathcal{X}$. Here, we almost always have that Y is either discrete or continuous with $\mathcal{Y} \subset \mathbb{R}$. In this case, we have a sufficient statistic $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, and we model $Y \mid X = x$ via the *generalized linear model* (or conditional exponential family model) if it has density or probability mass function

$$p_\theta(y \mid x) = \exp\left(\phi(x,y)^\top \theta - A(\theta \mid x)\right) h(y), \tag{3.4.1}$$

where as before h is the carrier and (in the case that $\mathcal{Y} \subset \mathbb{R}^k$)

$$A(\theta \mid x) = \log \int \exp(\phi(x, y)^\top \theta) h(y) dy$$

or, in the discrete case,

$$A(\theta \mid x) = \log \sum_y \exp(\phi(x, y)^\top \theta) h(y).$$

The log partition function $A(\cdot \mid x)$ provides the same insights for the conditional models (3.4.1) as it does for the unconditional exponential family models in the preceding sections. Indeed, as in Propositions 3.2.1 and 3.2.2, the log partition $A(\cdot \mid x)$ is always $\mathcal{C}^\infty$ on its domain and convex. Moreover, it gives the expected moments of the sufficient statistic ϕ conditional on x, as

$$\nabla A(\theta \mid x) = \mathbb{E}_\theta[\phi(X, Y) \mid X = x],$$

from which we can (typically) extract the mean or other statistics of Y conditional on x.

Three standard examples will be our most frequent motivators throughout this book: linear regression, binary logistic regression, and multiclass logistic regression. We give these three, as well as describing two more important examples involving modeling count data through Poisson regression and making predictions for targets y known to live in a bounded set.

Example 3.4.1 (Linear regression)**:** In linear regression, we wish to predict $Y \in \mathbb{R}$ from a vector $X \in \mathbb{R}^d$, and assume that $Y \mid X = x$ follow the normal distribution $\mathsf{N}(\theta^\top x, \sigma^2)$. In this case, we have

$$\begin{aligned} p_\theta(y \mid x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - x^\top\theta)^2\right) \\ &= \exp\left(\frac{1}{\sigma^2} y x^\top \theta - \frac{1}{2\sigma^2}\theta^\top x x^\top \theta\right) \exp\left(-\frac{1}{2\sigma^2} y^2 + \frac{1}{2}\log(2\pi\sigma^2)\right), \end{aligned}$$

so that we have the exponential family representation (3.4.1) with $\phi(x, y) = \frac{1}{\sigma^2} xy$, $h(y) = \exp(-\frac{1}{2\sigma^2}y^2 + \frac{1}{2}\log(2\pi\sigma^2))$, and $A(\theta) = \frac{1}{2\sigma^2}\theta^\top x x^\top \theta$. As $\nabla A(\theta \mid x) = \mathbb{E}_\theta[\phi(X, Y) \mid X = x] = \frac{1}{\sigma^2} x \mathbb{E}_\theta[Y \mid X = x]$, we easily recover $\mathbb{E}_\theta[Y \mid X = x] = \theta^\top x$. $\diamond$

Frequently, we wish to predict binary or multiclass random variables Y. For example, consider a medical application in which we wish to assess the probability that, based on a set of covariates $x \in \mathbb{R}^d$ (say, blood pressure, height, weight, family history) and individual will have a heart attack in the next 5 years, so that $Y = 1$ indicates heart attack and $Y = -1$ indicates not. The next example shows how we might model this.

Example 3.4.2 (Binary logistic regression)**:** If $Y \in \{-1, 1\}$, we model

$$p_\theta(y \mid x) = \frac{\exp(y x^\top \theta)}{1 + \exp(y x^\top \theta)},$$

where the idea in the probability above is that if $x^\top\theta$ has the same sign as y, then the large $x^\top\theta y$ becomes the higher the probability assigned the label y; when $x^\top\theta y < 0$, the probability is small. Of course, we always have $p_\theta(y \mid x) + p_\theta(-y \mid x) = 1$, and using the identity

$$y x^\top \theta - \log(1 + \exp(y x^\top \theta)) = \frac{y + 1}{2} x^\top \theta - \log(1 + \exp(x^\top \theta))$$

we obtain the generalized linear model representation $\phi(x, y) = \frac{y+1}{2}x$ and $A(\theta \mid x) = \log(1 + \exp(x^\top\theta))$.
As an alternative, we could represent $Y \in \{0, 1\}$ by

$$p_\theta(y \mid x) = \frac{\exp(yx^\top\theta)}{1 + \exp(x^\top\theta)} = \exp\left(yx^\top\theta - \log(1 + e^{x^\top\theta})\right),$$

which has the simpler sufficient statistic $\phi(x, y) = xy$. $\diamond$

Instead of a binary prediction problem, in many cases we have a *multiclass* prediction problem, where we seek to predict a label Y for an object x belonging to one of k different classes. For example, in image recognition, we are given an image x and wish to identify the subject Y of the image, where Y ranges over k classes, such as birds, dogs, cars, trucks, and so on. This too we can model using exponential families.

Example 3.4.3 (Multiclass logistic regression)**:** In the case that we have a k-class prediction problem in which we wish to predict $Y \in \{1, \ldots, k\}$ from $X \in \mathbb{R}^d$, we assign parameters $\theta_y \in \mathbb{R}^d$ to each of the classes $y = 1, \ldots, k$. We then model

$$p_\theta(y \mid x) = \frac{\exp(\theta_y^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} = \exp\left(\theta_y^\top x - \log\left(\sum_{j=1}^k e^{\theta_j^\top x}\right)\right).$$

Here, the idea is that if $\theta_y^\top x > \theta_j^\top x$ for all $j \neq y$, then the model assigns higher probability to class y than any other class; the larger the gap between $\theta_y^\top x$ and $\theta_j^\top x$, the larger the difference in assigned probabilities. $\diamond$

Other approaches with these ideas allow us to model other situations. Poisson regression models are frequent choices for modeling count data. For example, consider an insurance company that wishes to issue premiums for shipping cargo in different seasons and on different routes, and so wishes to predict the number of times a given cargo ship will be damaged by waves over a period of service; we might represent this with a feature vector x encoding information about the ship to be insured, typical weather on the route it will take, and the length of time it will be in service. To model such counts $Y \in \{0, 1, 2, \ldots\}$, we turn to Poisson regression.

Example 3.4.4 (Poisson regression)**:** When $Y \in \mathbb{N}$ is a count, the Poisson distribution with rate $\lambda > 0$ gives $P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$. Poisson regression models λ via $e^{\theta^\top x}$, giving model

$$p_\theta(y \mid x) = \frac{1}{y!}\exp\left(yx^\top\theta - e^{\theta^\top x}\right),$$

so that we have carrier $h(y) = 1/y!$ and the simple sufficient statistic $yx^\top\theta$. The log partition function is $A(\theta \mid x) = e^{\theta^\top x}$. $\diamond$

Lastly, we consider a less standard example, but which highlights the flexibility of these models. Here, we assume a linear regression problem but in which we wish to predict values Y in a bounded range.

Example 3.4.5 (Bounded range regression)**:** Suppose that we know $Y \in [-b, b]$, but we wish to model it via an exponential family model with density

$$p_\theta(y \mid x) = \exp(yx^\top\theta - A(\theta \mid x))\mathbf{1}\left\{y \in [-b, b]\right\},$$

which is non-zero only for $-b \le y \le b$. Letting $s = x^\top \theta$ for shorthand, we have

$$\int_{-b}^{b} e^{ys} dy = \frac{1}{s}\left[e^{bs} - e^{-bs}\right],$$

where the limit as $s \to 0$ is $2b$; the (conditional) log partition function is thus

$$A(\theta \mid x) = \begin{cases} \log \frac{e^{b\theta^\top x} - e^{-b\theta^\top x}}{\theta^\top x} & \text{if } \theta^\top x \neq 0 \\ \log(2b) & \text{otherwise.} \end{cases}$$

While its functional form makes this highly non-obvious, our general results guarantee that $A(\theta \mid x)$ is indeed $\mathcal{C}^\infty$ and convex in θ. We have $\nabla A(\theta \mid x) = x\mathbb{E}_\theta[Y \mid X = x]$ because $\phi(x, y) = xy$, and we can therefore immediately recover $\mathbb{E}_\theta[Y \mid X = x]$. Indeed, set $s = \theta^\top x$, and without loss of generality assume $s \neq 0$. Then

$$\mathbb{E}[Y \mid x^\top \theta = s] = \frac{\partial}{\partial s} \log \frac{e^{bs} - e^{-bs}}{s} = \frac{b(e^{bs} + e^{-bs})}{e^{bs} - e^{-bs}} - \frac{1}{s},$$

which increases from $-b$ to b as $s = x^\top \theta$ increases from $-\infty$ to $+\infty$. $\diamond$

3.4.1 Fitting a generalized linear model from a sample

We briefly revisit the approach in Section 3.2.1 for fitting exponential family models in the context of generalized linear models. In this case, the analogue of the maximum likelihood problem (3.2.2) is to solve

$$\underset{\theta}{\text{minimize}} \quad -\sum_{i=1}^{n} \log p_\theta(Y_i \mid X_i) = \sum_{i=1}^{n} \left[-\phi(X_i, Y_i)^\top \theta + A(\theta \mid X_i)\right].$$

This is a convex optimization problem with $\mathcal{C}^\infty$ objective, so we can treat solving it as an (essentially) trivial problem unless the sample size n or dimension d of θ are astronomically large.

As in the moment matching equality (3.2.3), a necessary and sufficient condition for $\widehat{\theta}_n$ to minimize the above objective is that it achieves 0 gradient, that is,

$$\frac{1}{n}\sum_{i=1}^{n} \nabla A(\widehat{\theta}_n \mid X_i) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i, Y_i).$$

Once again, to find $\widehat{\theta}_n$ amounts to matching moments, as $\nabla A(\theta \mid X_i) = \mathbb{E}[\phi(X, Y) \mid X = X_i]$, and we still enjoy the convexity properties of the standard exponential family models.

In general, we of course do not expect any exponential family or generalized linear model (GLM) to have perfect fidelity to the world: all models are in accurate (but many are useful!). Nonetheless, we can still *fit* any of the GLM models in Examples 3.4.1–3.4.5 to data of the appropriate type. In particular, for the logarithmic loss $\ell(\theta; x, y) = -\log p_\theta(y \mid x)$, we can define the empirical loss

$$L_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; X_i, Y_i).$$

Then, as $n \to \infty$, we expect that $L_n(\theta) \to \mathbb{E}[\ell(\theta; X, Y)]$, so that the minimizing θ should give the best predictions possible according to the loss ℓ. We shall therefore often be interested in such convergence guarantees and the deviations of sample quantities (like L_n) from their population counterparts.

3.5 Lower bounds on testing a parameter's value

We give a bit of a preview here of the tools we will develop to prove fundamental limits in Part II of the book, an *hors d'oeuvres* that points to the techniques we develop. In Section 2.3.1, we presented Le Cam's method and used it in Example 2.3.2 to give a lower bound on the probability of error in a hypothesis test comparing two normal means. This approach extends beyond this simple case, and here we give another example applying it to exponential family models.

We give a stylized version of the problem. Let $\{P_\theta\}$ be an exponential family model with parameter $\theta \in \mathbb{R}^d$. Suppose for some vector $v \in \mathbb{R}^d$, we wish to test whether $v^\top \theta > 0$ or $v^\top \theta < 0$ in the model. For example, in the regression settings in Section 3.4, we may be interested in the effect of a treatment on health outcomes. Then the covariates x contain information about an individual with first index x_1 corresponding to whether the individual is treated or not, while Y measures the outcome of treatment; setting $v = e_1$, we then wish to test whether there is a positive treatment effect $\theta_1 = e_1^\top \theta > 0$ or negative.

Abstracting away the specifics of the scenario, we ask the following question: given an exponential family $\{P_\theta\}$ and a threshold t of interest, at what separation $\delta > 0$ does it become essentially impossible to test

$$v^\top \theta \le t \quad \text{versus} \quad v^\top \theta \ge t + \delta?$$

We give one approach to this using two-point hypothesis testing lower bounds. In this case, we consider testing sequences of two alternatives

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_{1,n} : \theta = \theta_n$$

as n grows, where we observe a sample X_1^n drawn i.i.d. either according to P_{θ_0} (i.e., H_0) or P_{θ_n} (i.e., $H_{1,n}$). By choosing θ_n in a way that makes the separation $v^\top(\theta_n - \theta_0)$ large but testing H_0 against $H_{1,n}$ challenging, we can then (roughly) identify the separation δ at which testing becomes impossible.

Proposition 3.5.1. *Let $\theta_0 \in \mathbb{R}^d$. Then there exists a sequence of parameters θ_n with $\|\theta_n - \theta_0\| = O(1\sqrt{n})$, separation*

$$v^\top(\theta_n - \theta_0) = \frac{1}{\sqrt{n}} \sqrt{v^\top \nabla^2 A(\theta_0)^{-1} v},$$

and for which

$$\inf_\Psi \left\{P_{\theta_0}(\Psi(X_1^n) \neq 0) + P_{\theta_n}(\Psi(X_1^n) \neq 1)\right\} \ge \frac{1}{2} + O(n^{-1/2}).$$

Proof Let $\Delta \in \mathbb{R}^d$ be a potential perturbation to $\theta_1 = \theta_0 + \Delta$, which gives separation $\delta = v^\top \theta_1 - v^\top \theta_0 = v^\top \Delta$. Let $P_0 = P_{\theta_0}$ and $P_1 = P_{\theta_1}$. Then the smallest summed probability of error in testing between P_0 and P_1 based on n observations X_1^n is

$$\inf_\Psi \left\{P_0(\Psi(X_1, \ldots, X_n) \neq 0) + P_1(\Psi(X_1, \ldots, X_n) \neq 1)\right\} = 1 - \|P_0^n - P_1^n\|_{\rm TV}$$

by Proposition 2.3.1. Following the approach of Example 2.3.2, we apply Pinsker's inequality (2.2.10) and use that the KL-divergence tensorizes to find

$$2\left\|P_0^n - P_1^n\right\|_{\rm TV}^2 \le n D_{\rm kl}\left(P_0 \| P_1\right) = n D_{\rm kl}\left(P_{\theta_0} \| P_{\theta_0 + \Delta}\right) = n D_A(\theta_0 + \Delta, \theta_0),$$

where the final equality follows from the equivalence between KL and first-order divergences for exponential families (Proposition 3.3.1).

To guarantee that the summed probability of error is at least $\frac{1}{2}$, that is, $\|P_0^n - P_1^n\|_{\rm TV} \le \frac{1}{2}$, it suffices to choose Δ satisfying $nD_A(\theta_0 + \Delta, \theta_0) \le \frac{1}{2}$. So to maximize the separation $v^\top \Delta$ while guaranteeing a constant probability of error, we (approximately) solve

$$\begin{array}{ll} \text{maximize} & v^\top \Delta \\ \text{subject to} & D_A(\theta_0 + \Delta, \theta_0) \le \frac{1}{2n}. \end{array}$$

Now, consider that $D_A(\theta_0 + \Delta, \theta_0) = \frac{1}{2}\Delta^\top \nabla^2 A(\theta_0)\Delta + O(\|\Delta\|^3)$. Ignoring the higher order term, we consider maximizing $v^\top \Delta$ subject to $\Delta^\top \nabla^2 A(\theta_0)\Delta \le \frac{1}{n}$. A Lagrangian calculation shows that this has solution

$$\Delta = \frac{1}{\sqrt{n}} \frac{1}{\sqrt{v^\top \nabla^2 A(\theta_0)^{-1} v}} \nabla^2 A(\theta_0)^{-1} v.$$

With this choice, we have separation $\delta = v^\top \Delta = \sqrt{v^\top \nabla^2 A(\theta_0)^{-1} v / n}$, and $D_A(\theta_0 + \Delta, \theta_0) = \frac{1}{2n} + O(1/n^{3/2})$. The summed probability of error is at least

$$1 - \|P_0^n - P_1^n\|_{\rm TV} \ge 1 - \sqrt{\frac{n}{4n} + O(n^{-1/2})} = 1 - \sqrt{\frac{1}{4}} + O(n^{-1/2}) = \frac{1}{2} + O(n^{-1/2})$$

as desired. □

Let us briefly sketch out why Proposition 3.5.1 is the "right" answer using the heuristics in Section 3.2.1. For an unknown parameter θ in the exponential family model P_θ, we observe $X_1, \ldots, X_n$, and wish to test whether $v^\top \theta \ge t$ for a given threshold t. Call our null $H_0 : v^\top \theta \le t$, and assume we wish to test at an asymptotic level $\alpha > 0$, meaning the probability the test falsely rejects H_0 is (as $n \to \infty$) is at most α. Assuming the heuristic (3.2.4), we have the approximate distributional equality

$$v^\top \widehat{\theta}_n \stackrel{\cdot}{\sim} \mathsf{N}\left(v^\top \theta, \frac{1}{n} v^\top \nabla^2 A(\widehat{\theta}_n)^{-1} v\right).$$

Note that we have $\widehat{\theta}_n$ on the right side of the distribution; it is possible to make this rigorous, but here we target only intuition building. A natural asymptotically level α test is then

$$T_n := \begin{cases} \text{Reject} & \text{if } v^\top \widehat{\theta}_n \ge t + z_{1-\alpha} \sqrt{v^\top \nabla^2 A(\widehat{\theta}_n)^{-1} v / n} \\ \text{Accept} & \text{otherwise,} \end{cases}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal, $\mathbb{P}(Z \ge z_{1-\alpha}) = \alpha$ for $Z \sim \mathsf{N}(0,1)$. Let θ_0 be such that $v^\top \theta_0 = t$, so H_0 holds. Then

$$P_{\theta_0}(T_n \text{ rejects}) = P_{\theta_0}\left(\sqrt{n} \cdot v^\top (\widehat{\theta}_n - \theta_0) \ge z_{1-\alpha} \sqrt{v^\top \nabla^2 A(\widehat{\theta}_n)^{-1} v}\right) \to \alpha.$$

At least heuristically, then, this separation $\delta = \sqrt{v^\top A(\theta_0)^{-1} v}/\sqrt{n}$ is the fundamental separation in parameter values at which testing becomes possible (or below which it is impossible).

As a brief and suggestive aside, the precise growth of the KL-divergence $D_{\rm kl}\left(P_{\theta_0 + \Delta} \| P_{\theta_0}\right) = \frac{1}{2}\Delta^\top \nabla^2 A(\theta_0)\Delta + O(\|\Delta\|^3)$ near θ_0 plays the fundamental role in both the lower bound and upper bound on testing. When the Hessian $\nabla^2 A(\theta_0)$ is "large," meaning it is very positive definite, distributions with small parameter distances are still well-separated in KL-divergence, making testing easy, while when $\nabla^2 A(\theta_0)$ is small (nearly indefinite), the KL-divergence can be small even for large parameter separations Δ and testing is hard. As a consequence, at least for exponential family models, the Fisher information (3.3.2), which we defined as $\nabla^2 A(\theta) = \mathrm{Cov}_\theta(\phi(X))$, plays a central role in testing and, as we see later, estimation.

3.6 Deferred proofs

We collect proofs that rely on background we do not assume for this book here.

3.6.1 Proof of Proposition 3.2.2

We follow Brown [39]. We demonstrate only the first-order differentiability using Lebesgue's dominated convergence theorem , as higher orders and the interchange of integration and differentiation are essentially identical. Demonstrating first-order complex differentiability is of course enough to show that A is analytic.[1] As the proof of Proposition 3.2.1 does not rely on analyticity of A, we may use its results. Thus, let $\Theta = \operatorname{dom} A(\cdot)$ in $\mathbb{R}^d$, which is convex. We assume Θ has non-empty interior (if the interior is empty, then the convexity of Θ means that it must lie in a lower dimensional subspace; we simply take the interior relative to that subspace and may proceed). We claim the following lemma, which is the key to applying dominated convergence; we state it first for $\mathbb{R}^d$.

Lemma 3.6.1. *Consider any collection $\{\theta_1, \ldots, \theta_m\} \subset \Theta$, and let $\Theta_0 = \operatorname{Conv}\{\theta_i\}_{i=1}^m$ and $C \subset \operatorname{int} \Theta_0$. Then for any $k \in \mathbb{N}$, there exists a constant $K = K(C, k, \{\theta_i\})$ such that for all $\theta_0 \in C$,*

$$\|x\|^k \exp(\langle \theta_0, x \rangle) \le K \max_{j \le m} \exp(\langle \theta_j, x \rangle).$$

Proof Let $\mathbb{B} = \{u \in \mathbb{R}^d \mid \|u\| \le 1\}$ be the unit ball in $\mathbb{R}^d$. For any $\epsilon > 0$, there exists a $K = K(\epsilon)$ such that $\|x\|^k \le K e^{\epsilon \|x\|}$ for all $x \in \mathbb{R}^d$. As $C \subset \operatorname{int} \operatorname{Conv}(\Theta_0)$, there exists an $\epsilon > 0$ such that for all $\theta_0 \in C$, $\theta_0 + 2\epsilon\mathbb{B} \subset \Theta_0$, and by construction, for any $u \in \mathbb{B}$ we can write $\theta_0 + 2\epsilon u = \sum_{j=1}^m \lambda_j \theta_j$ for some $\lambda \in \mathbb{R}_+^m$ with $\mathbf{1}^\top \lambda = 1$. We therefore have

$$\begin{aligned}
\|x\|^k \exp(\langle \theta_0, x \rangle) &\le \|x\|^k \sup_{u \in \mathbb{B}} \exp(\langle \theta_0 + \epsilon u, x \rangle) \\
&= \|x\|^k \exp(\epsilon \|x\|) \exp(\langle \theta_0, x \rangle) \le K \exp(2\epsilon \|x\|) \exp(\langle \theta_0, x \rangle) \\
&= K \sup_{u \in \mathbb{B}} \exp(\langle \theta_0 + 2\epsilon u, x \rangle).
\end{aligned}$$

But using the convexity of $t \mapsto \exp(t)$ and that $\theta_0 + 2\epsilon u \in \Theta_0$, the last quantity has upper bound

$$\sup_{u \in \mathbb{B}} \exp(\langle \theta_0 + 2\epsilon u, x \rangle) \le \max_{j \le m} \exp(\langle \theta_j, x \rangle).$$

This gives the desired claim. □

A similar result is possible with differences of exponentials:

Lemma 3.6.2. *Under the conditions of Lemma 3.6.1, there exists a K such that for any $\theta, \theta_0 \in C$*

$$\frac{e^{\langle \theta, x \rangle} - e^{\langle \theta_0, x \rangle}}{\|\theta - \theta_0\|} \le K \max_{j \le m} e^{\langle \theta_j, x \rangle}.$$

Proof We write

$$\frac{\exp(\langle \theta, x \rangle) - \exp(\langle \theta_0, x \rangle)}{\|\theta - \theta_0\|} = \frac{\exp(\langle \theta - \theta_0, x \rangle) - 1}{\|\theta - \theta_0\|} \exp(\langle \theta_0, x \rangle)$$

[1] For complex functions, Osgood's lemma shows that if A is continuous and holomorphic in each variable individually, it is holomorphic. For a treatment of such ideas in an engineering context, see, e.g. [92, Ch. 1].

so that the lemma is equivalent to showing that

$$\frac{|e^{\langle \theta - \theta_0, x\rangle} - 1|}{\|\theta - \theta_0\|} \le K \max_{j \le m} \exp(\langle \theta_j - \theta_0, x\rangle).$$

From this, we can assume without loss of generality that $\theta_0 = \mathbf{0}$ (by shifting). Now note that by convexity $e^{-a} \ge 1 - a$ for all $a \in \mathbb{R}$, so $1 - e^a \le |a|$ when $a \le 0$. Conversely, if $a > 0$, then $ae^a \ge e^a - 1$ (note that $\frac{d}{da}(ae^a) = ae^a + e^a \ge e^a$), so dividing by $\|x\|$, we see that

$$\frac{|e^{\langle \theta, x\rangle} - 1|}{\|\theta\| \, \|x\|} \le \frac{|e^{\langle \theta, x\rangle} - 1|}{|\langle \theta, x\rangle|} \le \frac{\max\{\langle \theta, x\rangle e^{\langle \theta, x\rangle}, |\langle \theta, x\rangle|\}}{|\langle \theta, x\rangle|} \le e^{\langle \theta, x\rangle} + 1.$$

As $\theta \in C$, Lemma 3.6.1 then implies that

$$\frac{|e^{\langle \theta, x\rangle} - 1|}{\|\theta\|} \le \|x\| \left(e^{\langle \theta, x\rangle} + 1\right) \le K \max_j e^{\langle \theta_j, x\rangle},$$

as desired. □

With the lemmas in hand, we can demonstrate a dominating function for the derivatives. Indeed, fix $\theta_0 \in \operatorname{int} \Theta$ and for $\theta \in \Theta$, define

$$g(\theta, x) = \frac{\exp(\langle \theta, x\rangle) - \exp(\langle \theta_0, x\rangle) - \exp(\langle \theta_0, x\rangle)\langle x, \theta - \theta_0\rangle}{\|\theta - \theta_0\|} = \frac{e^{\langle \theta, x\rangle} - e^{\langle \theta_0, x\rangle} - \langle \nabla e^{\langle \theta_0, x\rangle}, \theta - \theta_0\rangle}{\|\theta - \theta_0\|}.$$

Then $\lim_{\theta \to \theta_0} g(\theta, x) = 0$ by the differentiability of $t \mapsto e^t$. Lemmas 3.6.1 and 3.6.2 show that if we take any collection $\{\theta_j\}_{j=1}^m \subset \Theta$ for which $\theta \in \operatorname{int} \operatorname{Conv}\{\theta_j\}$, then for $C \subset \operatorname{int} \operatorname{Conv}\{\theta_j\}$, there exists a constant K such that

$$|g(\theta, x)| \le \frac{|\exp(\langle \theta, x\rangle) - \exp(\langle \theta_0, x\rangle)|}{\|\theta - \theta_0\|} + \|x\| \exp(\langle \theta_0, x\rangle) \le K \max_j \exp(\langle \theta_j, x\rangle)$$

for all $\theta \in C$. As $\int \max_j e^{\langle \theta_j, x\rangle} d\mu(x) \le \sum_{j=1}^m \int e^{\langle \theta_j, x\rangle} d\mu(x) < \infty$, the dominated convergence theorem thus implies that

$$\lim_{\theta \to \theta_0} \int g(\theta, x) d\mu(x) = 0,$$

and so $M(\theta) = \exp(A(\theta))$ is differentiable in θ, as

$$M(\theta) = M(\theta_0) + \left\langle \int x e^{\langle \theta_0, x\rangle} d\mu(x), \theta - \theta_0 \right\rangle + o(\|\theta - \theta_0\|).$$

It is evident that we have the derivative

$$\nabla M(\theta) = \int \nabla \exp(\langle \theta, x\rangle) d\mu(x).$$

Analyticity Over the subset $\Theta_{\mathbb{C}} := \{\theta + iz \mid \theta \in \Theta, z \in \mathbb{R}^d\}$ (where $i = \sqrt{-1}$ is the imaginary unit), we can extend the preceding results to demonstrate that A is analytic on $\Theta_{\mathbb{C}}$. Indeed, we first simply note that for $a, b \in \mathbb{R}$, $\exp(a + ib) = \exp(a)\exp(ib)$ and $|\exp(a + ib)| = \exp(a)$, i.e. $|e^z| = e^z$ for $z \in \mathbb{C}$, and so Lemmas 3.6.1 and 3.6.2 follow *mutatis-mutandis* as in the real case. These are enough for the application of the dominated convergence theorem above, and we use that $\exp(\cdot)$ is analytic to conclude that $\theta \mapsto M(\theta)$ is analytic on $\Theta_{\mathbb{C}}$.

3.7 Bibliography

3.8 Exercises

Exercise 3.1: In Example 3.1.4, give the sufficient statistic ϕ and an explicit formula for the log partition function $A(\theta, \Theta)$ so that we can write $p_{\theta,\Theta}(x) = \exp(\langle\theta, \phi_1(x)\rangle + \langle\Theta, \phi_2(x)\rangle - A(\theta, \Theta))$.

Exercise 3.2: Consider the binary logistic regression model in Example 3.4.2, and let $\ell(\theta; x, y) = -\log p_\theta(y \mid x)$ be the associated log loss.

(i) Give the Hessian $\nabla^2_\theta \ell(\theta; x, y)$.

(ii) Let $(x_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$ be a sample. Give a sufficient condition for the minimizer of the empirical log loss

$$L_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta; x_i, y_i)$$

to be unique that depends only on the vectors $\{x_i\}$. *Hint.* A convex function h is strictly convex if and only if its Hessian $\nabla^2 h$ is positive definite.

Part I

Concentration, information, stability, and generalization

Chapter 4

Concentration Inequalities

In many scenarios, it is useful to understand how a random variable X behaves by giving bounds on the probability that it deviates far from its mean or median. This can allow us to give prove that estimation and learning procedures will have certain performance, that different decoding and encoding schemes work with high probability, among other results. In this chapter, we give several tools for proving bounds on the probability that random variables are far from their typical values. We conclude the section with a discussion of basic uniform laws of large numbers and applications to empirical risk minimization and statistical learning, though we focus on the relatively simple cases we can treat with our tools.

4.1 Basic tail inequalities

In this first section, we have a simple to state goal: given a random variable X, how does X concentrate around its mean? That is, assuming w.l.o.g. that $\mathbb{E}[X] = 0$, how well can we bound

$$\mathbb{P}(X \geq t)?$$

We begin with the three most classical three inequalities for this purpose: the Markov, Chebyshev, and Chernoff bounds, which are all instances of the same technique.

The basic inequality off of which all else builds is Markov's inequality.

Proposition 4.1.1 (Markov's inequality)**.** *Let X be a nonnegative random variable, meaning that $X \geq 0$ with probability 1. Then*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof For any random variable, $\mathbb{P}(X \geq t) = \mathbb{E}[\mathbf{1}\{X \geq t\}] \leq \mathbb{E}[(X/t)\mathbf{1}\{X \geq t\}] \leq \mathbb{E}[X]/t$, as $X/t \geq 1$ whenever $X \geq t$. □

When we know more about a random variable than that its expectation is finite, we can give somewhat more powerful bounds on the probability that the random variable deviates from its typical values. The first step in this direction, Chebyshev's inequality, requires two moments, and when we have exponential moments, we can give even stronger results. As we shall see, each of these results is but an application of Proposition 4.1.1.

Proposition 4.1.2 (Chebyshev's inequality)**.** *Let X be a random variable with $\mathrm{Var}(X) < \infty$. Then*

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \frac{\mathrm{Var}(X)}{t^2} \quad \text{and} \quad \mathbb{P}(X - \mathbb{E}[X] \leq -t) \leq \frac{\mathrm{Var}(X)}{t^2}$$

for all $t \geq 0$.

Proof We prove only the upper tail result, as the lower tail is identical. We first note that $X - \mathbb{E}[X] \geq t$ implies that $(X - \mathbb{E}[X])^2 \geq t^2$. But of course, the random variable $Z = (X - \mathbb{E}[X])^2$ is nonnegative, so Markov's inequality gives $\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \mathbb{P}(Z \geq t^2) \leq \mathbb{E}[Z]/t^2$, and $\mathbb{E}[Z] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var}(X)$. □

If a random variable has a moment generating function—exponential moments—we can give bounds that enjoy very nice properties when combined with sums of random variables. First, we recall that

$$\varphi_X(\lambda) := \mathbb{E}[e^{\lambda X}]$$

is the moment generating function of the random variable X. Then we have the Chernoff bound.

Proposition 4.1.3. *For any random variable X, we have*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}} = \varphi_X(\lambda) e^{-\lambda t}$$

for all $\lambda \geq 0$.

Proof This is another application of Markov's inequality: for $\lambda > 0$, we have $e^{\lambda X} \geq e^{\lambda t}$ if and only if $X \geq t$, so that $\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq \mathbb{E}[e^{\lambda X}]/e^{\lambda t}$. □

In particular, taking the infimum over all $\lambda \geq 0$ in Proposition 4.1.3 gives the more standard Chernoff (large deviation) bound

$$\mathbb{P}(X \geq t) \leq \exp\left(\inf_{\lambda \geq 0} \log \varphi_X(\lambda) - \lambda t\right).$$

Example 4.1.4 (Gaussian random variables)**:** When X is a mean-zero Gaussian variable with variance σ^2, we have

$$\varphi_X(\lambda) = \mathbb{E}[\exp(\lambda X)] = \exp\left(\frac{\lambda^2 \sigma^2}{2}\right). \tag{4.1.1}$$

To see this, we compute the integral; we have

$$\begin{aligned}
\mathbb{E}[\exp(\lambda X)] & = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\lambda x - \frac{1}{2\sigma^2} x^2\right) dx \\
& = e^{\frac{\lambda^2\sigma^2}{2}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \lambda\sigma^2 x)^2\right) dx}_{=1},
\end{aligned}$$

because this is simply the integral of the Gaussian density.

As a consequence of the equality (4.1.1) and the Chernoff bound technique (Proposition 4.1.3), we see that for X Gaussian with variance σ^2, we have

$$\mathbb{P}(X \geq \mathbb{E}[X] + t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right) \quad \text{and} \quad \mathbb{P}(X \leq \mathbb{E}[X] - t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

for all $t \geq 0$. Indeed, we have $\log \varphi_{X-\mathbb{E}[X]}(\lambda) = \frac{\lambda^2\sigma^2}{2}$, and $\inf_\lambda\{\frac{\lambda^2\sigma^2}{2} - \lambda t\} = -\frac{t^2}{2\sigma^2}$, which is attained by $\lambda = \frac{t}{\sigma^2}$. $\diamond$

4.1.1 Sub-Gaussian random variables

Gaussian random variables are convenient for their nice analytical properties, but a broader class of random variables with similar moment generating functions are known as *sub-Gaussian* random variables.

Definition 4.1. *A random variable X is* sub-Gaussian with parameter σ^2 *if*

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

for all $\lambda \in \mathbb{R}$. We also say such a random variable is σ^2-sub-Gaussian.

Of course, Gaussian random variables satisfy Definition 4.1 with equality. This would be uninteresting if only Gaussian random variables satisfied this property; happily, that is not the case, and we detail several examples.

Example 4.1.5 (Random signs (Rademacher variables))**:** The random variable X taking values $\{-1, 1\}$ with equal property is 1-sub-Gaussian. Indeed, we have

$$\mathbb{E}[\exp(\lambda X)] = \frac{1}{2}e^\lambda + \frac{1}{2}e^{-\lambda} = \frac{1}{2}\sum_{k=0}^\infty \frac{\lambda^k}{k!} + \frac{1}{2}\sum_{k=0}^\infty \frac{(-\lambda)^k}{k!} = \sum_{k=0}^\infty \frac{\lambda^{2k}}{(2k)!} \leq \sum_{k=0}^\infty \frac{(\lambda^2)^k}{2^k k!} = \exp\left(\frac{\lambda^2}{2}\right),$$

as claimed. $\diamond$

Bounded random variables are also sub-Gaussian; indeed, we have the following example.

Example 4.1.6 (Bounded random variables)**:** Suppose that X is bounded, say $X \in [a, b]$. Then Hoeffding's lemma states that

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right),$$

so that X is $(b-a)^2/4$-sub-Gaussian.
We prove a somewhat weaker statement with a simpler argument, while Exercise 4.1 gives one approach to proving the above statement. First, let $\varepsilon \in \{-1, 1\}$ be a Rademacher variable, so that $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}$. We apply a so-called *symmetrization* technique—a common technique in probability theory, statistics, concentration inequalities, and Banach space research—to give a simpler bound. Indeed, let X' be an independent copy of X, so that $\mathbb{E}[X'] = \mathbb{E}[X]$. We have

$$\begin{aligned}\varphi_{X-\mathbb{E}[X]}(\lambda) &= \mathbb{E}\left[\exp(\lambda(X - \mathbb{E}[X']))\right] \leq \mathbb{E}\left[\exp(\lambda(X - X'))\right] \\ &= \mathbb{E}\left[\exp(\lambda\varepsilon(X - X'))\right],\end{aligned}$$

where the inequality follows from Jensen's inequality and the last equality is a conseqence of the fact that $X - X'$ is symmetric about 0. Using the result of Example 4.1.5,

$$\mathbb{E}\left[\exp(\lambda\varepsilon(X - X'))\right] \le \mathbb{E}\left[\exp\left(\frac{\lambda^2(X - X')}{2}\right)\right] \le \exp\left(\frac{\lambda^2(b-a)^2}{2}\right),$$

where the final inequality is immediate from the fact that $|X - X'| \le b - a$. $\diamond$

While Example 4.1.6 shows how a symmetrization technique can give sub-Gaussian behavior, more sophisticated techniques involving explicitly bounding the logarithm of the moment generating function of X, often by calculations involving *exponential tilts* of its density. In particular, letting X be mean zero for simplicity, if we let

$$\psi(\lambda) = \log \varphi_X(\lambda) = \log \mathbb{E}[e^{\lambda X}],$$

then

$$\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \quad \text{and} \quad \psi''(\lambda) = \frac{\mathbb{E}[X^2e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \frac{\mathbb{E}[Xe^{\lambda X}]^2}{\mathbb{E}[e^{\lambda X}]^2},$$

where we can interchange the order of taking expectations and derivatives whenever $\psi(\lambda)$ is finite. Notably, if X has density p_X (with respect to any base measure) then the random variable Y_λ with density

$$p_\lambda(y) = \frac{e^{\lambda y}}{\mathbb{E}[e^{\lambda X}]} p_X(y)$$

(with respect to the same base measure) satisfies

$$\psi'(\lambda) = \mathbb{E}[Y_\lambda] \quad \text{and} \quad \psi''(\lambda) = \mathbb{E}[Y_\lambda^2] - \mathbb{E}[Y_\lambda]^2 = \mathrm{Var}(Y_\lambda).$$

One can exploit this in many ways, which the exercises and coming chapters do. As a particular example, we can give sharper sub-Gaussian constants for Bernoulli random variables.

Example 4.1.7 (Bernoulli random variables)**:** Let X be $\mathsf{Bernoulli}(p)$, so that $X = 1$ with probability p and $X = 0$ otherwise. Then a strengthening of Hoeffding's lemma (also, essentially, due to Hoeffding) is that

$$\log \mathbb{E}[e^{\lambda(X-p)}] \le \frac{\sigma^2(p)}{2}\lambda^2 \quad \text{for} \quad \sigma^2(p) := \frac{1-2p}{2\log\frac{1-p}{p}}.$$

Here we take the limits as $p \to \{0, \frac{1}{2}, 1\}$ and have $\sigma^2(0) = 0$, $\sigma^2(1) = 0$, and $\sigma^2(\frac{1}{2}) = \frac{1}{4}$. Because $p \mapsto \sigma^2(p)$ is concave and symmetric about $p = \frac{1}{2}$, this inequality is always sharper than that of Example 4.1.6. Exercise 4.9 gives one proof of this bound exploiting exponential tilting. $\diamond$

Chernoff bounds for sub-Gaussian random variables are immediate; indeed, they have the same concentration properties as Gaussian random variables, a consequence of the nice analytical properties of their moment generating functions (that their logarithms are at most quadratic). Thus, using the technique of Example 4.1.4, we obtain the following proposition.

Proposition 4.1.8. *Let X be a σ^2-sub-Gaussian. Then for all $t \ge 0$ we have*

$$\mathbb{P}(X - \mathbb{E}[X] \ge t) \vee \mathbb{P}(X - \mathbb{E}[X] \le -t) \le \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Chernoff bounds extend naturally to sums of independent random variables, because moment generating functions of sums of independent random variables become products of moment generating functions.

Proposition 4.1.9. *Let $X_1, X_2, \ldots, X_n$ be independent σ_i^2-sub-Gaussian random variables. Then*

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right)\right] \le \exp\left(\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R},$$

that is, $\sum_{i=1}^n X_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian.

Proof We assume w.l.o.g. that the X_i are mean zero. We have by independence that and sub-Gaussianity that

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] = \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} X_i\right)\right]\mathbb{E}[\exp(\lambda X_n)] \le \exp\left(\frac{\lambda^2 \sigma_n^2}{2}\right)\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} X_i\right)\right].$$

Applying this technique inductively to $X_{n-1}, \ldots, X_1$, we obtain the desired result. □

Two immediate corollary to Propositions 4.1.8 and 4.1.9 show that sums of sub-Gaussian random variables concentrate around their expectations. We begin with a general concentration inequality.

Corollary 4.1.10. *Let X_i be independent σ_i^2-sub-Gaussian random variables. Then for all $t \ge 0$*

$$\max\left\{\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \ge t\right), \mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \le -t\right)\right\} \le \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Additionally, the classical Hoeffding bound, follows when we couple Example 4.1.6 with Corollary 4.1.10: if $X_i \in [a_i, b_i]$, then

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

To give another interpretation of these inequalities, let us assume that X_i are indepenent and σ^2-sub-Gaussian. Then we have that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \ge t\right) \le \exp\left(-\frac{nt^2}{2\sigma^2}\right),$$

or, for $\delta \in (0,1)$, setting $\exp(-\frac{nt^2}{2\sigma^2}) = \delta$ or $t = \frac{\sqrt{2\sigma^2 \log \frac{1}{\delta}}}{\sqrt{n}}$, we have that

$$\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \le \frac{\sqrt{2\sigma^2 \log \frac{1}{\delta}}}{\sqrt{n}} \quad \text{with probability at least } 1 - \delta.$$

There are a variety of other conditions equivalent to sub-Gaussianity, which we capture in the following theorem.

Theorem 4.1.11. *Let X be a mean-zero random variable and $\sigma^2 \geq 0$ be a constant. The following statements are all equivalent, meaning that there are numerical constant factors K_j such that if one statement (i) holds with parameter K_i, then statement (j) holds with parameter $K_j \leq CK_i$, where C is a numerical constant.*

(1) Sub-gaussian tails: $\mathbb{P}(|X| \geq t) \leq 2\exp(-\frac{t^2}{K_1\sigma^2})$ *for all* $t \geq 0$.

(2) Sub-gaussian moments: $\mathbb{E}[|X|^k]^{1/k} \leq K_2\sigma\sqrt{k}$ *for all* k.

(3) Super-exponential moment: $\mathbb{E}[\exp(X^2/(K_3\sigma^2))] \leq e$.

(4) Sub-gaussian moment generating function: $\mathbb{E}[\exp(\lambda X)] \leq \exp(K_4\lambda^2\sigma^2)$ *for all* $\lambda \in \mathbb{R}$.

Particularly, (1) implies (2) with $K_1 = 1$ and $K_2 \leq e^{1/e}$; (2) implies (3) with $K_2 = 1$ and $K_3 = e\sqrt{\frac{2}{e-1}} < 3$; (3) implies (4) with $K_3 = 1$ and $K_4 \leq \frac{3}{4}$; and (4) implies (1) with $K_4 = \frac{1}{2}$ and $K_1 \leq 2$.

This result is standard in the literature on concentration and random variables, but see Appendix 4.5.1 for a proof of this theorem.

For completeness, we can give a tighter result than part (3) of the preceding theorem, giving a concrete upper bound on squares of sub-Gaussian random variables. The technique used in the example, to introduce an independent random variable for auxiliary randomization, is a common and useful technique in probabilistic arguments (similar to our use of symmetrization in Example 4.1.6).

Example 4.1.12 (Sub-Gaussian squares)**:** Let X be a mean-zero σ^2-sub-Gaussian random variable. Then

$$\mathbb{E}[\exp(\lambda X^2)] \leq \frac{1}{[1 - 2\sigma^2\lambda]_+^{\frac{1}{2}}}, \tag{4.1.2}$$

and expression (4.1.2) holds with equality for $X \sim \mathsf{N}(0, \sigma^2)$.
To see this result, we focus on the Gaussian case first and assume (for this case) without loss of generality (by scaling) that $\sigma^2 = 1$. Assuming that $\lambda < \frac{1}{2}$, we have

$$\mathbb{E}[\exp(\lambda Z^2)] = \int \frac{1}{\sqrt{2\pi}} e^{-(\frac{1}{2}-\lambda)z^2} dz = \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1-2\lambda}{2}z^2} dz = \frac{\sqrt{2\pi}}{\sqrt{1-2\lambda}} \frac{1}{\sqrt{2\pi}},$$

the final equality a consequence of the fact that (as we know for normal random variables) $\int e^{-\frac{1}{2\sigma^2}z^2} dz = \sqrt{2\pi\sigma^2}$. When $\lambda \geq \frac{1}{2}$, the above integrals are all infinite, giving the equality in expression (4.1.2).
For the more general inequality, we recall that if Z is an independent $\mathsf{N}(0, 1)$ random variable, then $\mathbb{E}[\exp(tZ)] = \exp(\frac{t^2}{2})$, and so

$$\mathbb{E}[\exp(\lambda X^2)] = \mathbb{E}[\exp(\sqrt{2\lambda} XZ)] \overset{(i)}{\leq} \mathbb{E}\left[\exp(\lambda\sigma^2 Z^2)\right] \overset{(ii)}{=} \frac{1}{[1 - 2\sigma^2\lambda]_+^{\frac{1}{2}}},$$

where inequality (i) follows because X is sub-Gaussian, and inequality (ii) because $Z \sim \mathsf{N}(0, 1)$. $\diamond$

4.1.2 Sub-exponential random variables

A slightly weaker condition than sub-Gaussianity is for a random variable to be *sub-exponential*, which—for a mean-zero random variable—means that its moment generating function exists in a neighborhood of zero.

Definition 4.2. *A random variable X is* sub-exponential *with parameters (τ^2, b) if for all λ such that $|\lambda| \le 1/b$,*

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le \exp\left(\frac{\lambda^2\tau^2}{2}\right).$$

It is clear from Definition 4.2 that a σ^2-sub-Gaussian random variable is $(\sigma^2, 0)$-sub-exponential.

A variety of random variables are sub-exponential. As a first example, χ^2-random variables are sub-exponential with constant values for τ and b:

Example 4.1.13: Let $X = Z^2$, where $Z \sim \mathsf{N}(0,1)$. We claim that

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \le \exp(2\lambda^2) \;\; \text{for } \lambda \le \frac{1}{4}. \tag{4.1.3}$$

Indeed, for $\lambda < \frac{1}{2}$ we have that

$$\mathbb{E}[\exp(\lambda(Z^2 - \mathbb{E}[Z^2]))] = \exp\left(-\frac{1}{2}\log(1-2\lambda) - \lambda\right) \stackrel{(i)}{\le} \exp\left(\lambda + 2\lambda^2 - \lambda\right)$$

where inequality (i) holds for $\lambda \le \frac{1}{4}$, because $-\log(1-2\lambda) \le 2\lambda + 4\lambda^2$ for $\lambda \le \frac{1}{4}$. $\diamond$

As a second example, we can show that bounded random variables are sub-exponential. It is clear that this is the case as they are also sub-Gaussian; however, in many cases, it is possible to show that their parameters yield much tighter control over deviations than is possible using only sub-Gaussian techniques.

Example 4.1.14 (Bounded random variables are sub-exponential)**:** Suppose that X is a mean zero random variable taking values in $[-b, b]$ with variance $\sigma^2 = \mathbb{E}[X^2]$ (note that we are guaranteed that $\sigma^2 \le b^2$ in this case). We claim that

$$\mathbb{E}[\exp(\lambda X)] \le \exp\left(\frac{3\lambda^2\sigma^2}{5}\right) \;\; \text{for } |\lambda| \le \frac{1}{2b}. \tag{4.1.4}$$

To see this, note first that for $k \ge 2$ we have $\mathbb{E}[|X|^k] \le \mathbb{E}[X^2 b^{k-2}] = \sigma^2 b^{k-2}$. Then by an expansion of the exponential, we find

$$\begin{aligned}\mathbb{E}[\exp(\lambda X)] &= 1 + \mathbb{E}[\lambda X] + \frac{\lambda^2\mathbb{E}[X^2]}{2} + \sum_{k=3}^\infty \frac{\lambda^k \mathbb{E}[X^k]}{k!} \le 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^\infty \frac{\lambda^k\sigma^2 b^{k-2}}{k!} \\ &= 1 + \frac{\lambda^2\sigma^2}{2} + \lambda^2\sigma^2 \sum_{k=1}^\infty \frac{(\lambda b)^k}{(k+2)!} \stackrel{(i)}{\le} 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{10},\end{aligned}$$

inequality (i) holding for $\lambda \le \frac{1}{2b}$. Using that $1 + x \le e^x$ gives the result.
It is possible to give a slightly tighter result for $\lambda \ge 0$ In this case, we have the bound

$$\mathbb{E}[\exp(\lambda X)] \le 1 + \frac{\lambda^2\sigma^2}{2} + \lambda^2\sigma^2 \sum_{k=3}^\infty \frac{\lambda^{k-2}b^{k-2}}{k!} = 1 + \frac{\sigma^2}{b^2}\left(e^{\lambda b} - 1 - \lambda b\right).$$

Then using that $1 + x \le e^x$, we obtain *Bennett's moment generating inequality*, which is that

$$\mathbb{E}[e^{\lambda X}] \le \exp\left(\frac{\sigma^2}{b^2}\left(e^{\lambda b} - 1 - \lambda b\right)\right) \quad \text{for } \lambda \ge 0. \tag{4.1.5}$$

Inequality (4.1.5) always holds, and for λb near 0, we have $e^{\lambda b} - 1 - \lambda b \approx \frac{\lambda^2 b^2}{2}$. $\diamond$

In particular, if the variance $\sigma^2 \ll b^2$, the absolute bound on X, inequality (4.1.4) gives much tighter control on the moment generating function of X than typical sub-Gaussian bounds based only on the fact that $X \in [-b, b]$ allow.

More broadly, we can show a result similar to Theorem 4.1.11.

Theorem 4.1.15. *Let X be a random variable and $\sigma \ge 0$. Then—in the sense of Theorem 4.1.11—the following statements are all equivalent for suitable numerical constants $K_1, \ldots, K_4$.*

(1) Sub-exponential tails: $\mathbb{P}(|X| \ge t) \le 2\exp(-\frac{t}{K_1\sigma})$ for all $t \ge 0$

(2) Sub-exponential moments: $\mathbb{E}[|X|^k]^{1/k} \le K_2\sigma k$ for all $k \ge 1$.

(3) Existence of moment generating function: $\mathbb{E}[\exp(X/(K_3\sigma))] \le e$.

(4) If, in addition, $\mathbb{E}[X] = 0$, then $\mathbb{E}[\exp(\lambda X)] \le \exp(K_4\lambda^2\sigma^2)$ for all $|\lambda| \le K_4'/\sigma$.

In particular, if (2) holds with $K_2 = 1$, then (4) holds with $K_4 = 2e^2$ and $K_4' = \frac{1}{2e}$.

The proof, which is similar to that for Theorem 4.1.11, is presented in Section 4.5.2.

While the concentration properties of sub-exponential random variables are not quite so nice as those for sub-Gaussian random variables (recall Hoeffding's inequality, Corollary 4.1.10), we can give sharp tail bounds for sub-exponential random variables. We first give a simple bound on deviation probabilities.

Proposition 4.1.16. *Let X be a mean-zero (τ^2, b)-sub-exponential random variable. Then for all $t \ge 0$,*

$$\mathbb{P}(X \ge t) \vee \mathbb{P}(X \le -t) \le \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{\tau^2}, \frac{t}{b}\right\}\right).$$

Proof The proof is an application of the Chernoff bound technique; we prove only the upper tail as the lower tail is similar. We have

$$\mathbb{P}(X \ge t) \le \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}} \overset{(i)}{\le} \exp\left(\frac{\lambda^2\tau^2}{2} - \lambda t\right),$$

inequality (i) holding for $|\lambda| \le 1/b$. To minimize the last term in λ, we take $\lambda = \min\{\frac{t}{\tau^2}, 1/b\}$, which gives the result. $\square$

Comparing with sub-Gaussian random variables, which have $b = 0$, we see that Proposition 4.1.16 gives a similar result for small t—essentially the same concentration sub-Gaussian random variables—while for large t, the tails decrease only exponentially in t.

We can also give a tensorization identity similar to Proposition 4.1.9.

Proposition 4.1.17. *Let $X_1, \ldots, X_n$ be independent mean-zero sub-exponential random variables, where X_i is (σ_i^2, b_i)-sub-exponential. Then for any vector $a_i \in \mathbb{R}^n$, we have*

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] \le \exp\left(\frac{\lambda^2 \sum_{i=1}^n a_i^2 \sigma_i^2}{2}\right) \quad \text{for} \quad |\lambda| \le \frac{1}{b_*},$$

where $b_* = \max_i b_i |a_i|$. That is, $\langle a, X\rangle$ is $(\sum_{i=1}^n a_i^2 \sigma_i^2, \min_i \frac{1}{b_i |a_i|})$-sub-exponential.

Proof We apply an inductive technique similar to that used in the proof of Proposition 4.1.9. First, for any fixed i, we know that if $|\lambda| \le \frac{1}{b_i |a_i|}$, then $|a_i \lambda| \le \frac{1}{b_i}$ and so

$$\mathbb{E}[\exp(\lambda a_i X_i)] \le \exp\left(\frac{\lambda^2 a_i^2 \sigma_i^2}{2}\right).$$

Now, we inductively apply the preceding inequality, which applies so long as $|\lambda| \le \frac{1}{b_i |a_i|}$ for all i. We have

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n a_i X_i\right)\right] = \prod_{i=1}^n \mathbb{E}[\exp(\lambda a_i X_i)] \le \prod_{i=1}^n \exp\left(\frac{\lambda^2 a_i^2 \sigma_i^2}{2}\right),$$

which is our desired result. □

As in the case of sub-Gaussian random variables, a combination of the tensorization property—that the moment generating functions of sums of sub-exponential random variables are well-behaved—of Proposition 4.1.17 and the concentration inequality (4.1.16) immediately yields the following Bernstein-type inequality. (See also Vershynin [170].)

Corollary 4.1.18. *Let $X_1, \ldots, X_n$ be independent mean-zero (σ_i^2, b_i)-sub-exponential random variables (Definition 4.2). Define $b_* := \max_i b_i$. Then for all $t \ge 0$ and all vectors $a \in \mathbb{R}^n$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \ge t\right) \vee \mathbb{P}\left(\sum_{i=1}^n a_i X_i \le -t\right) \le \exp\left(-\frac{1}{2} \min\left\{\frac{t^2}{\sum_{i=1}^n a_i^2 \sigma_i^2}, \frac{t}{b_* \|a\|_\infty}\right\}\right).$$

It is instructive to study the structure of the bound of Corollary 4.1.18. Notably, the bound is similar to the Hoeffding-type bound of Corollary 4.1.10 (holding for σ^2-sub-Gaussian random variables) that

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \ge t\right) \le \exp\left(-\frac{t^2}{2 \|a\|_2^2 \sigma^2}\right),$$

so that for small t, Corollary 4.1.18 gives sub-Gaussian tail behavior. For large t, the bound is weaker. However, in many cases, Corollary 4.1.18 can give finer control than naive sub-Gaussian bounds. Indeed, suppose that the random variables X_i are i.i.d., mean zero, and satisfy $X_i \in [-b, b]$ with probability 1, but have variance $\sigma^2 = \mathbb{E}[X_i^2] \le b^2$ as in Example 4.1.14. Then Corollary 4.1.18 implies that

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \ge t\right) \le \exp\left(-\frac{1}{2} \min\left\{\frac{5}{6} \frac{t^2}{\sigma^2 \|a\|_2^2}, \frac{t}{2b \|a\|_\infty}\right\}\right). \tag{4.1.6}$$

When applied to a standard mean (and with a minor simplification that $5/12 < 1/3$) with $a_i = \frac{1}{n}$, we obtain the bound that $\frac{1}{n}\sum_{i=1}^n X_i \le t$ with probability at least $1-\exp(-n\min\{\frac{t^2}{3\sigma^2}, \frac{t}{4b}\})$. Written differently, we take $t = \max\{\sigma\sqrt{\frac{3\log\frac{1}{\delta}}{n}}, \frac{4b\log\frac{1}{\delta}}{n}\}$ to obtain

$$\frac{1}{n}\sum_{i=1}^n X_i \le \max\left\{\sigma\frac{\sqrt{3\log\frac{1}{\delta}}}{\sqrt{n}}, \frac{4b\log\frac{1}{\delta}}{n}\right\} \quad \text{with probability } 1-\delta.$$

The sharpest such bound possible via more naive Hoeffding-type bounds is $b\sqrt{2\log\frac{1}{\delta}}/\sqrt{n}$, which has substantially worse scaling.

Further conditions and examples

There are a number of examples and conditions sufficient for random variables to be sub-exponential. One common condition, the so-called *Bernstein* condition, controls the higher moments of a random variable X by its variance. In this case, we say that X satisfies the b-Bernstein condition if

$$|\mathbb{E}[(X-\mu)^k]| \le \frac{k!}{2}\sigma^2 b^{k-2} \quad \text{for } k = 3, 4, \ldots, \tag{4.1.7}$$

where $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}(X) = \mathbb{E}[X^2] - \mu^2$. In this case, the following lemma controls the moment generating function of X. This result is essentially present in Theorem 4.1.15, but it provides somewhat tighter control with precise constants.

Lemma 4.1.19. *Let X be a random variable satisfying the Bernstein condition (4.1.7). Then*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right) \quad \textit{for } |\lambda| \le \frac{1}{b}.$$

Said differently, a random variable satisfying Condition (4.1.7) is $(\sqrt{2}\sigma, b/2)$-sub-exponential.

Proof Without loss of generality we assume $\mu = 0$. We expand the moment generating function by noting that

$$\begin{aligned}\mathbb{E}[e^{\lambda X}] = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^\infty \frac{\lambda^k\mathbb{E}[X^k]}{k!} &\overset{(i)}{\le} 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^\infty |\lambda b|^{k-2} \\ &= 1 + \frac{\lambda^2\sigma^2}{2}\frac{1}{[1-b|\lambda|]_+}\end{aligned}$$

where inequality (i) used the Bernstein condition (4.1.7). Noting that $1+x \le e^x$ gives the result. □

As one final example, we return to Bennett's inequality (4.1.5) from Example 4.1.14.

Proposition 4.1.20 (Bennett's inequality)**.** *Let X_i be independent mean-zero random variables with $\mathrm{Var}(X_i) = \sigma_i^2$ and $|X_i| \le b$. Then for $h(t) := (1+t)\log(1+t) - t$ and $\sigma^2 := \sum_{i=1}^n \sigma_i^2$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le \exp\left(-\frac{\sigma^2}{b^2}h\left(\frac{bt}{\sigma^2}\right)\right).$$

Proof We assume without loss of generality that $\mathbb{E}[X] = 0$. Using the standard Chernoff bound argument coupled with inequality (4.1.5), we see that

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t \sum\right) \leq \exp\left(\sum_{i=1}^n \frac{\sigma_i^2}{b^2}\left(e^{\lambda b} - 1 - \lambda b\right) - \lambda t\right).$$

Letting $h(t) = (1+t)\log(1+t) - t$ as in the statement of the proposition and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$, we minimize over $\lambda \geq 0$, setting $\lambda = \frac{1}{b}\log(1 + \frac{bt}{\sigma^2})$. Substituting into our Chernoff bound application gives the proposition. □

A slightly more intuitive writing of Bennett's inequality is to use averages, in which case for $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ the average of the variances,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{n\sigma^2}{b} h\left(\frac{bt}{\sigma^2}\right)\right).$$

It is possible to show that

$$\frac{n\sigma^2}{b} h\left(\frac{bt}{\sigma^2}\right) \geq \frac{nt^2}{2\sigma^2 + \frac{2}{3}bt},$$

which gives rise to the classical Bernstein inequality that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{nt^2}{2\sigma^2 + \frac{2}{3}bt}\right). \tag{4.1.8}$$

4.1.3 Orlicz norms

Sub-Gaussian and sub-exponential random variables are examples of a broader class of random variables belonging to what are known as *Orlicz-spaces*. For these, we take any convex function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\psi(0) = 0$ and $\psi(t) \to \infty$ as $t \uparrow \infty$, a class called the Orlicz functions. Then the Orlicz norm of a random variable X is

$$\|X\|_\psi := \inf\left\{t > 0 \mid \mathbb{E}[\psi(|X|/t)] \leq 1\right\}. \tag{4.1.9}$$

That this is a norm is not completely trivial, though a few properties are immediate: clearly $\|aX\|_\psi = |a|\,\|X\|_\psi$, and we have $\|X\|_\psi = 0$ if and only if $X = 0$ with probability 1. The key result is that in fact, $\|\cdot\|_\psi$ is actually convex, which then guarantees that it is a norm.

Proposition 4.1.21. *The function $\|\cdot\|_\psi$ is convex on the space of random variables.*

Proof Because ψ is convex and non-decreasing, $x \mapsto \psi(|x|)$ is convex as well. (Convince yourself of this.) Thus, its *perspective transform* $\text{pers}(\psi)(t, |x|) := t\psi(|x|/t)$ is jointly convex in both $t \geq 0$ and x (see Appendix B.3.3). This joint convexity of $\widetilde{\psi}$ implies that for any random variables X_0 and X_1 and t_0, t_1,

$$\mathbb{E}[\text{pers}(\psi)(\lambda t_0 + (1-\lambda)t_1, |\lambda X_0 + (1-\lambda)X_1|)] \leq \lambda\mathbb{E}[\text{pers}(\psi)(t_0, |X_0|)] + (1-\lambda)\mathbb{E}[\text{pers}(\psi)(t_1, |X_1|)].$$

Now note that $\mathbb{E}[\psi(|X|/t)] \leq 1$ if and only if $t\mathbb{E}[\psi(|X|/t)] \leq t$. □

Because $\|\cdot\|_\psi$ is convex and positively homogeneous, we certainly have

$$\|X + Y\|_\psi = 2 \|(X + Y)/2\|_\psi \le \|X\|_\psi + \|Y\|_\psi ,$$

that is, the triangle inequality holds.

We can recover several standard norms on random variables, including some we have already implicitly used. The first are the classical L^p norms, where we take $\psi(t) = t^p$, where we see that

$$\inf\{t > 0 \mid \mathbb{E}[|X|^p/t^p] \le 1\} = \mathbb{E}[|X|^p]^{1/p}.$$

We also have what we term the *sub-Gaussian* and *sub-Exponential* norms, typically denoted by considering the functions

$$\psi_p(x) := \exp\left(|x|^p\right) - 1.$$

These induce the *Orlicz ψ_p-norms*, as for $p \ge 1$, these are convex (as they are the composition of the increasing convex function $\exp(\cdot)$ applied to the nonnegative convex function $|\cdot|^p$). Theorem 4.1.11 shows that we have a natural *sub-Gaussian* norm

$$\|X\|_{\psi_2} := \inf\left\{t > 0 \mid \mathbb{E}[\exp(X^2/t^2)] \le 2\right\}, \tag{4.1.10}$$

while Theorem 4.1.15 shows a natural *sub-exponential norm* (or Orlicz ψ_1-norm)

$$\|X\|_{\psi_1} := \inf\left\{t > 0 \mid \mathbb{E}[\exp(|X|/t)] \le 2\right\}. \tag{4.1.11}$$

Many relationships follow immediately from the definitions (4.1.10) and (4.1.11). For example, any sub-Gaussian random variable (whether or not it is mean zero) has a square that is sub-exponential:

Lemma 4.1.22. *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential, and moreover,*

$$\|X\|_{\psi_2}^2 = \left\|X^2\right\|_{\psi_1}.$$

(This is immediate by definition.) By tracing through the arguments in the proofs of Theorems 4.1.11 and 4.1.15, we can also see that an alternative definition of the two norms could be

$$\sup_{k \in \mathbb{N}} \frac{1}{\sqrt{k}} \mathbb{E}[|X|^k]^{1/k} \quad \text{and} \quad \sup_{k \in \mathbb{N}} \frac{1}{k} \mathbb{E}[|X|^k]^{1/k}$$

for the sub-Gaussian and sub-exponential norms $\|X\|_{\psi_2}$ and $\|X\|_{\psi_1}$, respectively. They are all equivalent.

4.1.4 First applications of concentration: random projections

In this section, we investigate the use of concentration inequalities in random projections. As motivation, consider nearest-neighbor (or k-nearest-neighbor) classification schemes. We have a sequence of data points as pairs (u_i, y_i), where the vectors $u_i \in \mathbb{R}^d$ have labels $y_i \in \{1, \ldots, L\}$, where L is the number of possible labels. Given a new point $u \in \mathbb{R}^d$ that we wish to label, we find the k-nearest neighbors to u in the sample $\{(u_i, y_i)\}_{i=1}^n$, then assign u the majority label of these k-nearest neighbors (ties are broken randomly). Unfortunately, it can be prohibitively expensive to store high-dimensional vectors and search over large datasets to find near vectors; this has motivated a line of work in computer science on fast methods for nearest neighbors based on reducing the

dimension while preserving essential aspects of the dataset. This line of research begins with Indyk and Motwani [112], and continuing through a variety of other works, including Indyk [111] and work on locality-sensitive hashing by Andoni et al. [6], among others. The original approach is due to Johnson and Lindenstrauss, who used the results in the study of Banach spaces [117]; our proof follows a standard argument.

The most specific variant of this problem is as follows: we have n points $u_1, \ldots, u_n$, and we could like to construct a mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^m$, where $m \ll d$, such that

$$\left\|\Phi u_i - \Phi u_j\right\|^2 \in (1 \pm \epsilon) \left\|u_i - u_j\right\|^2.$$

Depending on the norm chosen, this task may be impossible; for the Euclidean (ℓ_2) norm, however, such an embedding is easy to construct using Gaussian random variables and with $m = O(\frac{1}{\epsilon^2} \log n)$. This embedding is known as the Johnson-Lindenstrauss embedding. Note that this size m is *independent* of the dimension d, only depending on the number of points n.

Example 4.1.23 (Johnson-Lindenstrauss)**:** Let the matrix $\Phi \in \mathbb{R}^{m \times d}$ be defined as follows:

$$\Phi_{ij} \stackrel{\text{iid}}{\sim} \mathsf{N}(0, 1/m),$$

and let $\Phi_i \in \mathbb{R}^d$ denote the ith row of this matrix. We claim that

$$m \geq \frac{8}{\epsilon^2} \left[2 \log n + \log \frac{1}{\delta}\right] \quad \text{implies} \quad \left\|\Phi u_i - \Phi u_j\right\|_2^2 \in (1 \pm \epsilon) \left\|u_i - u_j\right\|_2^2$$

for all pairs u_i, u_j with probability at least $1 - \delta$. In particular, $m \gtrsim \frac{\log n}{\epsilon^2}$ is sufficient to achieve accurate dimension reduction with high probability.

To see this, note that for any fixed vector u,

$$\frac{\langle \Phi_i, u \rangle}{\|u\|_2} \sim \mathsf{N}(0, 1/m), \quad \text{and} \quad \frac{\|\Phi u\|_2^2}{\|u\|_2^2} = \sum_{i=1}^m \langle \Phi_i, u / \|u\|_2 \rangle^2$$

is a sum of independent scaled χ^2-random variables. In particular, we have $\mathbb{E}[\|\Phi u / \|u\|_2\|_2^2] = 1$, and using the χ^2-concentration result of Example 4.1.13 yields

$$\begin{aligned}
\mathbb{P}\left(\left|\|\Phi u\|_2^2 / \|u\|_2^2 - 1\right| \geq \epsilon\right) &= \mathbb{P}\left(m \left|\|\Phi u\|_2^2 / \|u\|_2^2 - 1\right| \geq m\epsilon\right) \\
&\leq 2 \inf_{|\lambda| \leq \frac{1}{4}} \exp\left(2m\lambda^2 - \lambda m \epsilon\right) = 2 \exp\left(-\frac{m\epsilon^2}{8}\right),
\end{aligned}$$

the last inequality holding for $\epsilon \in [0, 1]$. Now, using the union bound applied to each of the pairs (u_i, u_j) in the sample, we have

$$\mathbb{P}\left(\text{there exist } i \neq j \text{ s.t. } \left|\|\Phi(u_i - u_j)\|_2^2 - \|u_i - u_j\|_2^2\right| \geq \epsilon \|u_i - u_j\|_2^2\right) \leq 2\binom{n}{2} \exp\left(-\frac{m\epsilon^2}{8}\right).$$

Taking $m \geq \frac{8}{\epsilon^2} \log \frac{n^2}{\delta} = \frac{16}{\epsilon^2} \log n + \frac{8}{\epsilon^2} \log \frac{1}{\delta}$ yields that with probability at least $1 - \delta$, we have $\|\Phi u_i - \Phi u_j\|_2^2 \in (1 \pm \epsilon) \|u_i - u_j\|_2^2$. $\diamond$

Computing low-dimensional embeddings of high-dimensional data is an area of active research, and more recent work has shown how to achieve sharper constants [57] and how to use more structured matrices to allow substantially faster computation of the embeddings Φu (see, for example, Achlioptas [1] for early work in this direction, and Ailon and Chazelle [3] for the so-called "Fast Johnson-Lindenstrauss transform").

4.1.5 A second application of concentration: codebook generation

We now consider a (very simplified and essentially un-implementable) view of encoding a signal for transmission and generation of a codebook for transmitting said signal. Suppose that we have a set of words, or signals, that we wish to transmit; let us index them by $i \in \{1, \ldots, m\}$, so that there are m total signals we wish to communicate across a *binary symmetric channel* Q, meaning that given an input bit $x \in \{0, 1\}$, Q outputs a $z \in \{0, 1\}$ with $Q(Z = x \mid x) = 1 - \epsilon$ and $Q(Z = 1 - x \mid x) = \epsilon$, for some $\epsilon < \frac{1}{2}$. (For simplicity, we assume Q is *memoryless*, meaning that when the channel is used multiple times on a sequence $x_1, \ldots, x_n$, its outputs $Z_1, \ldots, Z_n$ are conditionally independent: $Q(Z_{1:n} = z_{1:n} \mid x_{1:n}) = Q(Z_1 = z_1 \mid x_1) \cdots Q(Z_n = z_n \mid x_n)$.)

We consider a simplified block coding scheme, where we for each i we associate a codeword $x_i \in \{0, 1\}^d$, where d is a dimension (block length) to be chosen. Upon sending the codeword over the channel, and receiving some $z^{\rm rec} \in \{0, 1\}^d$, we decode by choosing

$$i^* \in \operatorname*{argmax}_{i \in [m]} Q(Z = z^{\rm rec} \mid x_i) = \operatorname*{argmin}_{i \in [m]} \|z^{\rm rec} x_i\|_1 , \tag{4.1.12}$$

the maximum likelihood decoder. We now investigate how to choose a collection $\{x_1, \ldots, x_m\}$ of such codewords and give finite sample bounds on its probability of error. In fact, by using concentration inequalities, we can show that a randomly drawn codebook of fairly small dimension is likely to enjoy good performance.

Intuitively, if our codebook $\{x_1, \ldots, x_m\} \subset \{0, 1\}^d$ is *well-separated*, meaning that each pair of words x_i, x_k satisfies $\|x_i - x_k\|_1 \geq cd$ for some numerical constant $c > 0$, we should be unlikely to make a mistake. Let us make this precise. We mistake word i for word k only if the received signal Z satisfies $\|Z - x_i\|_1 \geq \|Z - x_k\|_1$, and letting $J = \{j \in [d] : x_{ij} \neq x_{kj}\}$ denote the set of at least $c \cdot d$ indices where x_i and x_k differ, we have

$$\|Z - x_i\|_1 \geq \|Z - x_k\|_1 \quad \text{if and only if} \quad \sum_{j \in J} |Z_j - x_{ij}| - |Z_j - x_{kj}| \geq 0.$$

If x_i is the word being sent and x_i and x_k differ in position j, then $|Z_j - x_{ij}| - |Z_j - x_{kj}| \in \{-1, 1\}$, and is equal to -1 with probability $(1 - \epsilon)$ and 1 with probability ϵ. That is, we have $\|Z - x_i\|_1 \geq \|Z - x_k\|_1$ if and only if

$$\sum_{j \in J} |Z_j - x_{ij}| - |Z_j - x_{kj}| + |J|(1 - 2\epsilon) \geq |J|(1 - 2\epsilon) \geq cd(1 - 2\epsilon),$$

and the expectation $\mathbb{E}_Q[|Z_j - x_{ij}| - |Z_j - x_{kj}| \mid x_i] = -(1 - 2\epsilon)$ when $x_{ij} \neq x_{kj}$. Using the Hoeffding bound, then, we have

$$Q(\|Z - x_i\|_1 \geq \|Z - x_k\|_1 \mid x_i) \leq \exp\left(-\frac{|J|(1 - 2\epsilon)^2}{2}\right) \leq \exp\left(-\frac{cd(1 - 2\epsilon)^2}{2}\right),$$

where we have used that there are at least $|J| \geq cd$ indices differing between x_i and x_k. The probability of making a mistake at all is thus at most $m \exp(-\frac{1}{2} cd(1 - 2\epsilon)^2)$ if our codebook has separation $c \cdot d$.

For low error decoding to occur with extremely high probability, it is thus sufficient to choose a set of code words $\{x_1, \ldots, x_m\}$ that is well separated. To that end, we state a simple lemma.

Lemma 4.1.24. *Let X_i, $i = 1, \ldots, m$ be drawn independently and uniformly on the d-dimensional hypercube $\mathcal{H}_d := \{0,1\}^d$. Then for any $t \geq 0$,*

$$\mathbb{P}\left(\exists\ i, j\ s.t.\ \|X_i - X_j\|_1 < \frac{d}{2} - dt\right) \leq \binom{m}{2} \exp\left(-2dt^2\right) \leq \frac{m^2}{2} \exp\left(-2dt^2\right).$$

Proof First, let us consider two independent draws X and X' uniformly on the hypercube. Let $Z = \sum_{j=1}^d \mathbf{1}\left\{X_j \neq X'_j\right\} = d_{\rm ham}(X, X') = \|X - X'\|_1$. Then $\mathbb{E}[Z] = \frac{d}{2}$. Moreover, Z is an i.i.d. sum of Bernoulli $\frac{1}{2}$ random variables, so that by our concentration bounds of Corollary 4.1.10, we have

$$\mathbb{P}\left(\left\|X - X'\right\|_1 \leq \frac{d}{2} - t\right) \leq \exp\left(-\frac{2t^2}{d}\right).$$

Using a union bound gives the remainder of the result. □

Rewriting the lemma slightly, we may take $\delta \in (0,1)$. Then

$$\mathbb{P}\left(\exists\ i, j \text{ s.t. } \|X_i - X_j\|_1 < \frac{d}{2} - \sqrt{d \log \frac{1}{\delta} + d \log m}\right) \leq \delta.$$

As a consequence of this lemma, we see two things:

(i) If $m \leq \exp(d/16)$, or $d \geq 16 \log m$, then taking $\delta \uparrow 1$, there at least exists a codebook $\{x_1, \ldots, x_m\}$ of words that are all separated by at least $d/4$, that is, $\|x_i - x_j\|_1 \geq \frac{d}{4}$ for all i, j.

(ii) By taking $m \leq \exp(d/32)$, or $d \geq 32 \log m$, and $\delta = e^{-d/32}$, then with probability at least $1 - e^{-d/32}$—exponentially large in d—a randomly drawn codebook has all its entries separated by at least $\|x_i - x_j\|_1 \geq \frac{d}{4}$.

Summarizing, we have the following result: choose a codebook of m codewords $x_1, \ldots, x_m$ uniformly at random from the hypercube $\mathcal{H}_d = \{0,1\}^d$ with

$$d \geq \max\left\{32 \log m, \frac{8 \log \frac{m}{\delta}}{(1 - 2\epsilon)^2}\right\}.$$

Then with probability at least $1 - 1/m$ over the draw of the codebook, the probability we make a mistake in transmission of any given symbol i over the channel Q is at most δ.

4.2 Martingale methods

The next set of tools we consider constitute our first look at argument sbased on *stability*, that is, how quantities that do not change very much when a single observation changes should concentrate. In this case, we would like to understand more general quantities than sample means, developing a few of the basic cools to understand when functions $f(X_1, \ldots, X_n)$ of independent random variables X_i concentrate around their expectations. Roughly, we expect that if changing the value of one x_i does not significantly change $f(x_1^n)$ much—it is stable—then it should exhibit good concentration properties.

To develop the tools to do this, we go throuhg an approach based on martingales, a deep subject in probability theory. We give a high-level treatment of martingales, taking an approach that does not require measure-theoretic considerations, providing references at the end of the chapter. We begin by providing a definition.

Definition 4.3. *Let $M_1, M_2, \ldots$ be an $\mathbb{R}$-valued sequence of random variables. They are a* martingale *if there exist another sequence of random variables $\{Z_1, Z_2, \ldots\} \subset \mathcal{Z}$ and sequence of functions $f_n : \mathcal{Z}^n \to \mathbb{R}$ such that*

$$\mathbb{E}[M_n \mid Z_1^{n-1}] = M_{n-1} \quad \text{and} \quad M_n = f_n(Z_1^n)$$

for all $n \in \mathbb{N}$. We say that the sequence M_n is adapted *to $\{Z_n\}$.*

In general, the sequence $Z_1, Z_2, \ldots$ is a sequence of increasing σ-fields $\mathcal{F}_1, \mathcal{F}_2, \ldots$, and M_n is $\mathcal{F}_n$-measurable, but Definition 4.3 is sufficienet for our purposes. We also will find it convenient to study *differences* of martingales, so that we make the following

Definition 4.4. *Let $D_1, D_2, \ldots$ be a sequence of random variables. They form a* martingale difference sequence *if $M_n := \sum_{i=1}^n D_i$ is a martingale.*

Equivalently, there is a sequence of random variables Z_n and functions $g_n : \mathcal{Z}^n \to \mathbb{R}$ such that

$$\mathbb{E}[D_n \mid Z_1^{n-1}] = 0 \quad \text{and} \quad D_n = g_n(Z_1^n)$$

for all $n \in \mathbb{N}$.

There are numerous examples of martingale sequences. The classical one is the symmetric random walk.

Example 4.2.1: Let $D_n \in \{\pm 1\}$ be uniform and independent. Then D_n form a martingale difference sequence adapted to themselves (that is, we may take $Z_n = D_n$), and $M_n = \sum_{i=1}^n D_i$ is a martingale. $\diamond$

A more sophisticated example, to which we will frequently return and that suggests the potential usefulness of martingale constructions, is the *Doob martingale* associated with a function f.

Example 4.2.2 (Doob martingales)**:** Let $f : \mathcal{X}^n \to \mathbb{R}$ be an otherwise arbitrary function, and let $X_1, \ldots, X_n$ be arbitrary random variables. The Doob martingale is defined by the difference sequence

$$D_i := \mathbb{E}[f(X_1^n) \mid X_1^i] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}].$$

By inspection, the D_i are functions of X_1^i, and we have

$$\begin{aligned}\mathbb{E}[D_i \mid X_1^{i-1}] &= \mathbb{E}[\mathbb{E}[f(X_1^n) \mid X_1^i] \mid X_1^{i-1}] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}] \\ &= \mathbb{E}[f(X_1^n) \mid X_1^{i-1}] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}] = 0\end{aligned}$$

by the tower property of expectations. Thus, the D_i satisfy Definition 4.4 of a martingale difference sequence, and moreover, we have

$$\sum_{i=1}^n D_i = f(X_1^n) - \mathbb{E}[f(X_1^n)],$$

and so the Doob martingale captures exactly the difference between f and its expectation. $\diamond$

4.2.1 Sub-Gaussian martingales and Azuma-Hoeffding inequalities

With these motivating ideas introduced, we turn to definitions, providing generalizations of our concentration inequalities for sub-Gaussian sums to sub-Gaussian martingales, which we define.

Definition 4.5. *Let $\{D_n\}$ be a martingale difference sequence adapted to $\{Z_n\}$. Then D_n is a* σ_n^2-sub-Gaussian *martingale difference if*

$$\mathbb{E}[\exp(\lambda D_n) \mid Z_1^{n-1}] \le \exp\left(\frac{\lambda^2 \sigma_n^2}{2}\right)$$

for all n and $\lambda \in \mathbb{R}$.

Immediately from the definition, we have the Azuma-Hoeffding inequalities, which generalize the earlier tensorization identities for sub-Gaussian random variables.

Theorem 4.2.3 (Azuma-Hoeffding)**.** *Let $\{D_n\}$ be a σ_n^2-sub-Gaussian martingale difference sequence. Then $M_n = \sum_{i=1}^n D_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian, and moreover,*

$$\max\left\{\mathbb{P}(M_n \ge t), \mathbb{P}(M_n \le -t)\right\} \le \exp\left(-\frac{nt^2}{2\sum_{i=1}^n \sigma_i^2}\right) \quad \text{for all } t \ge 0.$$

Proof The proof is essentially immediate: letting Z_n be the sequence to which the D_n are adapted, we write

$$\begin{aligned}
\mathbb{E}[\exp(\lambda M_n)] &= \mathbb{E}\left[\prod_{i=1}^n e^{\lambda D_i}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^n e^{\lambda D_i} \mid Z_1^{n-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{n-1} e^{\lambda D_i} \mid Z_1^{n-1}\right] \mathbb{E}[e^{\lambda D_n} \mid Z_1^{n-1}]\right]
\end{aligned}$$

because $D_1, \ldots, D_{n-1}$ are functions of Z_1^{n-1}. Then we use Definition 4.5, which implies that $\mathbb{E}[e^{\lambda D_n} \mid Z_1^{n-1}] \le e^{\lambda^2 \sigma_n^2/2}$, and we obtain

$$\mathbb{E}[\exp(\lambda M_n)] \le \mathbb{E}\left[\prod_{i=1}^{n-1} e^{\lambda D_i}\right] \exp\left(\frac{\lambda^2 \sigma_n^2}{2}\right).$$

Repeating the same argument for $n-1, n-2, \ldots, 1$ gives that

$$\log \mathbb{E}[\exp(\lambda M_n)] \le \frac{\lambda^2}{2} \sum_{i=1}^n \sigma_i^2$$

as desired.

The second claims are simply applications of Chernoff bounds via Proposition 4.1.8 and that $\mathbb{E}[M_n] = 0$. □

As an immediate corollary, we recover Proposition 4.1.9, as sums of independent random variables form martingales via $M_n = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$. A second corollary gives what is typically termed the Azuma inequality:

Corollary 4.2.4. *Let D_i be a bounded difference martingale difference sequence, meaning that $|D_i| \le c$. Then $M_n = \sum_{i=1}^n D_i$ satisfies*

$$\mathbb{P}(n^{-1/2} M_n \ge t) \vee \mathbb{P}(n^{-1/2} M_n \le -t) \le \exp\left(-\frac{t^2}{2c^2}\right) \quad \text{for } t \ge 0.$$

Thus, bounded random walks are (with high probability) within $\pm\sqrt{n}$ of their expectations after n steps.

There exist extensions of these inequalities to the cases where we control the variance of the martingales; see Freedman [87].

4.2.2 Examples and bounded differences

We now develop several example applications of the Azuma-Hoeffding inequalities (Theorem 4.2.3), applying them most specifically to functions satisfying certain stability conditions.

We first define the collections of functions we consider.

Definition 4.6 (Bounded differences)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}$ for some space $\mathcal{X}$. Then f satisfies* bounded differences *with constants c_i if for each $i \in \{1, \ldots, n\}$, all $x_1^n \in \mathcal{X}^n$, and $x_i' \in \mathcal{X}$ we have*

$$|f(x_1^{i-1}, x_i, x_{i+1}^n) - f(x_1^{i-1}, x_i', x_{i+1}^n)| \le c_i.$$

The classical inequality relating bounded differences and concentration is McDiarmid's inequality, or the bounded differences inequality.

Proposition 4.2.5 (Bounded differences inequality)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}$ satisfy bounded differences with constants c_i, and let X_i be independent random variables. $f(X_1^n) - \mathbb{E}[f(X_1^n)]$ is $\frac{1}{4}\sum_{i=1}^n c_i^2$-sub-Gaussian, and*

$$\mathbb{P}\left(f(X_1^n) - \mathbb{E}[f(X_1^n)] \ge t\right) \vee \mathbb{P}\left(f(X_1^n) - \mathbb{E}[f(X_1^n)] \le -t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof The basic idea is to show that the Doob martingale (Example 4.2.2) associated with f is $c_i^2/4$-sub-Gaussian, and then to simply apply the Azuma-Hoeffding inequality. To that end, define $D_i = \mathbb{E}[f(X_1^n) \mid X_1^i] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}]$ as before, and note that $\sum_{i=1}^n D_i = f(X_1^n) - \mathbb{E}[f(X_1^n)]$. The random variables

$$L_i := \inf_x \mathbb{E}[f(X_1^n) \mid X_1^{i-1}, X_i = x] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}]$$
$$U_i := \sup_x \mathbb{E}[f(X_1^n) \mid X_1^{i-1}, X_i = x] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}]$$

evidently satisfy $L_i \le D_i \le U_i$, and moreover, we have

$$\begin{aligned} U_i - L_i &\le \sup_{x_1^{i-1}} \sup_{x,x'} \left\{\mathbb{E}[f(X_1^n) \mid X_1^{i-1} = x_1^{i-1}, X_i = x] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1} = x_1^{i-1}, X_i = x']\right\} \\ &= \sup_{x_1^{i-1}} \sup_{x,x'} \int \left(f(x_1^{i-1}, x, x_{i+1}^n) - f(x_1^{i-1}, x', x_{i+1}^n)\right) dP(x_{i+1}^n) \le c_i, \end{aligned}$$

where we have used the independence of the X_i and Definition 4.6 of bounded differences. Consequently, we have by Hoeffding's Lemma (Example 4.1.6) that $\mathbb{E}[e^{\lambda D_i} \mid X_1^{i-1}] \le \exp(\lambda^2 c_i^2/8)$, that is, the Doob martingale is $c_i^2/4$-sub-Gaussian.

The remainder of the proof is simply Theorem 4.2.3. □

A number of quantities satisfy the conditions of Proposition 4.2.5, and we give two examples here; we will revisit them more later.

Example 4.2.6 (Bounded random vectors)**:** Let $\mathbb{B}$ be a Banach space—a complete normed vector space—with norm $\|\cdot\|$. Let X_i be independent bounded random vectors in $\mathbb{B}$ satisfying $\mathbb{E}[X_i] = 0$ and $\|X_i\| \le c$. We claim that the quantity

$$f(X_1^n) := \left\| \frac{1}{n} \sum_{i=1}^n X_i \right\|$$

satisfies bounded differences. Indeed, we have by the triangle inequality that

$$|f(x_1^{i-1}, x, x_{i+1}^n) - f(x_1^{i-1}, x', x_{i+1}^n)| \le \frac{1}{n} \left\| x - x' \right\| \le \frac{2c}{n}.$$

Consequently, if X_i are indpendent, we have

$$\mathbb{P}\left(\left| \left\| \frac{1}{n} \sum_{i=1}^n X_i \right\| - \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n X_i \right\| \right] \right| \ge t \right) \le 2 \exp\left(-\frac{nt^2}{2c^2} \right) \tag{4.2.1}$$

for all $t \ge 0$. That is, the norm of (bounded) random vectors in an essentially arbitrary vector space concentrates extremely quickly about its expectation.
The challenge becomes to control the *expectation* term in the concentration bound (4.2.1), which can be a bit challenging. In certain cases—for example, when we have a Euclidean structure on the vectors X_i—it can be easier. Indeed, let us specialize to the case that $X_i \in \mathcal{H}$, a (real) Hilbert space, so that there is an inner product $\langle \cdot, \cdot \rangle$ and the norm satisfies $\|x\|^2 = \langle x, x \rangle$ for $x \in \mathcal{H}$. Then Cauchy-Schwarz implies that

$$\mathbb{E}\left[\left\| \sum_{i=1}^n X_i \right\| \right]^2 \le \mathbb{E}\left[\left\| \sum_{i=1}^n X_i \right\|^2 \right] = \sum_{i,j} \mathbb{E}[\langle X_i, X_j \rangle] = \sum_{i=1}^n \mathbb{E}[\|X_i\|^2].$$

That is assuming the X_i are independent and $\mathbb{E}[\|X_i\|^2] \le \sigma^2$, inequality (4.2.1) becomes

$$\mathbb{P}\left(\left\| \overline{X}_n \right\| \ge \frac{\sigma}{\sqrt{n}} + t \right) + \mathbb{P}\left(\left\| \overline{X}_n \right\| \le -\frac{\sigma}{\sqrt{n}} - t \right) \le 2 \exp\left(-\frac{nt^2}{2c^2} \right)$$

where $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. ◇

We can specialize Example 4.2.6 to a situation that is very important for treatments of concentration, sums of random vectors, and generalization bounds in machine learning.

Example 4.2.7 (Rademacher complexities)**:** This example is actually a special case of Example 4.2.6, but its frequent uses justify a more specialized treatment and consideration. Let $\mathcal{X}$ be some space, and let $\mathcal{F}$ be some collection of functions $f : \mathcal{X} \to \mathbb{R}$. Let $\varepsilon_i \in \{-1, 1\}$ be a collection of independent random sign vectors. Then the *empirical Rademacher complexity* of $\mathcal{F}$ is

$$R_n(\mathcal{F} \mid x_1^n) := \frac{1}{n} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i f(x_i) \right],$$

where the expectation is over only the random signs ε_i. (In some cases, depending on context and convenience, one takes the absolute value $|\sum_i \varepsilon_i f(x_i)|$.) The *Rademacher complexity* of $\mathcal{F}$ is

$$R_n(\mathcal{F}) := \mathbb{E}[R_n(\mathcal{F} \mid X_1^n)],$$

the expectation of the empirical Rademacher complexities.
If $f : \mathcal{X} \to [b_0, b_1]$ for all $f \in \mathcal{F}$, then the Rademacher complexity satisfies bounded differences, because for any two sequences x_1^n and z_1^n differing in only element j, we have

$$n|R_n(\mathcal{F} \mid x_1^n) - R_n(\mathcal{F} \mid z_1^n)| \le \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \varepsilon_i(f(x_i) - f(z_i))\right] = \mathbb{E}[\sup_{f\in\mathcal{F}} \varepsilon_i(f(x_j) - f(z_j))] \le b_1 - b_0.$$

Consequently, the empirical Rademacher complexity satisfies $R_n(\mathcal{F} \mid X_1^n) - R_n(\mathcal{F})$ is $\frac{(b_1-b_0)^2}{4n}$-sub-Gaussian by Theorem 4.2.3. $\diamond$

These examples warrant more discussion, and it is possible to argue that many variants of these random variables are well-concentrated. For example, instead of functions we may simply consider an arbitrary set $\mathcal{A} \subset \mathbb{R}^n$ and define the random variable

$$Z(\mathcal{A}) := \sup_{a\in\mathcal{A}} \langle a, \varepsilon\rangle = \sup_{a\in\mathcal{A}} \sum_{i=1}^n a_i\varepsilon_i.$$

As a function of the random signs ε_i, we may write $Z(\mathcal{A}) = f(\varepsilon)$, and this is then a function satisfying $|f(\varepsilon) - f(\varepsilon')| \le \sup_{a\in\mathcal{A}} |\langle a, \varepsilon - \varepsilon'\rangle|$, so that if ε and ε' differ in index i, we have $|f(\varepsilon) - f(\varepsilon')| \le 2\sup_{a\in\mathcal{A}} |a_i|$. That is, $Z(\mathcal{A}) - \mathbb{E}[Z(\mathcal{A})]$ is $\sum_{i=1}^n \sup_{a\in\mathcal{A}} |a_i|^2$-sub-Gaussian.

Example 4.2.8 (Rademacher complexity as a random vector)**:** This view of Rademacher complexity shows how we may think of Rademacher complexities as norms on certain spaces. Indeed, if we consider a vector space $\mathcal{L}$ of linear functions on $\mathcal{F}$, then we can define the $\mathcal{F}$-seminorm on $\mathcal{L}$ by $\|L\|_\mathcal{F} := \sup_{f\in\mathcal{F}} |L(f)|$. In this case, we may consider the symmetrized empirical distributions

$$P_n^0 := \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{1}_{X_i} \quad f \mapsto P_n^0 f := \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)$$

as elements of this vector space $\mathcal{L}$. (Here we have used $\mathbf{1}_{X_i}$ to denote the point mass at X_i.) Then the Rademacher complexity is nothing more than the expected norm of P_n^0, a random vector, as in Example 4.2.6. This view is somewhat sophisticated, but it shows that any general results we may prove about random vectors, as in Example 4.2.6, will carry over immediately to versions of the Rademacher complexity. $\diamond$

4.3 Uniformity and metric entropy

Now that we have explored a variety of concentration inequalities, we show how to put them to use in demonstrating that a variety of estimation, learning, and other types of procedures have nice convergence properties. We first give a somewhat general collection of results, then delve deeper by focusing on some standard tasks from machine learning.

4.3.1 Symmetrization and uniform laws

The first set of results we consider are *uniform laws of large numbers*, where the goal is to bound means uniformly over different classes of functions. Frequently, such results are called *Glivenko-Cantelli* laws, after the original Glivenko-Cantelli theorem, which shows that empirical distributions uniformly converge. We revisit these ideas in the next chapter, where we present a number of more advanced techniques based on ideas of metric entropy (or volume-like considerations); here we present the basic ideas using our stability and bounded differencing tools.

The starting point is to define what we mean by a uniform law of large numbers. To do so, we adopt notation (as in Example 4.2.8) we will use throughout the remainder of the book, reminding readers as we go. For a sample $X_1, \ldots, X_n$ on a space $\mathcal{X}$, we let

$$P_n := \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i}$$

denote the empirical distribution on $\{X_i\}_{i=1}^n$, where $\mathbf{1}_{X_i}$ denotes the point mass at X_i. Then for functions $f : \mathcal{X} \to \mathbb{R}$ (or more generally, any function f defined on $\mathcal{X}$), we let

$$P_n f := \mathbb{E}_{P_n}[f(X)] = \frac{1}{n} \sum_{i=1}^n f(X_i)$$

denote the empirical expectation of f evaluated on the sample, and we also let

$$Pf := \mathbb{E}_P[f(X)] = \int f(x) dP(x)$$

denote general expectations under a measure P. With this notation, we study *uniform laws of large numbers*, which consist of proving results of the form

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \to 0, \tag{4.3.1}$$

where convergence is in probability, expectation, almost surely, or with rates of convergence. When we view P_n and P as (infinite-dimensional) vectors on the space of maps from $\mathcal{F} \to \mathbb{R}$, then we may define the (semi)norm $\|\cdot\|_{\mathcal{F}}$ for any $L : \mathcal{F} \to \mathbb{R}$ by

$$\|L\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |L(f)|,$$

in which case Eq. (4.3.1) is equivalent to proving

$$\|P_n - P\|_{\mathcal{F}} \to 0.$$

Thus, roughly, we are simply asking questions about when random vectors converge to their expectations.[1]

The starting point of this investigation considers bounded random functions, that is, $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [a, b]$ for some $-\infty < a \le b < \infty$. In this case, the bounded differences inequality (Proposition 4.2.5) immediately implies that expectations of $\|P_n - P\|_{\mathcal{F}}$ provide strong guarantees on concentration of $\|P_n - P\|_{\mathcal{F}}$.

[1]Some readers may worry about measurability issues here. All of our applications will be in separable spaces, so that we may take suprema with abandon without worrying about measurability, and consequently we ignore this from now on.

Proposition 4.3.1. *Let $\mathcal{F}$ be as above. Then*

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{F}} \geq \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] + t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right) \quad \text{for } t \geq 0.$$

Proof Let P_n and P_n' be two empirical distributions, differing only in observation i (with X_i and X_i'). We observe that

$$\begin{aligned}\sup_{f\in\mathcal{F}} |P_n f - Pf| - \sup_{f\in\mathcal{F}} |P_n' f - Pf| &\leq \sup_{f\in\mathcal{F}} \left\{|P_n f - Pf| - |P_n' f - Pf|\right\} \\ &\leq \frac{1}{n} \sup_{f\in\mathcal{F}} |f(X_i) - f(X_i')| \leq \frac{b-a}{n}\end{aligned}$$

by the triangle inequality. An entirely parallel argument gives the converse lower bound of $-\frac{b-a}{n}$, and thus Proposition 4.2.5 gives the result. □

Proposition 4.3.1 shows that, to provide control over high-probability concentration of $\|P_n - P\|_{\mathcal{F}}$, it is (at least in cases where $\mathcal{F}$ is bounded) sufficient to control the expectation $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]$. We take this approach through the remainder of this section, developing tools to simplify bounding this quantity.

Our starting points consist of a few inequalities relating expectations to *symmetrized* quantities, which are frequently easier to control than their non-symmetrized parts. This symmetrization technique is widely used in probability theory, theoretical statistics, and machine learning. The key is that for centered random variables, symmetrized quantities have, to within numerical constants, similar expectations to their non-symmetrized counterparts. Thus, in many cases, it is equivalent to analyze the symmetized quantity and the initial quantity.

Proposition 4.3.2. *Let X_i be independent random vectors on a (Banach) space with norm $\|\cdot\|$ and let $\varepsilon_i\{-1, 1\}$ be independent random signs. Then for any $p \geq 1$,*

$$2^{-p}\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i(X_i - \mathbb{E}[X_i])\right\|^p\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\|^p\right] \leq 2^p\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i X_i\right\|^p\right]$$

In the proof of the upper bound, we could also show the bound

$$\mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\|^p\right] \leq 2^p\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i(X_i - \mathbb{E}[X_i])\right\|^p\right],$$

so we may analyze whichever is more convenient.

Proof We prove the right bound first. We introduce independent copies of the X_i and use these to symmetrize the quantity. Indeed, let X_i' be an independent copy of X_i, and use Jensen's inequality and the convexity of $\|\cdot\|^p$ to observe that

$$\mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\|^p\right] = \mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i'])\right\|^p\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - X_i')\right\|^p\right].$$

Now, note that the distribution of $X_i - X_i'$ is symmetric, so that $X_i - X_i' \stackrel{\text{dist}}{=} \varepsilon_i(X_i - X_i')$, and thus

$$\mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\|^p\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i(X_i - X_i')\right\|^p\right].$$

Multiplying and dividing by 2^p, Jensen's inequality then gives

$$\mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\|^p\right] \le 2^p \mathbb{E}\left[\left\|\frac{1}{2}\sum_{i=1}^n \varepsilon_i (X_i - X_i')\right\|^p\right]$$
$$\le 2^{p-1}\left[\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i X_i\right\|^p\right] + \mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i X_i'\right\|^p\right]\right]$$

as desired.

For the left bound in the proposition, let $Y_i = X_i - \mathbb{E}[X_i]$ be the centered version of the random variables. We break the sum over random variables into two parts, conditional on whether $\varepsilon_i = \pm 1$, using repeated conditioning. We have

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i Y_i\right\|^p\right] &= \mathbb{E}\left[\left\|\sum_{i:\varepsilon_i=1} Y_i - \sum_{i:\varepsilon=-1} Y_i\right\|^p\right] \\
&\le \mathbb{E}\left[2^{p-1}\mathbb{E}\left[\left\|\sum_{i:\varepsilon_i=1} Y_i\right\|^p \mid \varepsilon\right] + 2^{p-1}\mathbb{E}\left[\left\|\sum_{i:\varepsilon_i-1} Y_i\right\|^p \mid \varepsilon\right]\right] \\
&= 2^{p-1}\mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{i:\varepsilon_i=1} Y_i + \sum_{i:\varepsilon_i=-1}\mathbb{E}[Y_i]\right\|^p \mid \varepsilon\right] + \mathbb{E}\left[\left\|\sum_{i:\varepsilon_i=-1} Y_i + \sum_{i:\varepsilon_i=1}\mathbb{E}[Y_i]\right\|^p \mid \varepsilon\right]\right] \\
&\le 2^{p-1}\mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{i:\varepsilon_i=1} Y_i + \sum_{i:\varepsilon_i=-1} Y_i\right\|^p \mid \varepsilon\right] + \mathbb{E}\left[\left\|\sum_{i:\varepsilon_i=-1} Y_i + \sum_{i:\varepsilon_i=1} Y_i\right\|^p \mid \varepsilon\right]\right] \\
&= 2^p \mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^p\right].
\end{aligned}$$

□

We obtain as an immediate corollary a symmetrization bound for supremum norms on function spaces. In this corollary, we use the symmetrized empirical measure

$$P_n^0 := \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{1}_{X_i}, \quad P_n^0 f = \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i).$$

The expectation of $\left\|P_n^0\right\|_{\mathcal{F}}$ is of course the Rademacher complexity (Examples 4.2.7 and 4.2.8), and we have the following corollary.

Corollary 4.3.3. *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ and X_i be i.i.d. Then $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}] \le 2\mathbb{E}[\|P_n^0\|_{\mathcal{F}}]$.*

From Corollary 4.3.3, it is evident that by controlling the *expectation* of the symmetrized process $\mathbb{E}[\|P_n^0\|_{\mathcal{F}}]$ we can derive concentration inequalities and uniform laws of large numbers. For example, we immediately obtain that

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{F}} \ge 2\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] + t\right) \le \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

for all $t \ge 0$ whenever $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [a, b]$.

There are numerous examples of uniform laws of large numbers, many of which reduce to developing bounds on the expectation $\mathbb{E}[\|P_n^0\|_{\mathcal{F}}]$, which is frequently possible via more advanced techniques we develop in Chapter 6. A frequent application of these symmetrization ideas is to risk minimization problems, as we discuss in the coming section; for these, it will be useful for us to develop a few analytic and calculus tools. To better match the development of these ideas, we return to the notation of Rademacher complexities, so that $R_n(\mathcal{F}) := \mathbb{E}[\left\|P_n^0\right\|_{\mathcal{F}}]$. The first is a standard result, which we state for its historical value and the simplicity of its proof.

Proposition 4.3.4 (Massart's finite class bound)**.** *Let $\mathcal{F}$ be any collection of functions with $f : \mathcal{X} \to \mathbb{R}$, and assume that $\sigma_n^2 := n^{-1}\mathbb{E}[\max_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)^2] < \infty$. Then*

$$R_n(\mathcal{F}) \le \frac{\sqrt{2\sigma_n^2 \log |\mathcal{F}|}}{\sqrt{n}}.$$

Proof For each fixed x_1^n, the random variable $\sum_{i=1}^n \varepsilon_i f(x_i)$ is $\sum_{i=1}^n f(x_i)^2$-sub-Gaussian. Now, define $\sigma^2(x_1^n) := n^{-1} \max_{f\in\mathcal{F}} \sum_{i=1}^n f(x_i)^2$. Using the results of Exercise 4.7, that is, that $\mathbb{E}[\max_{j\le n} Z_j] \le \sqrt{2\sigma^2 \log n}$ if the Z_j are each σ^2-sub-Gaussian, we see that

$$R_n(\mathcal{F} \mid x_1^n) \le \frac{\sqrt{2\sigma^2(x_1^n) \log |\mathcal{F}|}}{\sqrt{n}}.$$

Jensen's inequality that $\mathbb{E}[\sqrt{\cdot}] \le \sqrt{\mathbb{E}[\cdot]}$ gives the result. □

A refinement of Massart's finite class bound applies when the classes are infinite but, on a collection $X_1, \ldots, X_n$, the functions $f \in \mathcal{F}$ may take on only a (smaller) number of values. In this case, we define the *empirical shatter coefficient* of a collection of points $x_1, \ldots, x_n$ by $\mathsf{S}_{\mathcal{F}}(x_1^n) := \text{card}\{(f(x_1), \ldots, f(x_n)) \mid f \in \mathcal{F}\}$, the number of distinct vectors of values $(f(x_1), \ldots, f(x_n))$ the functions $f \in \mathcal{F}$ may take. The *shatter coefficient* is the maximum of the empirical shatter coefficients over $x_1^n \in \mathcal{X}^n$, that is, $\mathsf{S}_{\mathcal{F}}(n) := \sup_{x_1^n} \mathsf{S}_{\mathcal{F}}(x_1^n)$. It is clear that $\mathsf{S}_{\mathcal{F}}(n) \le |\mathcal{F}|$ always, but by only counting distinct values, we have the following corollary.

Corollary 4.3.5 (A sharper variant of Massart's finite class bound)**.** *Let $\mathcal{F}$ be any collection of functions with $f : \mathcal{X} \to \mathbb{R}$, and assume that $\sigma_n^2 := n^{-1}\mathbb{E}[\max_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)^2] < \infty$. Then*

$$R_n(\mathcal{F}) \le \frac{\sqrt{2\sigma_n^2 \log \mathsf{S}_{\mathcal{F}}(n)}}{\sqrt{n}}.$$

Typical classes with small shatter coefficients include Vapnik-Chervonenkis classes of functions; we do not discuss these further here, instead referring to one of the many books in machine learning and empirical process theory in statistics.

The most important of the calculus rules we use are the *comparison inequalities* for Rademacher sums, which allow us to consider compositions of function classes and maintain small complexity measurers. We state the rule here; the proof is complex, so we defer it to Section 4.5.3

Theorem 4.3.6 (Ledoux-Talagrand Contraction)**.** *Let $T \subset \mathbb{R}^n$ be an arbitrary set and let $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz and satisfy $\phi_i(0) = 0$. Then for any nondecreasing convex function $\Phi : \mathbb{R} \to \mathbb{R}_+$,*

$$\mathbb{E}\left[\Phi\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{i=1}^n \phi_i(t_i)\varepsilon_i\right|\right)\right] \le \mathbb{E}\left[\Phi\left(\sup_{t\in T}\langle t, \varepsilon\rangle\right)\right].$$

A corollary to this theorem is suggestive of its power and applicability. Let $\phi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz, and for a function class $\mathcal{F}$ define $\phi \circ \mathcal{F} = \{\phi \circ f \mid f \in \mathcal{F}\}$. Then we have the following corollary about Rademacher complexities of contractive mappings.

Corollary 4.3.7. *Let $\mathcal{F}$ be an arbitrary function class and ϕ be L-Lipschitz. Then*

$$R_n(\phi \circ \mathcal{F}) \le 2LR_n(\mathcal{F}) + |\phi(0)|/\sqrt{n}.$$

Proof The result is an almost immediate consequence of Theorem 4.3.6; we simply recenter our functions. Indeed, we have

$$\begin{aligned} R_n(\phi \circ \mathcal{F} \mid x_1^n) &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\phi(f(x_i)) - \phi(0)) + \frac{1}{n}\sum_{i=1}^n \varepsilon_i \phi(0)\right|\right] \\ &\le \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\phi(f(x_i)) - \phi(0))\right|\right] + \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \phi(0)\right|\right] \\ &\le 2LR_n(\mathcal{F}) + \frac{|\phi(0)|}{\sqrt{n}}, \end{aligned}$$

where the final inequality follows by Theorem 4.3.6 (as $g(\cdot) = \phi(\cdot) - \phi(0)$ is Lipschitz and satisfies $g(0) = 0$) and that $\mathbb{E}[|\sum_{i=1}^n \varepsilon_i|] \le \sqrt{n}$. □

4.3.2 Metric entropy, coverings, and packings

When the class of functions $\mathcal{F}$ under consideration is finite, the union bound more or less provides guarantees that $P_n f$ is uniformly close to Pf for all $f \in \mathcal{F}$. When $\mathcal{F}$ is infinite, however, we require a different set of tools for addressing uniform laws. In many cases, because of the application of the bounded differences inequality in Proposition 4.3.1, all we really need to do is to control the expectation $\mathbb{E}[\|P_n^0\|_{\mathcal{F}}]$, though the techniques we develop here will have broader use and can sometimes directly guarantee concentration.

The basic object we wish to control is a measure of the size of the space on which we work. To that end, we modify notation a bit to simply consider arbitrary vectors $\theta \in \Theta$, where Θ is a non-empty set with an associated (semi)metric ρ. For many purposes in estimation (and in our optimality results in the further parts of the book), a natural way to measure the size of the set is via the number of balls of a fixed radius $\delta > 0$ required to cover it.

Definition 4.7 (Covering number)**.** *Let Θ be a set with (semi)metric ρ. A* δ-cover *of the set Θ with respect to ρ is a set $\{\theta_1, \ldots, \theta_N\}$ such that for any point $\theta \in \Theta$, there exists some $v \in \{1, \ldots, N\}$ such that $\rho(\theta, \theta_v) \le \delta$. The* δ-covering number *of Θ is*

$$N(\delta, \Theta, \rho) := \inf\left\{N \in \mathbb{N} \;:\; \text{there exists a } \delta\text{-cover } \theta_1, \ldots, \theta_N \text{ of } \Theta\right\}.$$

The *metric entropy* of the set Θ is simply the logarithm of its covering number $\log N(\delta, \Theta, \rho)$. We can define a related measure—more useful for constructing our lower bounds—of size that relates to the number of disjoint balls of radius $\delta > 0$ that can be placed into the set Θ.

Definition 4.8 (Packing number)**.** *A* δ-packing *of the set Θ with respect to ρ is a set $\{\theta_1, \ldots, \theta_M\}$ such that for all distinct $v, v' \in \{1, \ldots, M\}$, we have $\rho(\theta_v, \theta_v') \ge \delta$. The* δ-packing number *of Θ is*

$$M(\delta, \Theta, \rho) := \sup\left\{M \in \mathbb{N} \;:\; \text{there exists a } \delta\text{-packing } \theta_1, \ldots, \theta_M \text{ of } \Theta\right\}.$$

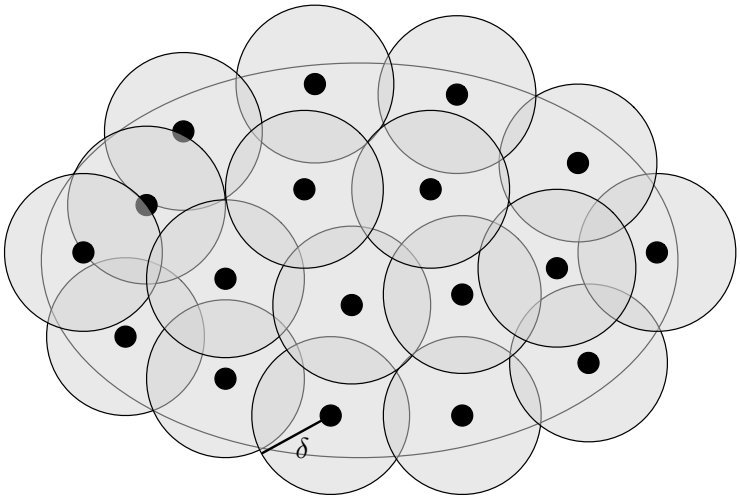

Figure 4.1. A δ-covering of the elliptical set by balls of radius δ.

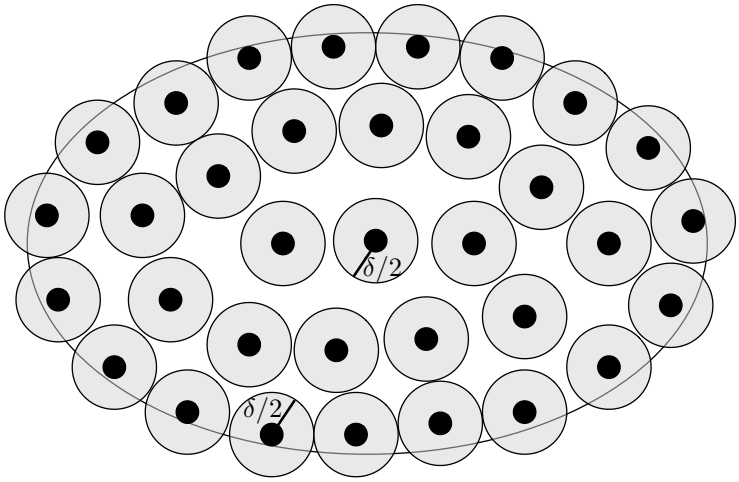

Figure 4.2. A δ-packing of the elliptical set, where balls have radius $\delta/2$. No balls overlap, and each center of the packing satisfies $\|\theta_v - \theta_{v'}\| \geq \delta$.

Figures 4.1 and 4.2 give examples of (respectively) a covering and a packing of the same set.

An exercise in proof by contradiction shows that the packing and covering numbers of a set are in fact closely related:

Lemma 4.3.8. *The packing and covering numbers satisfy the following inequalities:*

$$M(2\delta, \Theta, \rho) \leq N(\delta, \Theta, \rho) \leq M(\delta, \Theta, \rho).$$

We leave derivation of this lemma to Exercise 4.11, noting that it shows that (up to constant factors) packing and covering numbers have the same scaling in the radius δ. As a simple example, we see for any interval $[a, b]$ on the real line that in the usual absolute distance metric, $N(\delta, [a, b], |\cdot|) \asymp (b - a)/\delta$.

As one example of the metric entropy, consider a set of functions $\mathcal{F}$ with reasonable covering numbers (metric entropy) in $\|\cdot\|_\infty$-norm.

Example 4.3.9 (The "standard" covering number guarantee)**:** Let $\mathcal{F}$ consist of functions $f : \mathcal{X} \to [-b, b]$ and let the metric ρ be $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)|$. Then

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |P_n f - Pf| \geq t\right) \leq \exp\left(-\frac{nt^2}{18b^2} + \log N(t/3, \mathcal{F}, \|\cdot\|_\infty)\right). \tag{4.3.2}$$

So as long as the covering numbers $N(t, \mathcal{F}, \|\cdot\|_\infty)$ grow sub-exponentially in t—so that $\log N(t) \ll nt^2$—we have the (essentially) sub-Gaussian tail bound (4.3.2). Example 4.4.11 gives one typical case. Indeed, fix a minimal $t/3$-cover of $\mathcal{F}$ in $\|\cdot\|_\infty$ of size $N := N(t/3, \mathcal{F}, \|\cdot\|_\infty)$, calling the covering functions $f_1, \ldots, f_N$. Then for any $f \in \mathcal{F}$ and the function f_i satisfying $\|f - f_i\|_\infty \le t/2$, we have

$$|P_n f - Pf| \le |P_n f - P_n f_i| + |P_n f_i - Pf_i| + |Pf_i - Pf| \le |P_n f_i - Pf_i| + \frac{2t}{3}.$$

The Azuma-Hoeffding inequality (Theorem 4.2.3) guarantees (by a union bound) that

$$\mathbb{P}\left(\max_{i \le N} |P_n f_i - Pf_i| \ge t\right) \le \exp\left(-\frac{nt^2}{2b^2} + \log N\right).$$

Combine this bound (replacing t with $t/3$) to obtain inequality (4.3.2). $\diamond$

Given the relationships between packing, covering, and size of sets Θ, we would expect there to be relationships between volume, packing, and covering numbers. This is indeed the case, as we now demonstrate for arbitrary norm balls in finite dimensions.

Lemma 4.3.10. *Let $\mathbb{B}$ denote the unit $\|\cdot\|$-ball in $\mathbb{R}^d$. Then*

$$\left(\frac{1}{\delta}\right)^d \le N(\delta, \mathbb{B}, \|\cdot\|) \le \left(1 + \frac{2}{\delta}\right)^d.$$

Proof We prove the lemma via a volumetric argument. For the lower bound, note that if the points $v_1, \ldots, v_N$ are a δ-cover of $\mathbb{B}$, then

$$\mathrm{Vol}(\mathbb{B}) \le \sum_{i=1}^N \mathrm{Vol}(\delta\mathbb{B} + v_i) = N \,\mathrm{Vol}(\delta\mathbb{B}) = N \,\mathrm{Vol}(\mathbb{B})\delta^d.$$

In particular, $N \ge \delta^{-d}$. For the upper bound on $N(\delta, \mathbb{B}, \|\cdot\|)$, let $\mathcal{V}$ be a δ-packing of $\mathbb{B}$ with maximal cardinality, so that $|\mathcal{V}| = M(\delta, \mathbb{B}, \|\cdot\|) \ge N(\delta, \mathbb{B}, \|\cdot\|)$ (recall Lemma 4.3.8). Notably, the collection of δ-balls $\{\delta\mathbb{B} + v_i\}_{i=1}^M$ cover the ball $\mathbb{B}$ (as otherwise, we could put an additional element in the packing $\mathcal{V}$), and moreover, the balls $\{\frac{\delta}{2}\mathbb{B} + v_i\}$ are all disjoint by definition of a packing. Consequently, we find that

$$M\left(\frac{\delta}{2}\right)^d \mathrm{Vol}(\mathbb{B}) = M \,\mathrm{Vol}\left(\frac{\delta}{2}\mathbb{B}\right) \le \mathrm{Vol}\left(\mathbb{B} + \frac{\delta}{2}\mathbb{B}\right) = \left(1 + \frac{\delta}{2}\right)^d \mathrm{Vol}(\mathbb{B}).$$

Rewriting, we obtain

$$M(\delta, \mathbb{B}, \|\cdot\|) \le \left(\frac{2}{\delta}\right)^d \left(1 + \frac{\delta}{2}\right)^d \frac{\mathrm{Vol}(\mathbb{B})}{\mathrm{Vol}(\mathbb{B})} = \left(1 + \frac{2}{\delta}\right)^d,$$

completing the proof. $\square$

Let us give one application of Lemma 4.3.10 to concentration of random matrices; we explore more in the exercises as well. We can generalize the definition of sub-Gaussian random variables to *sub-Gaussian random vectors*, where we say that $X \in \mathbb{R}^d$ is a σ^2-sub-Gaussian vector if

$$\mathbb{E}[\exp(\langle u, X - \mathbb{E}[X]\rangle)] \le \exp\left(\frac{\sigma^2}{2}\|u\|_2^2\right) \tag{4.3.3}$$

for all $u \in \mathbb{R}^d$. For example, $X \sim \mathsf{N}(0, I_d)$ is immediately 1-sub-Gaussian, and $X \in [-b, b]^d$ with independent entries is b^2-sub-Gaussian. Now, suppose that X_i are independent isotropic random vectors, meaning that $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i X_i^\top] = I_d$, and that they are also σ^2-sub-Gaussian. Then by an application of Lemma 4.3.10, we can give concentration guarantees for the sample covariance $\Sigma_n := \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ for the operator norm $\|A\|_{\rm op} := \sup\{\langle u, Av\rangle \mid \|u\|_2 = \|v\|_2 = 1\}$.

Proposition 4.3.11. *Let X_i be independent isotropic and σ^2-sub-Gaussian vectors. Then there is a numerical constant C such that the sample covariance $\Sigma_n := \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ satisfies*

$$\|\Sigma_n - I_d\|_{\rm op} \le C\sigma^2 \left[\frac{d + \log\frac{1}{\delta}}{n} + \sqrt{\frac{d + \log\frac{1}{\delta}}{n}}\right]$$

with probability at least $1 - \delta$.

Proof We begin with an intermediate lemma.

Lemma 4.3.12. *Let A be symmetric and $\{u_i\}_{i=1}^N$ be an ϵ-cover of the unit ℓ_2 ball $\mathbb{B}_2^d$. Then*

$$(1 - 2\epsilon)\|A\|_{\rm op} \le \max_{i \le N} \langle u_i, Au_i\rangle \le \|A\|_{\rm op}.$$

Proof The second inequality is trivial. Fix any $u \in \mathbb{B}_2^d$. Then for the i such that $\|u - u_i\|_2 \le \epsilon$, we have

$$\langle u, Au\rangle = \langle u - u_i, Au\rangle + \langle u_i, Au\rangle = 2\langle u - u_i, Au\rangle + \langle u_i, Au_i\rangle \le 2\epsilon\|A\|_{\rm op} + \langle u_i, Au_i\rangle$$

by definition of the operator norm. Taking a supremum over u gives the final result. □

Let the matrix $E_i = X_i X_i^\top - I$, and define the average error $\overline{E}_n = \frac{1}{n}E_i$. Then with this lemma in hand, we see that for any ϵ-cover $\mathcal{N}$ of the ℓ_2-ball $\mathbb{B}_2^d$,

$$(1 - 2\epsilon)\left\|\overline{E}_n\right\|_{\rm op} \le \max_{u \in \mathcal{N}} \langle u, \overline{E}_n u\rangle.$$

Now, note that $\langle u, E_i u\rangle = \langle u, X_i\rangle^2 - \|u\|_2^2$ is sub-exponential, as it is certainly mean 0 and, moreover, is the square of a sub-Gaussian; in particular, Theorem 4.1.15 shows that there is a numerical constant $C < \infty$ such that

$$\mathbb{E}[\exp(\lambda\langle u, E_i u\rangle)] \le \exp\left(C\lambda^2\sigma^4\right) \quad \text{for} \quad |\lambda| \le \frac{1}{C\sigma^2}.$$

Taking $\epsilon = \frac{1}{4}$ in our covering $\mathcal{N}$, then,

$$\mathbb{P}(\left\|\overline{E}_n\right\|_{\rm op} \ge t) \le \mathbb{P}\left(\max_{u \in \mathcal{N}} \langle u, \overline{E}_n u\rangle \ge t/2\right) \le |\mathcal{N}| \cdot \max_{u \in \mathcal{N}} \mathbb{P}\left(\langle u, n\overline{E}_n u\rangle \ge nt/2\right)$$

by a union bound. As sums of sub-exponential random variable remain sub-exponential, Corollary 4.1.18 implies

$$\mathbb{P}\left(\left\|\overline{E}_n\right\|_{\rm op} \ge t\right) \le |\mathcal{N}| \exp\left(-c\min\left\{\frac{nt^2}{\sigma^4}, \frac{nt}{\sigma^2}\right\}\right),$$

where $c > 0$ is a numerical constant. Finally, we apply Lemma 4.3.10, which guarantees that $|\mathcal{N}| \le 9^d$, and then take t to scale as the maximum of $\sigma^2\frac{d + \log\frac{1}{\delta}}{n}$ and $\sigma^2\sqrt{\frac{d + \log\frac{1}{\delta}}{n}}$. □

4.4 Generalization bounds

We now build off of our ideas on uniform laws of large numbers and Rademacher complexities to demonstrate their applications in statistical machine learning problems, focusing on *empirical risk minimization* procedures and related problems. We consider a setting as follows: we have a sample $Z_1, \ldots, Z_n \in \mathcal{Z}$ drawn i.i.d. according to some (unknown) distribution P, and we have a collection of functions $\mathcal{F}$ from which we wish to select an f that "fits" the data well, according to some loss measure $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$. That is, we wish to find a function $f \in \mathcal{F}$ minimizing the *risk*

$$L(f) := \mathbb{E}_P[\ell(f, Z)]. \tag{4.4.1}$$

In general, however, we only have access to the risk via the empirical distribution of the Z_i, and we often choose f by minimizing the empirical risk

$$\widehat{L}_n(f) := \frac{1}{n} \sum_{i=1}^n \ell(f, Z_i). \tag{4.4.2}$$

As written, this formulation is quite abstract, so we provide a few examples to make it somewhat more concrete.

Example 4.4.1 (Binary classification problems)**:** One standard problem—still abstract—that motivates the formulation (4.4.1) is the *binary classification problem*. Here the data Z_i come in pairs (X, Y), where $X \in \mathcal{X}$ is some set of covariates (independent variables) and $Y \in \{-1, 1\}$ is the label of example X. The function class $\mathcal{F}$ consists of functions $f : \mathcal{X} \to \mathbb{R}$, and the goal is to find a function f such that

$$\mathbb{P}(\mathrm{sign}(f(X)) \neq Y)$$

is small, that is, minimizing the risk $\mathbb{E}[\ell(f, Z)]$ where the loss is the 0-1 loss, $\ell(f, (x, y)) = \mathbf{1}\{f(x)y \leq 0\}$. $\diamond$

Example 4.4.2 (Multiclass classification)**:** The multiclass classifcation problem is identical to the binary problem, but instead of $Y \in \{-1, 1\}$ we assume that $Y \in [k] = \{1, \ldots, k\}$ for some $k \geq 2$, and the function class $\mathcal{F}$ consists of (a subset of) functions $f : \mathcal{X} \to \mathbb{R}^k$. The goal is to find a function f such that, if $Y = y$ is the correct label for a datapoint x, then $f_y(x) > f_l(x)$ for all $l \neq y$. That is, we wish to find $f \in \mathcal{F}$ minimizing

$$\mathbb{P}\left(\exists\ l \neq Y \text{ such that } f_l(X) \geq f_Y(X)\right).$$

In this case, the loss function is the zero-one loss $\ell(f, (x, y)) = \mathbf{1}\{\max_{l \neq y} f_l(x) \geq f_y(x)\}$. $\diamond$

Example 4.4.3 (Binary classification with linear functions)**:** In the standard statistical learning setting, the data x belong to $\mathbb{R}^d$, and we assume that our function class $\mathcal{F}$ is indexed by a set $\Theta \subset \mathbb{R}^d$, so that $\mathcal{F} = \{f_\theta : f_\theta(x) = \theta^\top x, \theta \in \Theta\}$. In this case, we may use the zero-one loss, the convex hinge loss, or the (convex) logistic loss, which are variously $\ell_{\mathrm{zo}}(f_\theta, (x, y)) := \mathbf{1}\left\{y\theta^\top x \leq 0\right\}$, and the convex losses

$$\ell_{\mathrm{hinge}}(f_\theta, (x, y)) = \left[1 - yx^\top\theta\right]_+ \quad \text{and} \quad \ell_{\mathrm{logit}}(f_\theta, (x, y)) = \log(1 + \exp(-yx^\top\theta)).$$

The hinge and logistic losses, as they are convex, are substantially computationally easier to work with, and they are common choices in applications. $\diamond$

The main motivating question that we ask is the following: given a sample $Z_1, \ldots, Z_n$, if we choose some $\widehat{f}_n \in \mathcal{F}$ based on this sample, can we guarantee that it generalizes to unseen data? In particular, can we guarantee that (with high probability) we have the empirical risk bound

$$\widehat{L}_n(\widehat{f}_n) = \frac{1}{n} \sum_{i=1}^n \ell(\widehat{f}_n, Z_i) \le R(\widehat{f}_n) + \epsilon \tag{4.4.3}$$

for some small ϵ? If we allow $\widehat{f}_n$ to be arbitrary, then this becomes clearly impossible: consider the classification example 4.4.1, and set $\widehat{f}_n$ to be the "hash" function that sets $\widehat{f}_n(x) = y$ if the pair (x, y) was in the sample, and otherwise $\widehat{f}_n(x) = -1$. Then clearly $\widehat{L}_n(\widehat{f}_n) = 0$, while there is no useful bound on $R(\widehat{f}_n)$.

4.4.1 Finite and countable classes of functions

In order to get bounds of the form (4.4.3), we require a few assumptions that are not too onerous. First, throughout this section, we will assume that for any fixed function f, the loss $\ell(f, Z)$ is σ^2-sub-Gaussian, that is,

$$\mathbb{E}_P \left[\exp \left(\lambda (\ell(f, Z) - L(f)) \right) \right] \le \exp \left(\frac{\lambda^2 \sigma^2}{2} \right) \tag{4.4.4}$$

for all $f \in \mathcal{F}$. (Recall that the risk functional $L(f) = \mathbb{E}_P[\ell(f, Z)]$.) For example, if the loss is the zero-one loss from classification problems, inequality (4.4.4) is satisfied with $\sigma^2 = \frac{1}{4}$ by Hoeffding's lemma. In order to guarantee a bound of the form (4.4.4) for a function $\widehat{f}$ chosen dependent on the data, in this section we give uniform bounds, that is, we would like to bound

$$\mathbb{P} \left(\text{there exists } f \in \mathcal{F} \text{ s.t. } L(f) > \widehat{L}_n(f) + t \right) \quad \text{or} \quad \mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \widehat{L}_n(f) - R(f) \right| > t \right).$$

Such uniform bounds are certainly sufficient to guarantee that the empirical risk is a good proxy for the true risk L, even when $\widehat{f}_n$ is chosen based on the data.

Now, recalling that our set of functions or predictors $\mathcal{F}$ is finite or countable, let us suppose that for each $f \in \mathcal{F}$, we have a complexity measure $c(f)$—a penalty—such that

$$\sum_{f \in \mathcal{F}} e^{-c(f)} \le 1. \tag{4.4.5}$$

This inequality should look familiar to the Kraft inequality—which we will see in the coming chapters—from coding theory. As soon as we have such a penalty function, however, we have the following result.

Theorem 4.4.4. *Let the loss ℓ, distribution P on $\mathcal{Z}$, and function class $\mathcal{F}$ be such that $\ell(f, Z)$ is σ^2-sub-Gaussian for each $f \in \mathcal{F}$, and assume that the complexity inequality (4.4.5) holds. Then with probability at least $1 - \delta$ over the sample $Z_{1:n}$,*

$$L(f) \le \widehat{L}_n(f) + \sqrt{2\sigma^2 \frac{\log \frac{1}{\delta} + c(f)}{n}} \quad \text{for all } f \in \mathcal{F}.$$

Proof First, we note that by the usual sub-Gaussian concentration inequality (Corollary 4.1.10) we have for any $t \geq 0$ and any $f \in \mathcal{F}$ that

$$\mathbb{P}\left(L(f) \geq \widehat{L}_n(f) + t\right) \leq \exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

Now, if we replace t by $\sqrt{t^2 + 2\sigma^2 c(f)/n}$, we obtain

$$\mathbb{P}\left(L(f) \geq \widehat{L}_n(f) + \sqrt{t^2 + 2\sigma^2 c(f)/n}\right) \leq \exp\left(-\frac{nt^2}{2\sigma^2} - c(f)\right).$$

Then using a union bound, we have

$$\begin{aligned}\mathbb{P}\left(\exists\, f \in \mathcal{F} \text{ s.t. } L(f) \geq \widehat{L}_n(f) + \sqrt{t^2 + 2\sigma^2 c(f)/n}\right) &\leq \sum_{f \in \mathcal{F}} \exp\left(-\frac{nt^2}{2\sigma^2} - c(f)\right) \\ &= \exp\left(-\frac{nt^2}{2\sigma^2}\right) \underbrace{\sum_{f \in \mathcal{F}} \exp(-c(f))}_{\leq 1}.\end{aligned}$$

Setting $t^2 = 2\sigma^2 \log \frac{1}{\delta}/n$ gives the result. □

As one classical example of this setting, suppose that we have a finite class of functions $\mathcal{F}$. Then we can set $c(f) = \log|\mathcal{F}|$, in which case we clearly have the summation guarantee (4.4.5), and we obtain

$$L(f) \leq \widehat{L}_n(f) + \sqrt{2\sigma^2 \frac{\log\frac{1}{\delta} + \log|\mathcal{F}|}{n}} \text{ uniformly for } f \in \mathcal{F}$$

with probability at least $1 - \delta$. To make this even more concrete, consider the following example.

Example 4.4.5 (Floating point classifiers)**:** We implement a linear binary classifier using double-precision floating point values, that is, we have $f_\theta(x) = \theta^\top x$ for all $\theta \in \mathbb{R}^d$ that may be represented using d double-precision floating point numbers. Then for each coordinate of θ, there are at most 2^{64} representable numbers; in total, we must thus have $|\mathcal{F}| \leq 2^{64d}$. Thus, for the zero-one loss $\ell_{\text{zo}}(f_\theta, (x, y)) = \mathbf{1}\left\{\theta^\top xy \leq 0\right\}$, we have

$$L(f_\theta) \leq \widehat{L}_n(f_\theta) + \sqrt{\frac{\log\frac{1}{\delta} + 45d}{2n}}$$

for all representable classifiers simultaneously, with probability at least $1 - \delta$, as the zero-one loss is $1/4$-sub-Gaussian. (Here we have used that $64 \log 2 < 45$.) ◇

We also note in passing that by replacing δ with $\delta/2$ in the bounds of Theorem 4.4.4, a union bound yields the following two-sided corollary.

Corollary 4.4.6. *Under the conditions of Theorem 4.4.4, we have*

$$\left|\widehat{L}_n(f) - L(f)\right| \leq \sqrt{2\sigma^2 \frac{\log\frac{2}{\delta} + c(f)}{n}} \text{ for all } f \in \mathcal{F}$$

with probability at least $1 - \delta$.

4.4.2 Large classes

When the collection of functions is (uncountably) infinite, it can be more challenging to obtain strong generalization bounds, though there still exist numerous tools for these ideas. The most basic, of which we will give examples, leverage covering number bounds (essentially, as in Example 4.3.9). We return in the next chapter to alternative approaches based on randomization and divergence measures, which provide guarantees with somewhat similar structure to those we present here.

Let us begin by considering a few examples, after which we provide examples showing how to derive explicit bounds using Rademacher complexities.

Example 4.4.7 (Rademacher complexity of the ℓ_2-ball)**:** Let $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \le r\}$, and consider the class of linear functionals $\mathcal{F} := \{f_\theta(x) = \theta^T x, \theta \in \Theta\}$. Then

$$R_n(\mathcal{F} \mid x_1^n) \le \frac{r}{n}\sqrt{\sum_{i=1}^n \|x_i\|_2^2},$$

because we have

$$R_n(\mathcal{F} \mid x_1^n) = \frac{r}{n}\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2\right] \le \frac{r}{n}\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2^2\right]} = \frac{r}{n}\sqrt{\sum_{i=1}^n \|x_i\|_2^2},$$

as desired. $\diamond$

In high-dimensional situations, it is sometimes useful to consider more restrictive function classes, for example, those indexed by vectors in an ℓ_1-ball.

Example 4.4.8 (Rademacher complexity of the ℓ_1-ball)**:** In contrast to the previous example, suppose that $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \le r\}$, and consider the linear class $\mathcal{F} := \{f_\theta(x) = \theta^T x, \theta \in \Theta\}$. Then

$$R_n(\mathcal{F} \mid x_1^n) = \frac{r}{n}\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_\infty\right].$$

Now, each coordinate j of $\sum_{i=1}^n \varepsilon_i x_i$ is $\sum_{i=1}^n x_{ij}^2$-sub-Gaussian, and thus using that $\mathbb{E}[\max_{j \le d} Z_j] \le \sqrt{2\sigma^2 \log d}$ for arbitrary σ^2-sub-Gaussian Z_j (see Exercise 4.7), we have

$$R_n(\mathcal{F} \mid x_1^n) \le \frac{r}{n}\sqrt{2\log(2d)\max_j \sum_{i=1}^n x_{ij}^2}.$$

To facilitate comparison with Example 4.4.8, suppose that the vectors x_i all satisfy $\|x_i\|_\infty \le b$. In this case, the preceding inequality implies that $R_n(\mathcal{F} \mid x_1^n) \le rb\sqrt{2\log(2d)}/\sqrt{n}$. In contrast, the ℓ_2-norm of such x_i may satisfy $\|x_i\|_2 = b\sqrt{d}$, so that the bounds of Example 4.4.7 scale instead as $rb\sqrt{d}/\sqrt{n}$, which can be exponentially larger. $\diamond$

These examples are sufficient to derive a few sophisticated risk bounds. We focus on the case where we have a loss function applied to some class with reasonable Rademacher complexity, in

which case it is possible to recenter the loss class and achieve reasonable complexity bounds. The coming proposition does precisely this in the case of margin-based binary classification. Consider points $(x, y) \in \mathcal{X} \times \{\pm 1\}$, and let $\mathcal{F}$ be an arbitrary class of functions $f : \mathcal{X} \to \mathbb{R}$ and $\mathcal{L} = \{(x, y) \mapsto \ell(yf(x))\}_{f \in \mathcal{F}}$ be the induced collection of losses. As a typical example, we might have $\ell(t) = [1 - t]_+$, $\ell(t) = e^{-t}$, or $\ell(t) = \log(1 + e^{-t})$. We have the following proposition.

Proposition 4.4.9. *Let $\mathcal{F}$ and $\mathcal{X}$ be such that $\sup_{x \in \mathcal{X}} |f(x)| \le M$ for $f \in \mathcal{F}$ and assume that ℓ is L-Lipschitz. Define the empirical and population risks $\widehat{L}_n(f) := P_n \ell(Yf(X))$ and $L(f) := P\ell(Yf(X))$. Then*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\widehat{L}_n(f) - L(f)| \ge 4LR_n(\mathcal{F}) + t\right) \le 2 \exp\left(-\frac{nt^2}{2L^2M^2}\right) \quad \text{for } t \ge 0.$$

Proof We may recenter the class $\mathcal{L}$, that is, replace $\ell(\cdot)$ with $\ell(\cdot) - \ell(0)$, without changing $\widehat{L}_n(f) - L(f)$. Call this class $\mathcal{L}_0$, so that $\|P_n - P\|_{\mathcal{L}} = \|P_n - P\|_{\mathcal{L}_0}$. This recentered class satisfies bounded differences with constant $2ML$, as $|\ell(yf(x)) - \ell(y'f(x'))| \le L|yf(x) - y'f(x')| \le 2LM$, as in the proof of Proposition 4.3.1. Applying Proposition 4.3.1 and then Corollary 4.3.3 and gives that $\mathbb{P}(\sup_{f \in \mathcal{F}} |\widehat{L}_n(f) - L(f)| \ge 2R_n(\mathcal{L}_0) + t) \le \exp(-\frac{nt^2}{2M^2L^2})$ for $t \ge 0$. Then applying the contraction inequality (Theorem 4.3.6) yields $R_n(\mathcal{L}_0) \le 2LR_n(\mathcal{F})$, giving the result. □

Let us give a few example applications of these ideas.

Example 4.4.10 (Support vector machines and hinge losses)**:** In the support vector machine problem, we receive data $(X_i, Y_i) \in \mathbb{R}^d \times \{\pm 1\}$, and we seek to minimize average of the losses $\ell(\theta; (x, y)) = \left[1 - y\theta^T x\right]_+$. We assume that the space $\mathcal{X}$ has $\|x\|_2 \le b$ for $x \in \mathcal{X}$ and that $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \le r\}$. Applying Proposition 4.4.9 gives

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |P_n \ell(\theta; (X, Y)) - P\ell(\theta; (X, Y))| \ge 4R_n(\mathcal{F}_\Theta) + t\right) \le \exp\left(-\frac{nt^2}{2r^2b^2}\right),$$

where $\mathcal{F}_\Theta = \{f_\theta(x) = \theta^T x\}_{\theta \in \Theta}$. Now, we apply Example 4.4.7, which implies that

$$R_n(\phi \circ \mathcal{F}_\Theta) \le 2R_n(\mathcal{F}_\theta) \le \frac{2rb}{\sqrt{n}}.$$

That is, we have

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |P_n \ell(\theta; (X, Y)) - P\ell(\theta; (X, Y))| \ge \frac{4rb}{\sqrt{n}} + t\right) \le \exp\left(-\frac{nt^2}{2(rb)^2}\right),$$

so that P_n and P become close at rate roughly $rb/\sqrt{n}$ in this case. ◇

Example 4.4.10 is what is sometimes called a "dimension free" convergence result—there is no esxplicit dependence on the dimension d of the problem, except as the radii r and b make explicit. One consequence of this is that if x and θ instead belong to a Hilbert space (potentiall infinite dimensional) with inner product $\langle \cdot, \cdot \rangle$ and norm $\|x\|^2 = \langle x, x \rangle$, but for which we are guaranteed that $\|\theta\| \le r$ and similarly $\|x\| \le b$, then the result still applies. Extending this to other function classes is reasonably straightforward, and we present a few examples in the exercises.

When we do not have the simplifying structure of $\ell(yf(x))$ identified in the preceding examples, we can still provide guarantees of generalization using the covering number guarantees introduced in Section 4.3.2. The most common and important case is when we have a Lipschitzian loss function in an underlying parameter θ.

Example 4.4.11 (Lipschitz functions over a norm-bounded parameter space)**:** Consider the parametric loss minimization problem

$$\minimize_{\theta \in \Theta} \; L(\theta) := \mathbb{E}[\ell(\theta; Z)]$$

for a loss function ℓ that is M-Lipschitz (with respect to the norm $\|\cdot\|$) in its argument, where for normalization we assume $\inf_{\theta \in \Theta} \ell(\theta, z) = 0$ for each z. Then the metric entropy of Θ bounds the metric entropy of the loss class $\mathcal{F} := \{z \mapsto \ell(\theta, z)\}_{\theta \in \Theta}$ for the supremum norm $\|\cdot\|_\infty$. Indeed, for any pair θ, θ', we have

$$\sup_z |\ell(\theta, z) - \ell(\theta', z)| \le M \left\|\theta - \theta'\right\|,$$

and so an ϵ-cover of Θ is an $M\epsilon$-cover of $\mathcal{F}$ in supremum norm. In particular,

$$N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \le N(\epsilon/M, \Theta, \|\cdot\|).$$

Assume that $\Theta \subset \{\theta \mid \|\theta\| \le b\}$ for some finite b. Then Lemma 4.3.10 guarantees that $\log N(\epsilon, \Theta, \|\cdot\|) \le d \log(1 + 2/\epsilon) \lesssim d \log \frac{1}{\epsilon}$, and so the classical covering number argument in Example 4.3.9 gives

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |P_n \ell(\theta, Z) - P\ell(\theta, Z)| \ge t\right) \le \exp\left(-c \frac{nt^2}{b^2 M^2} + Cd \log \frac{M}{t}\right),$$

where c, C are numerical constants. In particular, taking $t^2 \asymp \frac{M^2 b^2 d}{n} \log \frac{n}{\delta}$ gives that

$$|P_n \ell(\theta, Z) - P\ell(\theta, Z)| \lesssim \frac{Mb\sqrt{d \log \frac{n}{\delta}}}{\sqrt{n}}$$

with probability at least $1 - \delta$. $\diamond$

4.4.3 Structural risk minimization and adaptivity

In general, for a given function class $\mathcal{F}$, we can always decompose the excess risk into the *approximation/estimation* error decomposition. That is, let

$$L^* = \inf_f L(f),$$

where the preceding infimum is taken across *all* (measurable) functions. Then we have

$$L(\widehat{f}_n) - L^* = \underbrace{L(\widehat{f}_n) - \inf_{f \in \mathcal{F}} L(f)}_{\text{estimation}} + \underbrace{\inf_{f \in \mathcal{F}} L(f) - L^*}_{\text{approximation}}. \tag{4.4.6}$$

There is often a tradeoff between these two, analogous to the bias/variance tradeoff in classical statistics; if the approximation error is very small, then it is likely hard to guarantee that the estimation error converges quickly to zero, while certainly a constant function will have low estimation

error, but may have substantial approximation error. With that in mind, we would like to develop procedures that, rather than simply attaining good performance for the class $\mathcal{F}$, are guaranteed to trade-off in an appropriate way between the two types of error. This leads us to the idea of *structural risk minimization*.

In this scenario, we assume we have a sequence of classes of functions, $\mathcal{F}_1, \mathcal{F}_2, \ldots$, of increasing complexity, meaning that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$. For example, in a linear classification setting with vectors $x \in \mathbb{R}^d$, we might take a sequence of classes allowing increasing numbers of non-zeros in the classification vector θ:

$$\mathcal{F}_1 := \left\{f_\theta(x) = \theta^\top x \text{ such that } \|\theta\|_0 \le 1\right\}, \; \mathcal{F}_2 := \left\{f_\theta(x) = \theta^\top x \text{ such that } \|\theta\|_0 \le 2\right\}, \ldots.$$

More broadly, let $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$ be a (possibly infinite) increasing sequence of function classes. We assume that for each $\mathcal{F}_k$ and each $n \in \mathbb{N}$, there exists a constant $C_{n,k}(\delta)$ such that we have the uniform generalization guarantee

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_k} \left|\widehat{L}_n(f) - L(f)\right| \ge C_{n,k}(\delta)\right) \le \delta \cdot 2^{-k}.$$

For example, by Corollary 4.4.6, if $\mathcal{F}$ is finite we may take

$$C_{n,k}(\delta) = \sqrt{2\sigma^2 \frac{\log|\mathcal{F}_k| + \log\frac{1}{\delta} + k\log 2}{n}}.$$

(We will see in subsequent sections of the course how to obtain other more general guarantees.)

We consider the following *structural risk minimization* procedure. First, given the empirical risk $\widehat{L}_n$, we find the model collection $\widehat{k}$ minimizing the penalized risk

$$\widehat{k} := \operatorname*{argmin}_{k\in\mathbb{N}} \left\{\inf_{f\in\mathcal{F}_k} \widehat{L}_n(f) + C_{n,k}(\delta)\right\}. \tag{4.4.7a}$$

We then choose $\widehat{f}$ to minimize the risk over the estimated "best" class $\mathcal{F}_{\widehat{k}}$, that is, set

$$\widehat{f} := \operatorname*{argmin}_{f\in\mathcal{F}_{\widehat{k}}} \widehat{L}_n(f). \tag{4.4.7b}$$

With this procedure, we have the following theorem.

Theorem 4.4.12. *Let $\widehat{f}$ be chosen according to the procedure (4.4.7a)–(4.4.7b). Then with probability at least $1-\delta$, we have*

$$L(\widehat{f}) \le \inf_{k\in\mathbb{N}} \inf_{f\in\mathcal{F}_k} \left\{L(f) + 2C_{n,k}(\delta)\right\}.$$

Proof First, we have by the assumed guarantee on $C_{n,k}(\delta)$ that

$$\begin{aligned}
&\mathbb{P}\left(\exists\, k \in \mathbb{N} \text{ and } f \in \mathcal{F}_k \text{ such that } \sup_{f\in\mathcal{F}_k} \left|\widehat{L}_n(f) - L(f)\right| \ge C_{n,k}(\delta)\right) \\
&\quad \le \sum_{k=1}^\infty \mathbb{P}\left(\exists\, f \in \mathcal{F}_k \text{ such that } \sup_{f\in\mathcal{F}_k} \left|\widehat{L}_n(f) - L(f)\right| \ge C_{n,k}(\delta)\right) \le \sum_{k=1}^\infty \delta \cdot 2^{-k} = \delta.
\end{aligned}$$

On the event that $\sup_{f\in\mathcal{F}_k} |\widehat{L}_n(f) - L(f)| < C_{n,k}(\delta)$ for all k, which occurs with probability at least $1-\delta$, we have

$$L(\widehat{f}) \le \widehat{L}_n(f) + C_{n,\widehat{k}}(\delta) = \inf_{k\in\mathbb{N}} \inf_{f\in\mathcal{F}_k} \left\{\widehat{L}_n(f) + C_{n,k}(\delta)\right\} \le \inf_{k\in\mathbb{N}} \inf_{f\in\mathcal{F}_k} \left\{L(f) + 2C_{n,k}(\delta)\right\}$$

by our choice of $\widehat{f}$. This is the desired result. □

We conclude with a final example, using our earlier floating point bound from Example 4.4.5, coupled with Corollary 4.4.6 and Theorem 4.4.12.

Example 4.4.13 (Structural risk minimization with floating point classifiers)**:** Consider again our floating point example, and let the function class $\mathcal{F}_k$ consist of functions defined by at most k double-precision floating point values, so that $\log|\mathcal{F}_k| \le 45d$. Then by taking

$$C_{n,k}(\delta) = \sqrt{\frac{\log\frac{1}{\delta} + 65k\log 2}{2n}}$$

we have that $|\widehat{L}_n(f) - L(f)| \le C_{n,k}(\delta)$ simultaneously for all $f \in \mathcal{F}_k$ and all $\mathcal{F}_k$, with probability at least $1-\delta$. Then the empirical risk minimization procedure (4.4.7) guarantees that

$$L(\widehat{f}) \le \inf_{k\in\mathbb{N}} \left\{ \inf_{f\in\mathcal{F}_k} L(f) + \sqrt{\frac{2\log\frac{1}{\delta} + 91k}{n}} \right\}.$$

Roughly, we trade between small risk $L(f)$—as the risk $\inf_{f\in\mathcal{F}_k} L(f)$ must be decreasing in k—and the estimation error penalty, which scales as $\sqrt{(k + \log\frac{1}{\delta})/n}$. ◇

4.5 Technical proofs

4.5.1 Proof of Theorem 4.1.11

(1) implies (2) Let $K_1 = 1$. Using the change of variables identity that for a nonnegative random variable Z and any $k \ge 1$ we have $\mathbb{E}[Z^k] = k\int_0^\infty t^{k-1}\mathbb{P}(Z \ge t)dt$, we find

$$\mathbb{E}[|X|^k] = k\int_0^\infty t^{k-1}\mathbb{P}(|X| \ge t)dt \le 2k\int_0^\infty t^{k-1}\exp\left(-\frac{t^2}{\sigma^2}\right)dt = k\sigma^k\int_0^\infty u^{k/2-1}e^{-u}du,$$

where for the last inequality we made the substitution $u = t^2/\sigma^2$. Noting that this final integral is $\Gamma(k/2)$, we have $\mathbb{E}[|X|^k] \le k\sigma^k\Gamma(k/2)$. Because $\Gamma(s) \le s^s$ for $s \ge 1$, we obtain

$$\mathbb{E}[|X|^k]^{1/k} \le k^{1/k}\sigma\sqrt{k/2} \le e^{1/e}\sigma\sqrt{k}.$$

Thus (2) holds with $K_2 = e^{1/e}$.

(2) implies (3) Let $\sigma = \|X\|_{\psi_2} = \sup_{k\ge1} k^{-\frac{1}{2}}\mathbb{E}[|X|^k]^{1/k}$, so that $K_2 = 1$ and $\mathbb{E}[|X|^k] \le k^{\frac{k}{2}}\sigma$ for all k. For $K_3 \in \mathbb{R}_+$, we thus have

$$\mathbb{E}[\exp(X^2/(K_3\sigma^2))] = \sum_{k=0}^\infty \frac{\mathbb{E}[X^{2k}]}{k!K_3^{2k}\sigma^{2k}} \le \sum_{k=0}^\infty \frac{\sigma^{2k}(2k)^k}{k!K_3^{2k}\sigma^{2k}} \overset{(i)}{\le} \sum_{k=0}^\infty \left(\frac{2e}{K_3^2}\right)^k$$

where inequality (i) follows because $k! \ge (k/e)^k$, or $1/k! \le (e/k)^k$. Noting that $\sum_{k=0}^\infty \alpha^k = \frac{1}{1-\alpha}$, we obtain (3) by taking $K_3 = e\sqrt{2/(e-1)} \approx 2.933$.

(3) implies (4) Let us take $K_3 = 1$. We claim that (4) holds with $K_4 = \frac{3}{4}$. We prove this result for both small and large λ. First, note the (highly non-standard, but true!) inequality that $e^x \le x + e^{\frac{9x^2}{16}}$ for all x. Then we have

$$\mathbb{E}[\exp(\lambda X)] \le \underbrace{\mathbb{E}[\lambda X]}_{=0} + \mathbb{E}\left[\exp\left(\frac{9\lambda^2 X^2}{16}\right)\right]$$

Now note that for $|\lambda| \le \frac{4}{3\sigma}$, we have $9\lambda^2\sigma^2/16 \le 1$, and so by Jensen's inequality,

$$\mathbb{E}\left[\exp\left(\frac{9\lambda^2 X^2}{16}\right)\right] = \mathbb{E}\left[\exp(X^2/\sigma^2)^{\frac{9\lambda^2\sigma^2}{16}}\right] \le e^{\frac{9\lambda^2\sigma^2}{16}}.$$

For large λ, we use the simpler Fenchel-Young inequality, that is, that $\lambda x \le \frac{\lambda^2}{2c} + \frac{cx^2}{2}$, valid for all $c \ge 0$. Then we have for any $0 \le c \le 2$ that

$$\mathbb{E}[\exp(\lambda X)] \le e^{\frac{\lambda^2\sigma^2}{2c}} \mathbb{E}\left[\exp\left(\frac{cX^2}{2\sigma^2}\right)\right] \le e^{\frac{\lambda^2\sigma^2}{2c}} e^{\frac{c}{2}},$$

where the final inequality follows from Jensen's inequality. If $|\lambda| \ge \frac{4}{3\sigma}$, then $\frac{1}{2} \le \frac{9}{32}\lambda^2\sigma^2$, and we have

$$\mathbb{E}[\exp(\lambda X)] \le \inf_{c\in[0,2]} e^{[\frac{1}{2c} + \frac{9c}{32}]\lambda^2\sigma^2} = \exp\left(\frac{3\lambda^2\sigma^2}{4}\right).$$

(4) implies (1) This is the content of Proposition 4.1.8, with $K_4 = \frac{1}{2}$ and $K_1 = 2$.

4.5.2 Proof of Theorem 4.1.15

(1) implies (2) As in the proof of Theorem 4.1.11, we use that for a nonnegative random variable Z we have $\mathbb{E}[Z^k] = k\int_0^\infty t^{k-1}\mathbb{P}(Z \ge t)dt$. Let $K_1 = 1$. Then

$$\mathbb{E}[|X|^k] = k\int_0^\infty t^{k-1}\mathbb{P}(|X| \ge t)dt \le 2k\int_0^\infty t^{k-1}\exp(-t/\sigma)dt = 2k\sigma^k\int_0^\infty u^{k-1}\exp(-u)du,$$

where we used the substitution $u = t/\sigma$. Thus we have $\mathbb{E}[|X|^k] \le 2\Gamma(k+1)\sigma^k$, and using $\Gamma(k+1) \le k^k$ yields $\mathbb{E}[|X|^k]^{1/k} \le 2^{1/k}k\sigma$, so that (2) holds with $K_2 \le 2$.

(2) implies (3) Let $K_2 = 1$, and note that

$$\mathbb{E}[\exp(X/(K_3\sigma))] = \sum_{k=0}^\infty \frac{\mathbb{E}[X^k]}{K_3^k\sigma^k k!} \le \sum_{k=0}^\infty \frac{k^k}{k!}\cdot\frac{1}{K_3^k} \overset{(i)}{\le} \sum_{k=0}^\infty \left(\frac{e}{K_3}\right)^k,$$

where inequality (i) used that $k! \ge (k/e)^k$. Taking $K_3 = e^2/(e-1) < 5$ gives the result.

(3) implies (1) If $\mathbb{E}[\exp(X/\sigma)] \le e$, then for $t \ge 0$

$$\mathbb{P}(X \ge t) \le \mathbb{E}[\exp(X/\sigma)]e^{-t/\sigma} \le e^{1-t/\sigma}.$$

With the same result for the negative tail, we have

$$\mathbb{P}(|X| \ge t) \le 2e^{1-t/\sigma} \wedge 1 \le 2e^{-\frac{2t}{5\sigma}},$$

so that (1) holds with $K_1 = \frac{5}{2}$.

(2) if and only if (4) Thus, we see that up to constant numerical factors, the definition $\|X\|_{\psi_1} = \sup_{k \geq 1} k^{-1} \mathbb{E}[|X|^k]^{1/k}$ has the equivalent statements

$$\mathbb{P}(|X| \geq t) \leq 2 \exp(-t/(K_1 \|X\|_{\psi_1})) \quad \text{and} \quad \mathbb{E}[\exp(X/(K_3 \|X\|_{\psi_1}))] \leq e.$$

Now, let us assume that (2) holds with $K_2 = 1$, so that $\sigma = \|X\|_{\psi_1}$ and that $\mathbb{E}[X] = 0$. Then we have $\mathbb{E}[X^k] \leq k^k \|X\|_{\psi_1}^k$, and

$$\mathbb{E}[\exp(\lambda X)] = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}[X^k]}{k!} \leq 1 + \sum_{k=2}^{\infty} \lambda^k \|X\|_{\psi_1}^k \cdot \frac{k^k}{k!} \leq 1 + \sum_{k=2}^{\infty} \lambda^k \|X\|_{\psi_1}^k e^k,$$

the final inequality following because $k! \geq (k/e)^k$. Now, if $|\lambda| \leq \frac{1}{2e\|X\|_{\psi_1}}$, then we have

$$\mathbb{E}[\exp(\lambda X)] \leq 1 + \lambda^2 e^2 \|X\|_{\psi_1} \sum_{k=0}^{\infty} (\lambda \|X\|_{\psi_1} e)^k \leq 1 + 2e^2 \|X\|_{\psi_1}^2 \lambda^2,$$

as the final sum is at most $\sum_{k=0}^{\infty} 2^{-k} = 2$. Using $1 + x \leq e^x$ gives that (2) implies (4). For the opposite direction, we may simply use that if (4) holds with $K_4 = 1$ and $K_4' = 1$, then $\mathbb{E}[\exp(X/\sigma)] \leq \exp(1)$, so that (3) holds.

4.5.3 Proof of Theorem 4.3.6

JCD Comment: I would like to write this. For now, check out Ledoux and Talagrand [129, Theorem 4.12] or Koltchinskii [122, Theorem 2.2].

4.6 Bibliography

A few references on concentration, random matrices, and entropies include Vershynin's extraordinarily readable lecture notes [170], upon which our proof of Theorem 4.1.11 is based, the comprehensive book of Boucheron, Lugosi, and Massart [34], and the more advanced material in Buldygin and Kozachenko [41]. Many of our arguments are based off of those of Vershynin and Boucheron et al. Kolmogorov and Tikhomirov [121] introduced metric entropy.

4.7 Exercises

Exercise 4.1 (Concentration of bounded random variables)**:** Let X be a random variable taking values in $[a, b]$, where $-\infty < a \leq b < \infty$. In this question, we show *Hoeffding's Lemma*, that is, that X is sub-Gaussian: for all $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right).$$

(a) Show that $\mathrm{Var}(X) \leq (\frac{b-a}{2})^2 = \frac{(b-a)^2}{4}$ for any random variable X taking values in $[a, b]$.

(b) Let

$$\varphi(\lambda) = \log \mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))].$$

Assuming that $\mathbb{E}[X] = 0$ (convince yourself that this is no loss of generality) show that

$$\varphi(0) = 0, \quad \varphi'(0) = 0, \quad \varphi''(t) = \frac{\mathbb{E}[X^2 e^{tX}]}{\mathbb{E}[e^{tX}]} - \frac{\mathbb{E}[Xe^{tX}]^2}{\mathbb{E}[e^{tX}]^2}.$$

(You may assume that derivatives and expectations commute, which they do in this case.)

(c) Construct a random variable Y_t, defined for $t \in \mathbb{R}$, such that $Y_t \in [a, b]$ and

$$\mathrm{Var}(Y_t) = \varphi''(t).$$

(You may assume X has a density for simplicity.)

(d) Using the result of part (c), show that $\varphi(\lambda) \le \frac{\lambda^2(b-a)^2}{8}$ for all $\lambda \in \mathbb{R}$.

Exercise 4.2: In this question, we show how to use Bernstein-type (sub-exponential) inequalities to give sharp convergence guarantees. Recall (Example 4.1.14, Corollary 4.1.18, and inequality (4.1.6)) that if X_i are independent bounded random variables with $|X_i - \mathbb{E}[X]| \le b$ for all i and $\mathrm{Var}(X_i) \le \sigma^2$, then

$$\max\left\{\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \ge \mathbb{E}[X] + t\right), \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \le \mathbb{E}[X] - t\right)\right\} \le \exp\left(-\frac{1}{2}\min\left\{\frac{5}{6}\frac{nt^2}{\sigma^2}, \frac{nt}{2b}\right\}\right).$$

We consider minimization of loss functions ℓ over finite function classes $\mathcal{F}$ with $\ell \in [0, 1]$, so that if $L(f) = \mathbb{E}[\ell(f, Z)]$ then $|\ell(f, Z) - L(f)| \le 1$. Throughout this question, we let

$$L^\star = \min_{f\in\mathcal{F}} L(f) \quad \text{and} \quad f^\star \in \operatorname*{argmin}_{f\in\mathcal{F}} L(f).$$

We will show that, roughly, a procedure based on picking an empirical risk minimizer is unlikely to choose a function $f \in \mathcal{F}$ with bad performance, so that we obtain faster concentration guarantees.

(a) Argue that for any $f \in \mathcal{F}$

$$\mathbb{P}\left(\widehat{L}(f) \ge L(f) + t\right) \vee \mathbb{P}\left(\widehat{L}(f) \le L(f) - t\right) \le \exp\left(-\frac{1}{2}\min\left\{\frac{5}{6}\frac{nt^2}{L(f)(1-L(f))}, \frac{nt}{2}\right\}\right).$$

(b) Define the set of "bad" prediction functions $\mathcal{F}_{\epsilon\,\mathrm{bad}} := \{f \in \mathcal{F} : L(f) \ge L^\star + \epsilon\}$. Show that for any fixed $\epsilon \ge 0$ and any $f \in \mathcal{F}_{2\epsilon\,\mathrm{bad}}$, we have

$$\mathbb{P}\left(\widehat{L}(f) \le L^\star + \epsilon\right) \le \exp\left(-\frac{1}{2}\min\left\{\frac{5}{6}\frac{n\epsilon^2}{L^\star(1-L^\star) + \epsilon(1-\epsilon)}, \frac{n\epsilon}{2}\right\}\right).$$

(c) Let $\widehat{f}_n \in \operatorname{argmin}_{f\in\mathcal{F}} \widehat{L}(f)$ denote the empirical minimizer over the class $\mathcal{F}$. Argue that it is likely to have good performance, that is, for all $\epsilon \ge 0$ we have

$$\mathbb{P}\left(L(\widehat{f}_n) \ge L(f^\star) + 2\epsilon\right) \le \mathrm{card}(\mathcal{F}) \cdot \exp\left(-\frac{1}{2}\min\left\{\frac{5}{6}\frac{n\epsilon^2}{L^\star(1-L^\star) + \epsilon(1-\epsilon)}, \frac{n\epsilon}{2}\right\}\right).$$

(d) Using the result of part (c), argue that with probability at least $1 - \delta$,

$$L(\widehat{f}_n) \le L(f^\star) + \frac{4 \log \frac{|\mathcal{F}|}{\delta}}{n} + \sqrt{\frac{12}{5}} \cdot \frac{\sqrt{L^\star(1 - L^\star) \cdot \log \frac{|\mathcal{F}|}{\delta}}}{\sqrt{n}}.$$

Why is this better than an inequality based purely on the boundedness of the loss ℓ, such as Theorem 4.4.4 or Corollary 4.4.6? What happens when there is a perfect risk minimizer $f^\star$?

Exercise 4.3 (Likelihood ratio bounds and concentration)**:** Consider a data release problem, where given a sample x, we release a sequence of data $Z_1, Z_2, \ldots, Z_n$ belonging to a discrete set $\mathcal{Z}$, where Z_i may depend on Z_1^{i-1} and x. We assume that the data has limited information about x in the sense that for any two samples x, x', we have the likelihood ratio bound

$$\frac{p(z_i \mid x, z_1^{i-1})}{p(z_i \mid x', z_1^{i-1})} \le e^\varepsilon.$$

Let us control the amount of "information" (in the form of an updated log-likelihood ratio) released by this sequential mechanism. Fix x, x', and define

$$L(z_1, \ldots, z_n) := \log \frac{p(z_1, \ldots, z_n \mid x)}{p(z_1, \ldots, z_n \mid x')}.$$

(a) Show that, assuming the data Z_i are drawn conditional on x,

$$\mathbb{P}\left(L(Z_1, \ldots, Z_n) \ge n\varepsilon(e^\varepsilon - 1) + t\right) \le \exp\left(-\frac{t^2}{2n\varepsilon^2}\right).$$

Equivalently, show that

$$\mathbb{P}\left(L(Z_1, \ldots, Z_n) \ge n\varepsilon(e^\varepsilon - 1) + \varepsilon\sqrt{2n \log(1/\delta)}\right) \le \delta.$$

(b) Let $\gamma \in (0, 1)$. Give the largest value of ε you can that is sufficient to guarantee that for any test $\Psi : \mathcal{Z}^n \to \{x, x'\}$, we have

$$P_x(\Psi(Z_1^n) \ne x) + P_{x'}(\Psi(Z_1^n) \ne x') \ge 1 - \gamma,$$

where P_x and $P_{x'}$ denote the sampling distribution of Z_1^n under x and x', respectively?

Exercise 4.4 (Marcinkiewicz-Zygmund inequality)**:** Let X_i be independent random variables with $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[|X_i|^p] < \infty$, where $1 \le p < \infty$. Prove that

$$\mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^p\right] \le C_p \mathbb{E}\left[\left(\sum_{i=1}^n |X_i|^2\right)^{p/2}\right]$$

where C_p is a constant (that depends on p). As a corollary, derive that if $\mathbb{E}[|X_i|^p] \le \sigma^p$ and $p \ge 2$, then

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i\right|^p\right] \le C_p \frac{\sigma^p}{n^{p/2}}.$$

That is, sample means converge quickly to zero in higher moments. *Hint:* For any fixed $x \in \mathbb{R}^n$, if ε_i are i.i.d. uniform signs $\varepsilon_i \in \{\pm 1\}$, then $\varepsilon^T x$ is sub-Gaussian.

Exercise 4.5 (Small balls and anti-concentration)**:** Let X be a nonnegative random variable satisfying $\mathbb{P}(X \le \epsilon) \le c\epsilon$ for some $c < \infty$ and all $\epsilon > 0$. Argue that if X_i are i.i.d. copies of X, then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \ge t\right) \ge 1 - \exp(-2n\left[1/2 - 2ct\right]_+^2)$$

for all t.

Exercise 4.6 (Lipschitz functions remain sub-Gaussian)**:** Let X be σ^2-sub-Gaussian and $f : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz, meaning that $|f(x) - f(y)| \le L|x - y|$ for all x, y. Prove that there exists a numerical constant $C < \infty$ such that $f(X)$ is $CL^2\sigma^2$-sub-Gaussian.

Exercise 4.7 (Sub-gaussian maxima)**:** Let $X_1, \ldots, X_n$ be σ^2-sub-gaussian (not necessarily independent) random variables. Show that

(a) $\mathbb{E}[\max_i X_i] \le \sqrt{2\sigma^2 \log n}$.

(b) There exists a numerical constant $C < \infty$ such that $\mathbb{E}[\max_i |X_i|^p] \le (Cp\sigma^2 \log k)^{p/2}$.

Exercise 4.8: Consider a binary classification problem with logistic loss $\ell(\theta; (x, y)) = \log(1 + \exp(-y\theta^T x))$, where $\theta \in \Theta := \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \le r\}$ and $y \in \{\pm 1\}$. Assume additionally that the space $\mathcal{X} \subset \{x \in \mathbb{R}^d \mid \|x\|_\infty \le b\}$. Define the empirical and population risks $\widehat{L}_n(\theta) := P_n \ell(\theta; (X, Y))$ and $L(\theta) := P\ell(\theta; (X, Y))$, and let $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \widehat{L}(\theta)$. Show that with probability at least $1 - \delta$ over $(X_i, Y_i) \stackrel{\text{iid}}{\sim} P$,

$$L(\widehat{\theta}_n) \le \inf_{\theta \in \Theta} L(\theta) + C\frac{rb\sqrt{\log \frac{d}{\delta}}}{\sqrt{n}}$$

where $C < \infty$ is a numerical constant (you need not specify this).

Exercise 4.9 (Sub-Gaussian constants of Bernoulli random variables)**:** In this exercise, we will derive sharp sub-Gaussian constants for Bernoulli random variables (cf. [106, Thm. 1] or [118, 24]), showing

$$\log \mathbb{E}[e^{t(X-p)}] \le \frac{1 - 2p}{4 \log \frac{1-p}{p}} t^2 \quad \text{for all } t \ge 0. \tag{4.7.1}$$

(a) Define $\varphi(t) = \log(\mathbb{E}[e^{t(X-p)}]) = \log((1-p)e^{-tp} + pe^{t(1-p)})$. Show that

$$\varphi'(t) = \mathbb{E}[Y_t] \quad \text{and} \quad \varphi''(t) = \operatorname{Var}(Y_t)$$

where $Y_t = (1 - p)$ with probability $q(t) := \frac{pe^{t(1-p)}}{pe^{t(1-p)} + (1-p)e^{-tp}}$ and $Y_t = -p$ otherwise.

(b) Show that $\varphi'(0) = 0$ and that if $p > \frac{1}{2}$, then $\operatorname{Var}(Y_t) \le \operatorname{Var}(Y_0) = p(1-p)$. Conclude that $\varphi(t) \le \frac{p(1-p)}{2} t^2$ for all $t \ge 0$.

(c) Argue that $p(1-p) \le \frac{1-2p}{2 \log \frac{1-p}{p}}$ for $p \in [0, 1]$. *Hint:* Let $p = \frac{1+\delta}{2}$ for $\delta \in [0, 1]$, so that the inequality is equivalent to $\log \frac{1+\delta}{1-\delta} \le \frac{2\delta}{1-\delta^2}$. Then use that $\log(1 + \delta) = \int_0^\delta \frac{1}{1+u} du$.

(d) Let $C = 2 \log \frac{1-p}{p}$ and define $s = Ct = 2 \log \frac{1-p}{p} s$, and let

$$f(s) = \frac{1-2p}{2} Cs^2 + Cps - \log(1 - p + pe^{Cs}),$$

so that inequality (4.7.1) holds if and only if $f(s) \geq 0$ for all $s \geq 0$. Give $f'(s)$ and $f''(s)$.

(e) Show that $f(0) = f(1) = f'(0) = f'(1) = 0$, and argue that $f''(s)$ changes signs at most twice and that $f''(0) = f''(1) > 0$. Use this to show that $f(s) \geq 0$ for all $s \geq 0$.

JCD Comment: Perhaps use transportation inequalities to prove this bound, and also maybe give Ordentlich and Weinberger's "A Distribution Dependent Refinement of Pinsker's Inequality" as an exercise.

Exercise 4.10: Let $s(p) = \frac{1-2p}{\log \frac{1-p}{p}}$. Show that s is concave on $[0, 1]$.

Exercise 4.11: Prove Lemma 4.3.8.

JCD Comment: Add in some connections to the exponential family material. Some ideas:

1. A hypothesis test likelihood ratio for them (see page 40 of handwritten notes)
2. A full learning guarantee with convergence of Hessian and everything, e.g., for logistic regression?
3. In the Ledoux-Talagrand stuff, maybe worth going through example of logistic regression. Also, having working logistic example throughout? Helps clear up the structure and connect with exponential families.
4. Maybe an exercise for Lipschitz functions with random Lipschitz constants?

Chapter 5

Generalization and stability

Concentration inequalities provide powerful techniques for demonstrating when random objects that are functions of collections of independent random variables—whether sample means, functions with bounded variation, or collections of random vectors—behave similarly to their expectations. This chapter continues exploration of these ideas by incorporating the central thesis of this book: that information theory's connections to statistics center around measuring when (and how) two probability distributions get close to one another. On its face, we remain focused on the main objects of the preceding chapter, where we have a population probability distribution P on a space $\mathcal{X}$ and some collection of functions $f : \mathcal{X} \to \mathbb{R}$. We then wish to understand when we expect the empirical distribution

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}X_i,$$

defined by teh sample $X_i \stackrel{\text{iid}}{\sim} P$, to be close to the population P as measured by f. Following the notation we introduce in Section 4.3, for $Pf := \mathbb{E}_P[f(X)]$, we again ask to have

$$P_n f - Pf = \frac{1}{n} \sum_{i=1}^{n} \left(f(X_i) - \mathbb{E}_P[f(X)] \right)$$

to be small simultaneously for all f.

In this chapter, however, we develop a family of tools based around *PAC (probably approximately correct) Bayesian* bounds, where we slightly perturb the functions f of interest to average them in some way; when these perturbations keep $P_n f$ stable, we expect that $P_n f \approx Pf$, that is, the sample generalizes to the population. These perturbations allow us to bring the tools of the divergence measures we have developed to bear on the problems of convergence and generalization. Even more, they allow us to go beyond the "basic" concentration inequalities to situations with interaction, where a data analyst may evaluate some functions of P_n, then adaptively choose additional queries or analyses to do on the sample sample X_1^n. This breaks standard statistical analyses—which assume an *a priori* specified set of hypotheses or questions to be answered—but is possible to address once we can limit the information the analyses release in precise ways that information-theoretic tools allow. Modern work has also shown how to leverage these techniques, coupled with computation, to provide non-vacuous bounds on learning for complicated scenarios and models to which all classical bounds fail to apply, such as deep learning.

5.1 The variational representation of Kullback-Leibler divergence

The starting point of all of our generalization bounds is a surprisingly simply variational result, which relates expectations, moment generating functions, and the KL-divergence in one single equality. It turns out that this inequality, by relating means with moment generating functions and divergences, allows us to prove generalization bounds based on information-theoretic tools and stability.

Theorem 5.1.1 (Donsker-Varadhan variational representation)**.** *Let P and Q be distributions on a common space $\mathcal{X}$. Then*

$$D_{\rm kl}\left(P\|Q\right) = \sup_g \left\{\mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[e^{g(X)}]\right\},$$

where the supremum is taken over measurable functions $g : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_Q[e^{g(X)}] < \infty$.

We give one proof of this result and one sketch of a proof, which holds when the underlying space is discrete, that may be more intuitive: the first constructs a particular "tilting" of Q via the function e^g, and verifies the equality. The second relies on the discretization of the KL-divergence and may be more intuitive to readers familiar with convex optimization: essentially, we expect this result because the function $\log(\sum_{j=1}^k e^{x_j})$ is the convex conjugate of the negative entropy. (See also Exercise 5.1.)

Proof We may assume that P is absolutely continuous with respect to Q, meaning that $Q(A) = 0$ implies that $P(A) = 0$, as otherwise both sides are infinite by inspection. Thus, it is no loss of generality to let P and Q have densities p and q.

Attainment in the equality is easy: we simply take $g(x) = \log \frac{p(x)}{q(x)}$, so that $\mathbb{E}_Q[e^{g(X)}] = 1$. To show that the right hand side is never larger than $D_{\rm kl}\left(P\|Q\right)$ requires a bit more work. To that end, let g be any function such that $\mathbb{E}_Q[e^{g(X)}] < \infty$, and define the random variable $Z_g(x) = e^{g(x)}/\mathbb{E}_Q[e^{g(X)}]$, so that $\mathbb{E}_Q[Z] = 1$. Then using the absolute continuity of P w.r.t. Q, we have

$$\begin{aligned}
\mathbb{E}_P[\log Z_g] = \mathbb{E}_P\left[\log \frac{p(X)}{q(X)} + \log\left(Z_g(X)\frac{q(X)}{p(X)}\right)\right] &= D_{\rm kl}\left(P\|Q\right) + \mathbb{E}_P\left[\log\left(Z_g \frac{dQ}{dP}\right)\right] \\
&\le D_{\rm kl}\left(P\|Q\right) + \log \mathbb{E}_P\left[\frac{dQ}{dP} Z_g\right] \\
&= D_{\rm kl}\left(P\|Q\right) + \log \mathbb{E}_Q[Z_g].
\end{aligned}$$

As $\mathbb{E}_Q[Z_g] = 1$, using that $\mathbb{E}_P[\log Z_g] = \mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[e^{g(X)}]$ gives the result. □

Here is the second proof of Theorem 5.1.1, which applies when $\mathcal{X}$ is discrete and finite. That we can approximate KL-divergence by suprema over finite partitions (as in definition (2.2.1)) suggests that this approach works in general—which it can—but this requires some not completely trivial approximations of $\mathbb{E}_P[g]$ and $\mathbb{E}_Q[e^g]$ by discretized versions of their expectations, which makes things rather tedious.

Proof of Theorem 5.1.1, the finite case As we have assumed that P and Q have finite supports, which we identify with $\{1, \ldots, k\}$ and p.m.f.s $p, q \in \Delta_k = \{p \in \mathbb{R}_+^k \mid \langle \mathbf{1}, p\rangle = 1\}$. Define $f_q(v) = \log(\sum_{j=1}^k q_j e^{v_j})$, which is convex in v (recall Proposition 3.2.1). Then the supremum in the variational representation takes the form

$$h(p) := \sup_{v \in \mathbb{R}^k} \left\{\langle p, v\rangle - f_q(v)\right\}.$$

If we can take derivatives and solve for zero, we are guaranteed to achieve the supremum. To that end, note that

$$\nabla_v \{\langle p, v\rangle - f_q(v)\} = p - \left[\frac{q_i e^{v_i}}{\sum_{j=1}^k q_j e^{v_j}}\right]_{i=1}^k,$$

so that setting $v_j = \log \frac{p_j}{q_j}$ achieves $p - \nabla_v f_q(v) = p - p = 0$ and hence the supremum. Noting that $\log(\sum_{j=1}^k q_j \exp(\log \frac{p_j}{q_j})) = \log(\sum_{j=1}^k p_j) = 0$ gives $h(p) = D_{\rm kl}\left(p\|q\right)$. □

The Donsker-Varadhan variational representation already gives a hint that we can use some information-theoretic techniques to control the difference between an empirical sample and its expectation, at least in an average sense. In particular, we see that for any function g, we have

$$\mathbb{E}_P[g(X)] \le D_{\rm kl}\left(P\|Q\right) + \log \mathbb{E}_Q[e^{g(X)}]$$

for any random variable X. Now, changing this on its head a bit, suppose that we consider a collection of functions $\mathcal{F}$ and put two probability measures π and π_0 on $\mathcal{F}$, and consider $P_n f - Pf$, where we consider f a random variable $f \sim \pi$ or $f \sim \pi_0$. Then a consequence of the Donsker-Varadhan theorem is that

$$\int (P_n f - Pf) d\pi(f) \le D_{\rm kl}\left(\pi\|\pi_0\right) + \log \int \exp(P_n f - Pf) d\pi_0(f)$$

for any π, π_0. While this inequality is a bit naive—bounding a difference by an exponent seems wasteful—as we shall see, it has substantial applications when we can upper bound the KL-divergence $D_{\rm kl}\left(\pi\|\pi_0\right)$.

5.2 PAC-Bayes bounds

Probably-approximately-correct (PAC) Bayesian bounds proceed from a perspective similar to that of the covering numbers and covering entropies we develop in Section 4.3, where if for a collection of functions $\mathcal{F}$ there is a finite subset (a cover) $\{f_v\}$ such that each $f \in \mathcal{F}$ is "near" one of the f_v, then we need only control deviations of $P_n f$ from Pf for the elements of $\{f_v\}$. In PAC-Bayes bounds, we instead average functions f with other functions, and this averaging allows a similar family of guarantees and applications.

Let us proceed with the main results. Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$, and assume that each function f is σ^2-sub-Gaussian, which we recall (Definition 4.1) means that $\mathbb{E}[e^{\lambda(f(X) - Pf)}] \le \exp(\lambda^2\sigma^2/2)$ for all $\lambda \in \mathbb{R}$, where $Pf = \mathbb{E}_P[f(X)] = \int f(x) dP(x)$ denotes the expectation of f under P. The main theorem of this section shows that averages of the squared error $(P_n f - Pf)^2$ of the empirical distribution P_n to P converge quickly to zero for *all* averaging distributions π on functions $f \in \mathcal{F}$ so long as each f is σ^2-sub-Gaussian, with the caveat that we pay a cost for different choices of π. The key is that we choose some prior distribution π_0 on $\mathcal{F}$ first.

Theorem 5.2.1. *Let Π be the collection of all probability distributions on the set $\mathcal{F}$ and let π_0 be a fixed prior probability distribution on $f \in \mathcal{F}$. With probability at least $1 - \delta$,*

$$\int (P_n f - Pf)^2 d\pi(f) \le \frac{8\sigma^2}{3} \frac{D_{\rm kl}\left(\pi\|\pi_0\right) + \log \frac{2}{\delta}}{n} \quad \textit{simultaneously for all } \pi \in \Pi.$$

Proof The key is to combine Example 4.1.12 with the variational representation that Theorem 5.1.1 provides for KL-divergences. We state Example 4.1.12 as a lemma here.

Lemma 5.2.2. *Let Z be a σ^2-sub-Gaussian random variable. Then for $\lambda \geq 0$,*

$$\mathbb{E}[e^{\lambda Z^2}] \leq \frac{1}{\sqrt{[1 - 2\sigma^2\lambda]_+}}.$$

Without loss of generality, we assume that $Pf = 0$ for all $f \in \mathcal{F}$, and recall that $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$ is the empirical mean of f. Then we know that $P_n f$ is σ^2/n-sub-Gaussian, and Lemma 5.2.2 implies that $\mathbb{E}[\exp(\lambda(P_n f)^2)] \leq \left[1 - 2\lambda\sigma^2/n\right]_+^{-1/2}$ for any f, and thus for any prior π_0 on f we have

$$\mathbb{E}\left[\int \exp(\lambda(P_n f)^2) d\pi_0(f)\right] \leq \left[1 - 2\lambda\sigma^2/n\right]_+^{-1/2}.$$

Consequently, taking $\lambda = \lambda_n := \frac{3n}{8\sigma^2}$, we obtain

$$\mathbb{E}\left[\int \exp(\lambda_n(P_n f)^2) d\pi_0(f)\right] = \mathbb{E}\left[\int \exp\left(\frac{3n}{8\sigma^2}(P_n f)^2\right) d\pi_0(f)\right] \leq 2.$$

Markov's inequality thus implies that

$$\mathbb{P}\left(\int \exp\left(\lambda_n(P_n f)^2\right) d\pi_0(f) \geq \frac{2}{\delta}\right) \leq \delta, \tag{5.2.1}$$

where the probability is over $X_i \stackrel{\text{iid}}{\sim} P$.

Now, we use the Donsker-Varadhan equality (Theorem 5.1.1). Letting $\lambda > 0$, we define the function $g(f) = \lambda(P_n f)^2$, so that for any two distributions π and π_0 on $\mathcal{F}$, we have

$$\frac{1}{\lambda}\int g(f) d\pi(f) = \int (P_n f)^2 d\pi(f) \leq \frac{D_{\rm kl}\left(\pi \| \pi_0\right) + \log \int \exp(\lambda(P_n f)^2) d\pi_0(f)}{\lambda}.$$

This holds without any probabilistic qualifications, so using the application (5.2.1) of Markov's inequality with $\lambda = \lambda_n$, we thus see that with probability at least $1 - \delta$ over $X_1, \ldots, X_n$, simultaneously for all distributions π,

$$\int (P_n f)^2 d\pi(f) \leq \frac{8\sigma^2}{3} \frac{D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{2}{\delta}}{n}.$$

This is the desired result (as we have assumed that $Pf = 0$ w.l.o.g.). □

By Jensen's inequality (or Cauchy-Schwarz), it is immediate from Theorem 5.2.1 that we also have

$$\int |P_n f - Pf| d\pi(f) \leq \sqrt{\frac{8\sigma^2}{3} \frac{D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{2}{\delta}}{n}} \quad \text{simultaneously for all } \pi \in \Pi \tag{5.2.2}$$

with probability at least $1 - \delta$, so that $\mathbb{E}_\pi[|P_n f - Pf|]$ is with high probability of order $1/\sqrt{n}$. The inequality (5.2.2) is the original form of the PAC-Bayes bound due to McAllester, with slightly

sharper constants and improved logarithmic dependence. The key is that *stability*, in the form of a prior π_0 and posterior π closeness, allow us to achieve reasonably tight control over the deviations of random variables and functions with high probability.

Let us give an example, which is similar to many of our approaches in Section 4.4, to illustrate some of the approaches this allows. The basic idea is that by appropriate choice of prior π_0 and "posterior" π, whenever we have appropriately smooth classes of functions we achieve certain generalization guarantees.

Example 5.2.3 (A uniform law for Lipschitz functions)**:** Consider a case as in Section 4.4, where we let $L(\theta) = P\ell(\theta, Z)$ for some function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$. Let $\mathbb{B}_2^d = \{v \in \mathbb{R}^d \mid \|v\|_2 \le 1\}$ be the ℓ_2-ball in $\mathbb{R}^d$, and let us assume that $\Theta \subset r\mathbb{B}_2^d$ and additionally that $\theta \mapsto \ell(\theta, z)$ is M-Lipschitz for all $z \in \mathcal{Z}$. For simplicity, we assume that $\ell(\theta, z) \in [0, 2Mr]$ for all $\theta \in \Theta$ (we may simply relativize our bounds by replacing ℓ by $\ell(\cdot, z) - \inf_{\theta\in\Theta} \ell(\theta, z) \in [0, 2Mr]$).
If $\widehat{L}_n(\theta) = P_n\ell(\theta, Z)$, then Theorem 5.2.1 implies that

$$\int |\widehat{L}_n(\theta) - L(\theta)| d\pi(\theta) \le \sqrt{\frac{8M^2r^2}{3n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{2}{\delta}\right]}$$

for all π with probability at least $1 - \delta$. Now, let $\theta_0 \in \Theta$ be arbitrary, and for $\epsilon > 0$ (to be chosen later) take π_0 to be uniform on $(r + \epsilon)\mathbb{B}_2^d$ and π to be uniform on $\theta_0 + \epsilon\mathbb{B}_2^d$. Then we immediately see that $D_{\rm kl}\left(\pi\|\pi_0\right) = d\log(1+\frac{r}{\epsilon})$. Moreover, we have $\int \widehat{L}_n(\theta)d\pi(\theta) \in \widehat{L}_n(\theta_0) \pm M\epsilon$ and similarly for $L(\theta)$, by the M-Lipschitz continuity of ℓ. For any fixed $\epsilon > 0$, we thus have

$$|\widehat{L}_n(\theta_0) - L(\theta_0)| \le 2M\epsilon + \sqrt{\frac{2M^2r^2}{3n}\left[d\log\left(1 + \frac{r}{\epsilon}\right) + \log\frac{2}{\delta}\right]}$$

simultaneously for all $\theta_0 \in \Theta$, with probability at least $1 - \delta$. By choosing $\epsilon = \frac{rd}{n}$ we obtain that with probability at least $1 - \delta$,

$$\sup_{\theta\in\Theta} |\widehat{L}_n(\theta) - L(\theta)| \le \frac{2Mrd}{n} + \sqrt{\frac{8M^2r^2}{3n}\left[d\log\left(1 + \frac{n}{d}\right) + \log\frac{2}{\delta}\right]}.$$

Thus, roughly, with high probability we have $|\widehat{L}_n(\theta) - L(\theta)| \le O(1)Mr\sqrt{\frac{d}{n}\log\frac{n}{d}}$ for all θ. $\diamond$

On the one hand, the result in Example 5.2.3 is satisfying: it applies to any Lipschitz function and provides a uniform bound. On the other hand, when we compare to the results achievable for specially structured linear function classes, then applying Rademacher complexity bounds—such as Proposition 4.4.9 and Example 4.4.10—we have somewhat weaker results, in that they depend on the dimension explicitly, while the Rademacher bounds do not exhibit this explicit dependence. This means they can potentially apply in infinite dimensional spaces that Example 5.2.3 cannot. We will give an example presently showing how to address some of these issues.

5.2.1 Relative bounds

In many cases, it is useful to have bounds that provide somewhat finer control than the bounds we have presented. Recall from our discussion of sub-Gaussian and sub-exponential random variables, especially the Bennett and Bernstein-type inequalities (Proposition 4.1.20), that if a random

variable X satisfies $|X| \le b$ but $\mathrm{Var}(X) \le \sigma^2 \ll b^2$, then X concentrates more quickly about its mean than the convergence provided by naive application of sub-Gaussian concentration with sub-Gaussian parameter $b^2/8$. To that end, we investigate an alternative to Theorem 5.2.1 that allows somewhat sharper control.

The approach is similar to our derivation in Theorem 5.2.1, where we show that the moment generating function of a quantity like $P_n f - Pf$ is small (Eq. (5.2.1)) and then relate this—via the Donsker-Varadhan change of measure in Theorem 5.1.1—to the quantities we wish to control. In the next proposition, we provide relative bounds on the deviations of functions from their means. To make this precise, let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$, and let $\sigma^2(f) := \mathrm{Var}(f(X))$ be the variance of functions in $\mathcal{F}$. We assume the class satisfies the Bernstein condition (4.1.7) with parameter b, that is,

$$\left|\mathbb{E}\left[(f(X) - Pf)^k\right]\right| \le \frac{k!}{2}\sigma^2(f)b^{k-2} \quad \text{for } k = 3, 4, \ldots. \tag{5.2.3}$$

This says that the second moment of functions $f \in \mathcal{F}$ bounds—with the additional boundedness-type constant b—the higher moments of functions in f. We then have the following result.

Proposition 5.2.4. *Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$ satisfying the Bernstein condition (5.2.3). Then for any $|\lambda| \le \frac{1}{2b}$, with probability at least $1 - \delta$,*

$$\lambda \int Pf d\pi(f) - \lambda^2 \int \sigma^2(f) d\pi(f) \le \lambda \int P_n f d\pi(f) + \frac{1}{n}\left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta}\right]$$

simultaneously for all $\pi \in \Pi$.

Proof We begin with an inequality on the moment generating function of random variables satisfying the Bernstein condition (4.1.7), that is, that $|\mathbb{E}[(X - \mu)^k]| \le \frac{k!}{2}\sigma^2 b^{k-2}$ for $k \ge 2$. In this case, Lemma 4.1.19 implies that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \le \exp(\lambda^2\sigma^2)$$

for $|\lambda| \le 1/(2b)$. As a consequence, for any f in our collection $\mathcal{F}$, we see that if we define

$$\Delta_n(f, \lambda) := \lambda\left[P_n f - Pf - \lambda\sigma^2(f)\right],$$

we have that

$$\mathbb{E}[\exp(n\Delta_n(f, \lambda))] = \mathbb{E}[\exp(\lambda(f(X) - Pf) - \lambda^2\sigma^2(f))]^n \le 1$$

for all n, $f \in \mathcal{F}$, and $|\lambda| \le \frac{1}{2b}$. Then, for any fixed measure π_0 on $\mathcal{F}$, Markov's inequality implies that

$$\mathbb{P}\left(\int \exp(n\Delta_n(f, \lambda)) d\pi_0(f) \ge \frac{1}{\delta}\right) \le \delta. \tag{5.2.4}$$

Now, as in the proof of Theorem 5.2.1, we use the Donsker-Varadhan Theorem 5.1.1 (change of measure), which implies that

$$n \int \Delta_n(f, \lambda) d\pi(f) \le D_{\rm kl}\left(\pi \| \pi_0\right) + \log \int \exp(n\Delta_n(f, \lambda)) d\pi_0(f)$$

for all distributions π. Using inequality (5.2.4), we obtain that with probability at least $1 - \delta$,

$$\int \Delta_n(f, \lambda) d\pi(f) \le \frac{1}{n}\left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta}\right]$$

for all π. As this holds for any fixed $|\lambda| \le 1/(2b)$, this gives the desired result by rearranging. $\square$

We would like to optimize over the bound in Proposition 5.2.4 by choosing the "best" λ. If we *could* choose the optimal λ, by rearranging Proposition 5.2.4 we would obtain the bound

$$\begin{aligned}\mathbb{E}_\pi[Pf] &\le \mathbb{E}_\pi[P_n f] + \inf_{\lambda > 0} \left\{ \lambda \mathbb{E}_\pi[\sigma^2(f)] + \frac{1}{n\lambda} \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta} \right] \right\} \\ &= \mathbb{E}_\pi[P_n f] + 2 \sqrt{\frac{\mathbb{E}_\pi[\sigma^2(f)]}{n} \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta} \right]}\end{aligned}$$

simultaneously for all π, with probability at least $1 - \delta$. The problem with this approach is two-fold: first, we cannot arbitrarily choose λ in Proposition 5.2.4, and second, the bound above depends on the unknown population variance $\sigma^2(f)$. It is thus of interest to understand situations in which we can obtain similar guarantees, but where we can replace unknown population quantities on the right side of the bound with known quantities.

To that end, let us consider the following condition, a type of relative error condition related to the Bernstein condition (4.1.7): for each $f \in \mathcal{F}$,

$$\sigma^2(f) \le bPf. \tag{5.2.5}$$

This condition is most natural when each of the functions f take nonnegative values—for example, when $f(X) = \ell(\theta, X)$ for some loss function ℓ and parameter θ of a model. If the functions f are nonnegative and upper bounded by b, then we certainly have $\sigma^2(f) \le \mathbb{E}[f(X)^2] \le b\mathbb{E}[f(X)] = bPf$, so that Condition (5.2.5) holds. Revisiting Proposition 5.2.4, we rearrange to obtain the following theorem.

Theorem 5.2.5. *Let $\mathcal{F}$ be a collection of functions satisfying the Bernstein condition (5.2.3) as in Proposition 5.2.4, and in addition, assume the variance-bounding condition (5.2.5). Then for any $0 \le \lambda \le \frac{1}{2b}$, with probability at least $1 - \delta$,*

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_n f] + \frac{\lambda b}{1 - \lambda b} \mathbb{E}_\pi[P_n f] + \frac{1}{\lambda(1 - \lambda b)} \frac{1}{n} \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta} \right]$$

for all π.

Proof We use condition (5.2.5) to see that

$$\lambda \mathbb{E}_\pi[Pf] - \lambda^2 b \mathbb{E}_\pi[Pf] \le \lambda \mathbb{E}_\pi[Pf] - \lambda^2 \mathbb{E}_\pi[\sigma^2(f)],$$

apply Proposition 5.2.4, and divide both sides of the resulting inequality by $\lambda(1 - \lambda b)$. $\square$

To make this uniform in λ, thus achieving a tighter bound (so that we need not pre-select λ), we choose multiple values of λ and apply a union bound. To that end, let $1 + \eta = \frac{1}{1 - \lambda b}$, or $\eta = \frac{\lambda b}{1 - \lambda b}$ and $\frac{1}{\lambda b (1 - \lambda b)} = \frac{(1 + \eta)^2}{\eta}$, so that the inequality in Theorem 5.2.1 is equivalent to

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_n f] + \eta \mathbb{E}_\pi[P_n f] + \frac{(1 + \eta)^2}{\eta} \frac{b}{n} \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta} \right].$$

Using that our choice of $\eta \in [0, 1]$, this implies

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_n f] + \eta \mathbb{E}_\pi[P_n f] + \frac{1}{\eta}\frac{b}{n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{1}{\delta}\right] + \frac{3b}{n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{1}{\delta}\right].$$

Now, take $\eta_1 = 1/n, \ldots, \eta_n = 1$. Then by optimizing over $\eta \in \{\eta_1, \ldots, \eta_n\}$ (which is equivalent, to within a $1/n$ factor, to optimizing over $0 < \eta \le 1$) and applying a union bound, we obtain

Corollary 5.2.6. *Let the conditions of Theorem 5.2.5 hold. Then with probability at least $1 - \delta$,*

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_n f] + 2\sqrt{\frac{b\mathbb{E}_\pi[P_n f]}{n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{n}{\delta}\right]} + \frac{1}{n}\left(\mathbb{E}_\pi[P_n f] + 5b\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{n}{\delta}\right]\right),$$

simultaneously for all π on $\mathcal{F}$.

Proof By a union bound, we have

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_n f] + \eta \mathbb{E}_\pi[P_n f] + \frac{1}{\eta}\frac{b}{n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{n}{\delta}\right] + \frac{3b}{n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{n}{\delta}\right]$$

for each $\eta \in \{1/n, \ldots, 1\}$. We consider two cases. In the first, assume that $\mathbb{E}_\pi[P_n f] \le \frac{b}{n}(D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{n}{\delta}$. Then taking $\eta = 1$ above evidently gives the result. In the second, we have $\mathbb{E}_\pi[P_n f] > \frac{b}{n}(D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{n}{\delta})$, and we can set

$$\eta_\star = \sqrt{\frac{\frac{b}{n}(D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{n}{\delta})}{\mathbb{E}_\pi[P_n f]}} \in (0, 1).$$

Choosing η to be the smallest value η_k in $\{\eta_1, \ldots, \eta_n\}$ with $\eta_k \ge \eta_\star$, so that $\eta_\star \le \eta \le \eta_\star + \frac{1}{n}$ then implies the claim in the corollary. □

5.2.2 A large-margin guarantee

Let us revisit the loss minimization approaches central to Section 4.4 and Example 5.2.3 in the context of Corollary 5.2.6. We will investigate an approach to achieve convergence guarantees that are (nearly) independent of dimension, focusing on 0-1 losses in a binary classification problem. Consider a binary classification problem with data $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$, where we make predictions $\langle\theta, x\rangle$ (or its sign), and for a *margin penalty* $\gamma \ge 0$ we define the loss

$$\ell_\gamma(\theta; (x, y)) = \mathbf{1}\left\{\langle\theta, x\rangle y \le \gamma\right\}.$$

We call the quantity $\langle\theta, x\rangle y$ the *margin* of θ on the pair (x, y), noting that when the margin is large, $\langle\theta, x\rangle$ has the same sign as y and is "confident" (i.e. far from zero). For shorthand, let us define the expected and empirical losses at margin γ by

$$L_\gamma(\theta) := P\ell_\gamma(\theta; (X, Y)) \quad \text{and} \quad \widehat{L}_\gamma(\theta) := P_n\ell_\gamma(\theta; (X, Y)).$$

Consider the following scenario: the data x lie in a ball of radius b, so that $\|x\|_2 \le b$; note that the losses ℓ_γ and ℓ_0 satisfy the Bernstein (5.2.3) and self-bounding (5.2.5) conditions with constant 1 as they take values in $\{0, 1\}$. We then have the following proposition.

Proposition 5.2.7. *Let the above conditions on the data (x, y) hold and let the margin $\gamma > 0$ and radius $r < \infty$. Then with probability at least $1 - \delta$,*

$$P(\langle\theta, X\rangle Y \le 0) \le \left(1 + \frac{1}{n}\right) P_n(\langle\theta, X\rangle Y \le \gamma) + \sqrt{8}\frac{rb\log\frac{n}{\delta}}{\gamma\sqrt{n}}\sqrt{P_n(\langle\theta, X\rangle Y \le \gamma)} + C\frac{r^2b^2\log\frac{n}{\delta}}{\gamma^2 n}$$

simultaneously for all $\|\theta\|_2 \le r$, where C is a numerical constant independent of the problem parameters.

Proposition 5.2.7 provides a "dimension-free" guarantee—it depends only on the ℓ_2-norms $\|\theta\|_2$ and $\|x\|_2$—so that it can apply equally in infinite dimensional spaces. The key to the inequality is that if we can find a large margin predictor—for example, one achieved by a support vector machine or, more broadly, by minimizing a convex loss of the form

$$\minimize_{\|\theta\|_2 \le r} \frac{1}{n}\sum_{i=1}^n \phi(\langle X_i, \theta\rangle Y_i)$$

for some decreasing convex $\phi : \mathbb{R} \to \mathbb{R}_+$, e.g. $\phi(t) = [1 - t]_+$ or $\phi(t) = \log(1 + e^{-t})$—then we get strong generalization performance guarantees relative to the empirical margin γ. As one particular instantiation of this approach, suppose we can obtain a perfect classifier with positive margin: a vector θ with $\|\theta\|_2 \le r$ such that $\langle\theta, X_i\rangle Y_i \ge \gamma$ for each $i = 1, \ldots, n$. Then Proposition 5.2.7 guarantees that

$$P(\langle\theta, X\rangle Y \le 0) \le C\frac{r^2b^2\log\frac{n}{\delta}}{\gamma^2 n}$$

with probability at least $1 - \delta$.

Proof Let π_0 be $\mathsf{N}(0, \tau^2 I)$ for some $\tau > 0$ to be chosen, and let π be $\mathsf{N}(\widehat{\theta}, \tau^2 I)$ for some $\widehat{\theta} \in \mathbb{R}^d$ satisfying $\|\widehat{\theta}\|_2 \le r$. Then Corollary 5.2.6 implies that

$$\begin{aligned}
&\mathbb{E}_\pi[L_\gamma(\theta)] \\
&\le \mathbb{E}_\pi[\widehat{L}_\gamma(\theta)] + 2\sqrt{\frac{\mathbb{E}_\pi[\widehat{L}_\gamma(\theta)]}{n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{n}{\delta}\right]} + \frac{1}{n}\left(\mathbb{E}_\pi[\widehat{L}_\gamma(\theta)] + C\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{n}{\delta}\right]\right) \\
&\le \mathbb{E}_\pi[\widehat{L}_\gamma(\theta)] + 2\sqrt{\frac{\mathbb{E}_\pi[\widehat{L}_\gamma(\theta)]}{n}\left[\frac{r^2}{2\tau^2} + \log\frac{n}{\delta}\right]} + \frac{1}{n}\left(\mathbb{E}_\pi[\widehat{L}_\gamma(\theta)] + C\left[\frac{r^2}{2\tau^2} + \log\frac{n}{\delta}\right]\right)
\end{aligned}$$

simultaneously for all $\widehat{\theta}$ satisfying $\|\widehat{\theta}\|_2 \le r$ with probability at least $1 - \delta$, where we have used that $D_{\rm kl}\left(\mathsf{N}(\theta, \tau^2 I)\|\mathsf{N}(0, \tau^2 I)\right) = \|\theta\|_2^2/(2\tau^2)$.

Let us use the margin assumption. Note that if $Z \sim \mathsf{N}(0, \tau^2 I)$, then for any fixed θ_0, x, y we have

$$\ell_0(\theta_0; (x, y)) - \mathbb{P}(Z^\top x \ge \gamma) \le \mathbb{E}[\ell_\gamma(\theta_0 + Z; (x, y))] \le \ell_{2\gamma}(\theta_0; (x, y)) + \mathbb{P}(Z^\top x \ge \gamma)$$

where the middle expectation is over $Z \sim \mathsf{N}(0, \tau^2 I)$. Using the $\tau^2\|x\|_2^2$-sub-Gaussianity of $Z^\top x$, we can obtain immediately that if $\|x\|_2 \le b$, we have

$$\ell_0(\theta_0; (x, y)) - \exp\left(-\frac{\gamma^2}{2\tau^2 b^2}\right) \le \mathbb{E}[\ell_\gamma(\theta_0 + Z; (x, y))] \le \ell_{2\gamma}(\theta_0; (x, y)) + \exp\left(-\frac{\gamma^2}{2\tau^2 b^2}\right).$$

Returning to our earlier bound, we evidently have that if $\|x\|_2 \le b$ for all $x \in \mathcal{X}$, then with probability at least $1-\delta$, simultaneously for all $\theta \in \mathbb{R}^d$ with $\|\theta\|_2 \le r$,

$$L_0(\theta) \le \widehat{L}_{2\gamma}(\theta) + 2\exp\left(-\frac{\gamma^2}{2\tau^2 b^2}\right) + 2\sqrt{\frac{\widehat{L}_{2\gamma}(\theta) + \exp(-\frac{\gamma^2}{2\tau^2 b^2})}{n}\left[\frac{r^2}{2\tau^2} + \log\frac{n}{\delta}\right]} + \frac{1}{n}\left(\widehat{L}_{2\gamma}(\theta) + \exp\left(-\frac{\gamma^2}{2\tau^2 b^2}\right) + C\left[\frac{r^2}{2\tau^2} + \log\frac{n}{\delta}\right]\right).$$

Setting $\tau^2 = \frac{\gamma^2}{2b^2 \log n}$, we immediately see that for any choice of margin $\gamma > 0$, we have with probability at least $1-\delta$ that

$$L_0(\theta) \le \widehat{L}_{2\gamma}(\theta) + \frac{2b}{n} + 2\sqrt{\frac{1}{n}\left[\widehat{L}_{2\gamma}(\theta) + \frac{b}{n}\right]\left[\frac{r^2 b^2 \log n}{2\gamma^2} + \log\frac{n}{\delta}\right]} + \frac{1}{n}\left(\widehat{L}_{2\gamma}(\theta) + \frac{1}{n} + C\left[\frac{r^2 b^2 \log n}{2\gamma^2} + \log\frac{n}{\delta}\right]\right)$$

for all $\|\theta\|_2 \le r$.

Rewriting (replacing 2γ with γ) and recognizing that with no loss of generality we may take γ such that $rb \ge \gamma$ gives the claim of the proposition. □

5.2.3 A mutual information bound

An alternative perspective of the PAC-Bayesian bounds that Theorem 5.2.1 gives is to develop bounds based on mutual information, which is also central to the interactive data analysis setting in the next section. We present a few results along these lines here. Assume the setting of Theorem 5.2.1, so that $\mathcal{F}$ consists of σ^2-sub-Gaussian functions. Let us assume the following observational model: we observe $X_1^n \stackrel{\text{iid}}{\sim} P$, and then conditional on the sample X_1^n, draw a (random) function $F \in \mathcal{F}$ following the distribution $\pi(\cdot \mid X_1^n)$. Assuming the prior π_0 is fixed, Theorem 5.2.1 guarantees that with probability at least $1-\delta$ over X_1^n,

$$\mathbb{E}[(P_n F - PF)^2 \mid X_1^n] \le \frac{8\sigma^2}{3n}\left[D_{\rm kl}\left(\pi(\cdot \mid X_1^n) \| \pi_0\right) + \log\frac{2}{\delta}\right],$$

where the expectation is taken over $F \sim \pi(\cdot \mid X_1^n)$, leaving the sample fixed. Now, consider choosing π_0 to be the average over all samples X_1^n of π, that is, $\pi_0(\cdot) = \mathbb{E}_P[\pi(\cdot \mid X_1^n)]$, the expectation taken over $X_1^n \stackrel{\text{iid}}{\sim} P$. Then by definition of mutual information,

$$I(F; X_1^n) = \mathbb{E}_P\left[D_{\rm kl}\left(\pi(\cdot \mid X_1^n) \| \pi_0\right)\right],$$

and by Markov's inequality we have

$$\mathbb{P}(D_{\rm kl}\left(\pi(\cdot \mid X_1^n) \| \pi_0\right) \ge K \cdot I(F; X_1^n)) \le \frac{1}{K}$$

for all $K \ge 0$. Combining these, we obtain the following corollary.

Corollary 5.2.8. *Let F be chosen according to any distribution $\pi(\cdot \mid X_1^n)$ conditional on the sample X_1^n. Then with probability at least $1 - \delta_0 - \delta_1$ over the sample $X_1^n \stackrel{\text{iid}}{\sim} P$,*

$$\mathbb{E}[(P_n F - PF)^2 \mid X_1^n] \le \frac{8\sigma^2}{3n} \left[\frac{I(F; X_1^n)}{\delta_0} + \log \frac{2}{\delta_1} \right].$$

This corollary shows that if we have any procedure—say, a learning procedure or otherwise—that limits the information between a sample X_1^n and an output F, then we are guaranteed that F generalizes. Tighter analyses of this are possible, though not our focus here, just that already there should be an inkling that limiting information between input samples and outputs may be fruitful.

5.3 Interactive data analysis

A major challenge in modern data analysis is that analyses are often not the classical statistics and scientific method setting. In the scientific method—forgive me for being a pedant—one proposes a hypothesis, the status quo or some other belief, and then designs an experiment to falsify that hypothesis. Then, upon performing the experiment, there are only two options: either the experimental results contradict the hypothesis (that is, we must reject the null) so that the hypothesis is false, or the hypothesis remains consistent with available data. In the classical (Fisherian) statistics perspective, this typically means that we have a single null hypothesis H_0 before observing a sample, we draw a sample $X \in \mathcal{X}$, and then for some test statistic $T : \mathcal{X} \to \mathbb{R}$ with observed value $t_{\text{observed}} = T(X)$, we compute the probability under the null of observing something as extreme as what we observed, that is, the p-value $p = P_{H_0}(T(X) \ge t_{\text{observed}})$.

Yet modern data analyses are distant from this pristine perspective for many reasons. The simplest is that we often have a number of hypotheses we wish to test, not a single one. For example, in biological applications, we may wish to investigate the associations between the expression of number of genes and a particular phenotype or disease; each gene j then corresponds to a null hypothesis $H_{0,j}$ that gene j is independent of the phenotype. There are numerous approaches to addressing the challenges associated with such multiple testing problems—such as false discovery rate control, familywise error rate control, and others—with whole courses devoted to the challenges.

Even these approaches to multiple testing and high-dimensional problems do not truly capture modern data analyses, however. Indeed, in many fields, researchers use one or a few main datasets, writing papers and performing multiple analyses on the same dataset. For example, in medicine, the UK Biobank dataset [163] has several thousand citations (as of 2023), many of which build on one another, with early studies coloring the analyses in subsequent studies. Even in situations without a shared dataset, analyses present researchers with huge degrees of freedom and choice. A researcher may study a summary statistic of his or her sampled data, or a plot of a few simple relationships, performing some simple data exploration—which statisticians and scientists have advocated for 50 years, dating back at least to John Tukey!—but this means that there are huge numbers of *potential* comparisons a researcher might make (that he or she does not). This "garden of forking paths," as Gelman and Loken [91] term it, causes challenges even when researchers are not "p-hacking" or going on a "fishing expedition" to try to find publishable results. The problem in these studies and approaches is that, because we make decisions that may, even only in a small way, depend on the data observed, we have invalidated all classical statistical analyses.

To that end, we now consider *interactive* data analyses, where we perform data analyses sequentially, computing new functions on a fixed sample $X_1, \ldots, X_n$ after observing some initial

information about the sample. The starting point of our approach is similar to our analysis of PAC-Bayesian learning and generalization: we observe that if the function we decide to compute on the data X_1^n is chosen without much information about the data at hand, then its value on the sample should be similar to its values on the full population. This insight dovetails with what we have seen thus far, that appropriate "stability" in information can be useful and guarantee good future performance.

5.3.1 The interactive setting

We do not consider the interactive data analysis setting in full, rather, we consider a stylized approach to the problem, as it captures many of the challenges while being broad enough for different applications. In particular, we focus on the *statistical queries* setting, where a data analyst wishes to evaluate expectations

$$\mathbb{E}_P[\phi(X)] \tag{5.3.1}$$

of various functionals $\phi : \mathcal{X} \to \mathbb{R}$ under the population P using a sample $X_1^n \stackrel{\rm iid}{\sim} P$. Certainly, numerous problems problems are solvable using statistical queries (5.3.1). Means use $\phi(x) = x$, while we can compute variances using the two statistical queries $\phi_1(x) = x$ and $\phi_2(x) = x^2$, as $\mathrm{Var}(X) = \mathbb{E}_P[\phi_2(X)] - \mathbb{E}_P[\phi_1(X)]^2$.

Classical algorithms for the statistical query problem simply return sample means $P_n\phi := \frac{1}{n}\sum_{i=1}^n \phi(X_i)$ given a query $\phi : \mathcal{X} \to \mathbb{R}$. When the number of queries to be answered is not chosen adaptively, this means we can typically answer a large number relatively accurately; indeed, if we have a finite collection Φ of σ^2-sub-Gaussian $\phi : \mathcal{X} \to \mathbb{R}$, then we of course have

$$\mathbb{P}\left(\max_{\phi\in\Phi} |P_n\phi - P\phi| \ge \sqrt{\frac{2\sigma^2}{n}(\log(2|\Phi|) + t)}\right) \le e^{-t^2} \quad \text{for } t \ge 0$$

by Corollary 4.1.10 (sub-Gaussian concentration) and a union bound. Thus, so long as $|\Phi|$ is not exponential in the sample size n, we expect uniformly high accuracy.

Example 5.3.1 (Risk minimization via statistical queries)**:** Suppose that we are in the loss-minimization setting (4.4.2), where the losses $\ell(\theta, X_i)$ are convex and differentiable in θ. Then gradient descent applied to $\widehat{L}_n(\theta) = P_n\ell(\theta, X)$ will converge to a minimizing value of $\widehat{L}_n$. We can evidently implement gradient descent by a sequence of statistical queries $\phi(x) = \nabla_\theta \ell(\theta, x)$, iterating

$$\theta^{(k+1)}) = \theta^{(k)} - \alpha_k P_n \phi^{(k)}, \tag{5.3.2}$$

where $\phi^{(k)} = \nabla_\theta \ell(\theta^{(k)}, x)$ and α_k is a stepsize. $\diamond$

One issue with the example (5.3.1) is that we are *interacting* with the dataset, because each sequential query $\phi^{(k)}$ depends on the previous $k-1$ queries. (Our results on uniform convergence of empirical functionals and related ideas address many of these challenges, so that the result of the process (5.3.2) will be well-behaved regardless of the interactivity.)

We consider an interactive version of the statistical query estimation problem. In this version, there are two parties: an analyst (or statistician or learner), who issues queries $\phi : \mathcal{X} \to \mathbb{R}$, and a mechanism that answers the queries to the analyst. We index our functionals ϕ by $t \in \mathcal{T}$ for a (possibly infinite) set $\mathcal{T}$, so we have a collection $\{\phi_t\}_{t\in\mathcal{T}}$. In this context, we thus have the following scheme:

Input: Sample X_1^n drawn i.i.d. P, collection $\{\phi_t\}_{t\in\mathcal{T}}$ of possible queries
Repeat: for $k = 1, 2, \ldots$

i. Analyst chooses index $T_k \in \mathcal{T}$ and query $\phi := \phi_{T_k}$

ii. Mechanism responds with answer A_k approximating $P\phi = \mathbb{E}_P[\phi(X)]$ using X_1^n

Figure 5.1: The interactive statistical query setting

Of interest in the iteration 5.1 is that we *interactively* choose $T_1, T_2, \ldots, T_k$, where the choice T_i may depend on our approximations of $\mathbb{E}_P[\phi_{T_j}(X)]$ for $j < i$, that is, on the results of our previous queries. Even more broadly, the analyst may be able to choose the index T_k in alternative ways depending on the sample X_1^n, and our goal is to still be able to accurately compute expectations $P\phi_T = \mathbb{E}_P[\phi_T(X)]$ when the index T may depend on X_1^n. The setting in Figure 5.1 clearly breaks with the classical statistical setting in which an analysis is pre-specified before collecting data, but more closely captures modern data exploration practices.

5.3.2 Second moment errors and mutual information

The starting point of our derivation is the following result, which follows from more or less identical arguments to those for our PAC-Bayesian bounds earlier.

Theorem 5.3.2. *Let $\{\phi_t\}_{t\in\mathcal{T}}$ be a collection of σ^2-sub-Gaussian functions $\phi_t : \mathcal{X} \to \mathbb{R}$. Then for any random variable T and any $\lambda > 0$,*

$$\mathbb{E}[(P_n\phi_T - P\phi_T)^2] \le \frac{1}{\lambda}\left[I(X_1^n; T) - \frac{1}{2}\log\left[1 - 2\lambda\sigma^2/n\right]_+\right]$$

and

$$|\mathbb{E}[P_n\phi_T] - \mathbb{E}[P\phi_T]| \le \sqrt{\frac{2\sigma^2}{n} I(X_1^n; T)}$$

where the expectations are taken over T and the sample X_1^n.

Proof The proof is similar to that of our first basic PAC-Bayes result in Theorem 5.2.1. Let us assume w.l.o.g. that $P\phi_t = 0$ for all $t \in \mathcal{T}$, noting that then $P_n\phi_t$ is σ^2/n-sub-Gaussian. We prove the first result first. Lemma 5.2.2 implies that $\mathbb{E}[\exp(\lambda(P_n\phi_t)^2)] \le \left[1 - 2\lambda\sigma^2/n\right]_+^{-1/2}$ for each $t \in \mathcal{T}$. As a consequence, we obtain via the Donsker-Varadhan equality (Theorem 5.1.1) that

$$\begin{aligned}
\lambda\mathbb{E}\left[\int (P_n\phi_t)^2 d\pi(t)\right] &\overset{(i)}{\le} \mathbb{E}[D_{\rm kl}\left(\pi\|\pi_0\right)] + \mathbb{E}\left[\log \int \exp(\lambda(P_n\phi_t)^2) d\pi_0(t)\right] \\
&\overset{(ii)}{\le} \mathbb{E}[D_{\rm kl}\left(\pi\|\pi_0\right)] + \log \mathbb{E}\left[\int \exp(\lambda(P_n\phi_t)^2) d\pi_0(t)\right] \\
&\overset{(iii)}{\le} \mathbb{E}[D_{\rm kl}\left(\pi\|\pi_0\right)] - \frac{1}{2}\log\left[1 - 2\lambda\sigma^2/n\right]_+
\end{aligned}$$

for all distributions π on $\mathcal{T}$, which may depend on P_n, where the expectation $\mathbb{E}$ is taken over the sample $X_1^n \overset{\rm iid}{\sim} P$. (Here inequality (i) is Theorem 5.1.1, inequality (ii) is Jensen's inequality, and

inequality (*iii*) is Lemma 5.2.2.) Now, let π_0 be the marginal distribution on T (marginally over all observations X_1^n), and let π denote the posterior of T conditional on the sample X_1^n. Then $\mathbb{E}[D_{\rm kl}(\pi\|\pi_0)] = I(X_1^n;T)$ by definition of the mutual information, giving the bound on the squared error.

For the second result, note that the Donsker-Varadhan equality implies

$$\lambda\mathbb{E}\left[\int P_n\phi_t d\pi(t)\right] \le \mathbb{E}[D_{\rm kl}(\pi\|\pi_0)] + \log\int \mathbb{E}[\exp(\lambda P_n\phi_t)]d\pi_0(t) \le I(X_1^n;T) + \frac{\lambda^2\sigma^2}{2n}.$$

Dividing both sides by λ gives $\mathbb{E}[P_n\phi_T] \le \sqrt{2\sigma^2 I(X_1^n;T)/n}$, and performing the same analysis with $-\phi_T$ gives the second result of the theorem. □

The key in the theorem is that if the mutual information—the Shannon information—$I(X;T)$ between the sample X and T is small, then the expected squared error can be small. To make this a bit clearer, let us choose values for λ in the theorem; taking $\lambda = \frac{n}{2e\sigma^2}$ gives the following corollary.

Corollary 5.3.3. *Let the conditions of Theorem 5.3.2 hold. Then*

$$\mathbb{E}[(P_n\phi_T - P\phi_T)^2] \le \frac{2e\sigma^2}{n} I(X_1^n;T) + \frac{5\sigma^2}{4n}.$$

Consequently, if we can limit the amount of information any particular query T (i.e., ϕ_T) contains about the actual sample X_1^n, then guarantee reasonably high accuracy in the second moment errors $(P_n\phi_T - P\phi_T)^2$.

5.3.3 Limiting interaction in interactive analyses

Let us now return to the interactive data analysis setting of Figure 5.1, where we recall the stylized application of estimating mean functionals $P\phi$ for $\phi \in \{\phi_t\}_{t\in\mathcal{T}}$. To motivate a more careful approach, we consider a simple example to show the challenges that may arise even with only a single "round" of interactive data analysis. Naively answering queries accurately—using the mechanism $P_n\phi$ that simply computes the sample average—can easily lead to problems:

Example 5.3.4 (A stylized correlation analysis)**:** Consider the following stylized genetics experiment. We observe vectors $X \in \{-1,1\}^k$, where $X_j = 1$ if gene j is expressed and -1 otherwise. We also observe phenotypes $Y \in \{-1,1\}$, where $Y = 1$ indicates appearance of the phenotype. In our setting, we will assume that the vectors X are uniform on $\{-1,1\}^k$ and independent of Y, but an experimentalist friend of ours wishes to know if there exists a vector v with $\|v\|_2 = 1$ such that the correlation between v^TX and Y is high, meaning that v^TX is associated with Y. In our notation here, we have index set $\{v \in \mathbb{R}^k \mid \|v\|_2 = 1\}$, and by Example 4.1.6, Hoeffding's lemma, and the independence of the coordinates of X we have that v^TXY is $\|v\|_2^2/4 = 1/4$-sub-Gaussian. Now, we recall the fact that if Z_j, $j = 1,\ldots,k$, are σ^2-sub-Gaussian, then for any $p \ge 1$, we have

$$\mathbb{E}[\max_j |Z_j|^p] \le (Cp\sigma^2\log k)^{p/2}$$

for a numerical constant C. That is, powers of sub-Gaussian maxima grow at most logarithmically. Indeed, by Theorem 4.1.11, we have for any $q \ge 1$ by Hölder's inequality that

$$\mathbb{E}[\max_j |Z_j|^p] \le \mathbb{E}\left[\sum_j |Z_j|^{pq}\right]^{1/q} \le k^{1/q}(Cpq\sigma^2)^{p/2},$$

and setting $q = \log k$ gives the inequality. Thus, we see that for any *a priori* fixed $v_1, \ldots, v_k, v_{k+1}$, we have

$$\mathbb{E}[\max_j (v_j^T (P_n Y X))^2] \le O(1) \frac{\log k}{n}.$$

If instead we allow a *single* interaction, the problem is different. We issue queries associated with $v = e_1, \ldots, e_k$, the k standard basis vectors; then we simply set $V_{k+1} = P_n Y X / \left\| P_n Y X \right\|_2$. Then evidently

$$\mathbb{E}[(V_{k+1}^T (P_n Y X))^2] = \mathbb{E}[\left\| P_n Y X \right\|_2^2] = \frac{k}{n},$$

which is exponentially larger than in the non-interactive case. That is, if an analyst is allowed to interact with the dataset, he or she may be able to discover very large correlations that are certainly false in the population, which in this case has $PXY = 0$. $\diamond$

Example 5.3.4 shows that, without being a little careful, substantial issues may arise in interactive data analysis scenarios. When we consider our goal more broadly, which is to be able to provide accurate approximations to $P\phi$ for queries ϕ chosen adaptively for any population distribution P and $\phi : \mathcal{X} \to [-1, 1]$, it is possible to construct quite perverse situations, where if we compute sample expectations $P_n \phi$ exactly, one round of interaction is sufficient to find a query ϕ for which $P_n \phi - P\phi \ge 1$.

Example 5.3.5 (Exact query answering allows arbitrary corruption)**:** Suppose we draw a sample X_1^n of size n on a sample space $\mathcal{X} = [m]$ with $X_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}([m])$, where $m \ge 2n$. Let Φ be the collection of all functions $\phi : [m] \to [-1, 1]$, so that $\mathbb{P}(|P_n \phi - P\phi| \ge t) \le \exp(-nt^2/2)$ for any fixed ϕ. Suppose that in the interactive scheme in Fig. 5.1, we simply release answers $A = P_n \phi$. Consider the following query:

$$\phi(x) = n^{-x} \quad \text{for } x = 1, 2, \ldots, m.$$

Then by inspection, we see that

$$\begin{aligned} P_n \phi &= \sum_{j=1}^m n^{-j} \operatorname{card}(\{X_i \mid X_i = j\}) \\ &= \frac{1}{n} \operatorname{card}(\{X_i \mid X_i = 1\}) + \frac{1}{n^2} \operatorname{card}(\{X_i \mid X_i = 1\}) + \cdots + \frac{1}{n^m} \operatorname{card}(\{X_i \mid X_i = m\}). \end{aligned}$$

It is clear that given $P_n \phi$, we can reconstruct the sample counts exactly. Then if we define a second query $\phi_2(x) = 1$ for $x \in X_1^n$ and $\phi_2(x) = -1$ for $x \notin X_1^n$, we see that $P\phi_2 \le \frac{n}{m} - 1$, while $P_n \phi_2 = 1$. The gap is thus

$$\mathbb{E}[P_n \phi_2 - P\phi_2] \ge 2 - \frac{n}{m} \ge 1,$$

which is essentially as bad as possible. $\diamond$

More generally, when one performs an interactive data analysis (e.g. as in Fig. 5.1), adapting hypotheses while interacting with a dataset, it is not a question of statistical significance or multiplicity control for the analysis one does, but for *all the possible analyses* one might have done otherwise. Given the branching paths one might take in an analysis, it is clear that we require some care.

With that in mind, we consider the desiderata for techniques we might use to control information in the indices we select. We seek some type of *stability* in the information algorithms provide to a data analyst—intuitively, if small changes to a sample do not change the behavior of an analyst substantially, then we expect to obtain reasonable generalization bounds. If outputs of a particular analysis procedure carry little information about a particular sample (but instead provide information about a population), then Corollary 5.3.3 suggests that any estimates we obtain should be accurate.

To develop this stability theory, we require two conditions: first, that whatever quantity we develop for stability should *compose adaptively*, meaning that if we apply two (randomized) algorithms to a sample, then if both are appropriately stable, even if we choose the second algorithm because of the output of the first in arbitrary ways, they should remain jointly stable. Second, our notion should bound the mutual information $I(X_1^n; T)$ between the sample X_1^n and T. Lastly, we remark that this control on the mutual information has an additional benefit: by the data processing inequality, any downstream analysis we perform that depends only on T necessarily satisfies the same stability and information guarantees as T, because if we have the Markov chain $X_1^n \to T \to V$ then $I(X_1^n; V) \le I(X_1^n; T)$.

We consider randomized algorithms $A : \mathcal{X}^n \to \mathcal{A}$, taking values in our index set $\mathcal{A}$, where $A(X_1^n) \in \mathcal{A}$ is a random variable that depends on the sample X_1^n. For simplicity in derivation, we abuse notation in this section, and for random variables X and Y with distributions P and Q respectively, we denote

$$D_{\rm kl}\left(X \| Y\right) := D_{\rm kl}\left(P \| Q\right).$$

We then ask for a type of leave-one-out stability for the algorithms A, where A is insensitive to the changes of a single example (on average).

Definition 5.1. *Let $\varepsilon \ge 0$. A randomized algorithm $A : \mathcal{X}^n \to \mathcal{A}$ is ε-*KL*-stable if for each $i \in \{1, \ldots, n\}$ there is a randomized $A_i : \mathcal{X}^{n-1} \to \mathcal{A}$ such that for every sample $x_1^n \in \mathcal{X}^n$,*

$$\frac{1}{n} \sum_{i=1}^n D_{\rm kl}\left(A(x_1^n) \| A_i(x_{\setminus i})\right) \le \varepsilon.$$

Examples may be useful to understand Definition 5.1.

Example 5.3.6 (KL-stability in mean estimation: Gaussian noise addition)**:** Suppose we wish to estimate a mean, and that $x_i \in [-1, 1]$ are all real-valued. Then a natural statistic is to simply compute $A(x_1^n) = \frac{1}{n} \sum_{i=1}^n x_i$. In this case, without randomization, we will have infinite KL-divergence between $A(x_1^n)$ and $A_i(x_{\setminus i})$. If instead we set $A(x_1^n) = \frac{1}{n} \sum_{i=1}^n x_i + Z$ for $Z \sim \mathsf{N}(0, \sigma^2)$, and similarly $A_i = \frac{1}{n} \sum_{j \ne i} x_j + Z$, then we have (recall Example 2.1.7)

$$\frac{1}{n} \sum_{i=1}^n D_{\rm kl}\left(A(x_1^n) \| A(x_{\setminus i})\right) = \frac{1}{2n\sigma^2} \sum_{i=1}^n \frac{1}{n^2} x_i^2 \le \frac{1}{2\sigma^2 n^2},$$

so that a the sample mean of a bounded random variable perturbed with Guassian noise is $\varepsilon = \frac{1}{2\sigma^2 n^2}$-KL-stable. $\diamond$

We can consider other types of noise addition as well.

Example 5.3.7 (KL-stability in mean estimation: Laplace noise addition)**:** Let the conditions of Example 2.1.7 hold, but suppose instead of Gaussian noise we add scaled Laplace noise,

that is, $A(x_1^n) = \frac{1}{n}\sum_{i=1}^n x_i + Z$ for Z with density $p(z) = \frac{1}{2\sigma}\exp(-|z|/\sigma)$, where $\sigma > 0$. Then using that if $L_{\mu,\sigma}$ denotes the Laplace distribution with shape σ and mean μ, with density $p(z) = \frac{1}{2\sigma}\exp(-|z-\mu|/\sigma)$, we have

$$\begin{aligned} D_{\rm kl}\left(L_{\mu_0,\sigma}\|L_{\mu_1,\sigma}\right) &= \frac{1}{\sigma^2}\int_0^{|\mu_1-\mu_0|}\exp(-z/\sigma)(|\mu_1-\mu_0|-z)dz \\ &= \exp\left(-\frac{|\mu_1-\mu_0|}{\sigma}\right) - 1 + \frac{|\mu_1-\mu_0|}{\sigma} \le \frac{|\mu_1-\mu_0|^2}{2\sigma^2}, \end{aligned}$$

we see that in this case the sample mean of a bounded random variable perturbed with Laplace noise is $\varepsilon = \frac{1}{2\sigma^2 n^2}$-KL-stable, where σ is the shape parameter. $\diamond$

The two key facts are that KL-stable algorithms compose adaptively and that they bound mutual information in independent samples.

Lemma 5.3.8. *Let $A : \mathcal{X}^n \to \mathcal{A}_0$ and $A' : \mathcal{A}_0 \times \mathcal{X} \to \mathcal{A}_1$ be ε and ε'-KL-stable algorithms, respectively. Then the (randomized) composition $A' \circ A(x_1^n) = A'(A(x_1^n), x_1^n)$ is $\varepsilon + \varepsilon'$-KL-stable. Moreover, the pair $(A' \circ A(x_1^n), A(x_1^n))$ is $\varepsilon + \varepsilon'$-KL-stable.*

Proof Let A_i and A_i' be the promised sub-algorithms in Definition 5.1. We apply the data processing inequality, which implies for each i that

$$D_{\rm kl}\left(A'(A(x_1^n), x_1^n)\|A_i'(A_i(x_{\setminus i}), x_{\setminus i})\right) \le D_{\rm kl}\left(A'(A(x_1^n), x_1^n), A(x_1^n)\|A_i'(A_i(x_{\setminus i}), x_{\setminus i}), A_i(x_{\setminus i})\right).$$

We require a bit of notational trickery now. Fixing i, let $P_{A,A'}$ be the joint distribution of $A'(A(x_1^n), x_1^n)$ and $A(x_1^n)$ and $Q_{A,A'}$ the joint distribution of $A_i'(A_i(x_{\setminus i}), x_{\setminus i})$ and $A_i(x_{\setminus i})$, so that they are both distributions over $\mathcal{A}_1 \times \mathcal{A}_0$. Let $P_{A'|a}$ be the distribution of $A'(t, x_1^n)$ and similarly $Q_{A'|a}$ is the distribution of $A_i'(t, x_{\setminus i})$. Note that A', A_i' both "observe" x, so that using the chain rule (2.1.6) for KL-divergences, we have

$$\begin{aligned} D_{\rm kl}\left(A' \circ A, A\|A_i' \circ A_i, A_i\right) &= D_{\rm kl}\left(P_{A,A'}\|Q_{A,A'}\right) \\ &= D_{\rm kl}\left(P_A\|Q_A\right) + \int D_{\rm kl}\left(P_{A'|t}\|Q_{A'|t}\right) dP_A(t) \\ &= D_{\rm kl}\left(A\|A_i\right) + \mathbb{E}_A[D_{\rm kl}\left(A'(A, x_1^n)\|A_i'(A, x_1^n)\right)]. \end{aligned}$$

Summing this from $i = 1$ to n yields

$$\frac{1}{n}\sum_{i=1}^n D_{\rm kl}\left(A' \circ A\|A_i' \circ A_i\right) \le \frac{1}{n}\sum_{i=1}^n D_{\rm kl}\left(A\|A_i\right) + \mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^n D_{\rm kl}\left(A'(A, x_1^n)\|A_i'(A, x_1^n)\right)\right] \le \varepsilon + \varepsilon',$$

as desired. $\square$

The second key result is that KL-stable algorithms also bound the mutual information of a random function.

Lemma 5.3.9. *Let X_i be independent. Then for any random variable A,*

$$I(A; X_1^n) \le \sum_{i=1}^n I(A; X_i \mid X_{\setminus i}) = \sum_{i=1}^n \int D_{\rm kl}\left(A(x_1^n)\|A_i(x_{\setminus i})\right) dP(x_1^n),$$

where $A_i(x_{\setminus i}) = A(x_1^{i-1}, X_i, x_{i+1}^n)$ is the random realization of A conditional on $X_{\setminus i} = x_{\setminus i}$.

Proof Without loss of generality, we assume A and X are both discrete. In this case, we have

$$I(A; X_1^n) = \sum_{i=1}^n I(A; X_i \mid X_1^{i-1}) = \sum_{i=1}^n H(X_i \mid X_1^{i-1}) - H(X_i \mid A, X_1^{i-1}).$$

Now, because the X_i follow a product distribution, $H(X_i \mid X_1^{i-1}) = H(X_i)$, while $H(X_i \mid A, X_1^{i-1}) \geq H(X_i \mid A, X_{\setminus i})$ because conditioning reduces entropy. Consequently, we have

$$I(A; X_1^n) \leq \sum_{i=1}^n H(X_i) - H(X_i \mid A, X_{\setminus i}) = \sum_{i=1}^n I(A; X_i \mid X_{\setminus i}).$$

To see the final equality, note that

$$\begin{aligned} I(A; X_i \mid X_{\setminus i}) &= \int_{\mathcal{X}^{n-1}} I(A; X_i \mid X_{\setminus i} = x_{\setminus i}) dP(x_{\setminus i}) \\ &= \int_{\mathcal{X}^{n-1}} \int_{\mathcal{X}} D_{\rm kl}\left(A(x_1^n) \| A(x_{1:i-1}, X_i, x_{i+1:n})\right) dP(x_i) dP(x_{\setminus i}) \end{aligned}$$

by definition of mutual information as $I(X; Y) = \mathbb{E}_X[D_{\rm kl}\left(P_{Y|X} \| P_Y\right)]$. □

Combining Lemmas 5.3.8 and 5.3.9, we see (nearly) immediately that KL stability implies a mutual information bound, and consequently even interactive KL-stable algorithms maintain bounds on mutual information.

Proposition 5.3.10. *Let $A_1, \ldots, A_k$ be ε_i-KL-stable procedures, respectively, composed in any arbitrary sequence. Let X_i be independent. Then*

$$\frac{1}{n} I(A_1, \ldots, A_k; X_1^n) \leq \sum_{i=1}^k \varepsilon_i.$$

Proof Applying Lemma 5.3.9,

$$I(A_1^k; X_1^n) \leq \sum_{i=1}^n I(A_1^k; X_i \mid X_{\setminus i}) = \sum_{j=1}^k \sum_{i=1}^n I(A_j; X_i \mid X_{\setminus i}, A_1^{j-1}).$$

Fix an index j and for shorthand, let $A = A$ and $A' = (A_1, \ldots, A_{j-1})$ be the first $j-1$ procedures. Then expanding the final mutual information term and letting ν denote the distribution of A', we have

$$I(A; X_i \mid X_{\setminus i}, A') = \int D_{\rm kl}\left(A(a', x_1^n) \| \overline{A}(a', x_{\setminus i})\right) dP(x_i \mid A' = a', x_{\setminus i}) dP^{n-1}(x_{\setminus i}) d\nu(a' \mid x_{\setminus i})$$

where $A(a', x_1^n)$ is the (random) procedure A on inputs x_1^n and a', while $\overline{A}(a', x_{\setminus i})$ denotes the (random) procedure A on input $a', x_{\setminus i}, X_i$, and where the ith example X_i follows its disdtribution conditional on $A' = a'$ and $X_{\setminus i} = x_{\setminus i}$, as in Lemma 5.3.9. We then recognize that for each i, we have

$$\int D_{\rm kl}\left(A(a', x_1^n) \| \overline{A}(a', x_{\setminus i})\right) dP(x_i \mid a', x_{\setminus i}) \leq \int D_{\rm kl}\left(A(a', x_1^n) \| \widetilde{A}(a', x_{\setminus i})\right) dP(x_i \mid a', x_{\setminus i})$$

for *any* randomized function $\widetilde{A}$, as the marginal $\overline{A}$ in the lemma minimizes the average KL-divergence (recall Exercise 2.15). Now, sum over i and apply the definition of KL-stability as in Lemma 5.3.8. □

5.3.4 Error bounds for a simple noise addition scheme

Based on Proposition 5.3.10, to build an appropriately well-generalizing procedure we must build a mechanism for the interaction in Fig. 5.1 that maintains KL-stability. Using Example 5.3.6, this is not challenging for the class of bounded queries. Let $\Phi = \{\phi_t\}_{t\in\mathcal{T}}$ where $\phi_t : \mathcal{X} \to [-1, 1]$ be the collection of statistical queries taking values in $[-1, 1]$. Then based on Proposition 5.3.10 and Example 5.3.6, the following procedure is stable.

Input: Sample $X_1^n \in \mathcal{X}^n$ drawn i.i.d. P, collection $\{\phi_t\}_{t\in\mathcal{T}}$ of possible queries $\phi_t : \mathcal{X} \to [-1, 1]$
Repeat: for $k = 1, 2, \ldots$

i. Analyst chooses index $T_k \in \mathcal{T}$ and query $\phi := \phi_{T_k}$

ii. Mechanism draws independent $Z_k \sim \mathsf{N}(0, \sigma^2)$ and responds with answer

$$A_k := P_n\phi + Z_k = \frac{1}{n}\sum_{i=1}^n \phi(X_i) + Z_k.$$

Figure 5.2: Sequential Gaussian noise mechanism.

This procedure is evidently KL-stable, and based on Example 5.3.6 and Proposition 5.3.10, we have that

$$\frac{1}{n} I(X_1^n; T_1, \ldots, T_k, T_{k+1}) \le \frac{k}{2\sigma^2 n^2}$$

so long as the indices $T_i \in \mathcal{T}$ are chosen only as functions of $P_n\phi + Z_j$ for $j < i$, as the classical information processing inequality implies that

$$\frac{1}{n} I(X_1^n; T_1, \ldots, T_k, T_{k+1}) \le \frac{1}{n} I(X_1^n; A_1, \ldots, A_k)$$

because we have $X_1^n \to A_1 \to T_2$ and so on for the remaining indices. With this, we obtain the following theorem.

Theorem 5.3.11. *Let the indices T_i, $i = 1, \ldots, k+1$ be chosen in an arbitrary way using the procedure 5.2, and let $\sigma^2 > 0$. Then*

$$\mathbb{E}\left[\max_{j\le k}(A_j - P\phi_{T_j})^2\right] \le \frac{2ek}{\sigma^2 n^2} + \frac{10}{4n} + 4\sigma^2(\log k + 1).$$

By inspection, we can optimize over σ^2 by setting $\sigma^2 = \sqrt{k/(\log k + 1)}/n$, which yields the upper bound

$$\mathbb{E}\left[\max_{j\le k}(A_j - P\phi_{T_j})^2\right] \le \frac{10}{4n} + 10\frac{\sqrt{k(1 + \log k)}}{n}.$$

Comparing to Example 5.3.4, we see a substantial improvement. While we do not achieve accuracy scaling with $\log k$, as we would if the queried functionals ϕ_t were completely independent of the sample, we see that we achieve mean-squared error of order

$$\frac{\sqrt{k \log k}}{n}$$

for k adaptively chosen queries.

Proof To prove the result, we use a technique sometimes called the *monitor* technique. Roughly, the idea is that we can choose the index T_{k+1} in any way we desire as long as it is a function of the answers $A_1, \ldots, A_k$ and any other constants independent of the data. Thus, we may choose

$$T_{k+1} := T_{k^\star} \quad \text{where} \quad k^\star = \operatorname*{argmax}_{j \le k} \{|A_j - P\phi_{T_j}|\},$$

as this is a (downstream) function of the k different $\varepsilon = \frac{1}{2\sigma^2 n^2}$-KL-stable queries $T_1, \ldots, T_k$. As a consequence, we have from Corollary 5.3.3 (and the fact that the queries ϕ are 1-sub-Gaussian) that for $T = T_{k+1}$,

$$\mathbb{E}[(P_n\phi_T - P\phi_T)^2] \le \frac{2e}{n} I(X_1^n; T_{k+1}) + \frac{5}{4n} \le 2ek\varepsilon + \frac{5}{4n} = \frac{ek}{\sigma^2 n^2} + \frac{5}{4n}.$$

Now, we simply consider the independent noise addition, noting that $(a+b)^2 \le 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, so that

$$\begin{aligned} \mathbb{E}\left[\max_{j \le k} (A_j - P\phi_{T_j})^2\right] &\le 2\mathbb{E}[(P_n\phi_T - P\phi_T)^2] + 2\mathbb{E}\left[\max_{j \le k}\{Z_j^2\}\right] \\ &\le \frac{2ek}{\sigma^2 n^2} + \frac{10}{4n} + 4\sigma^2(\log k + 1), \end{aligned} \tag{5.3.3}$$

where inequality (5.3.3) is the desired result and follows by the following lemma.

Lemma 5.3.12. *Let W_j, $j = 1, \ldots, k$ be independent $\mathsf{N}(0,1)$. Then $\mathbb{E}[\max_j W_j^2] \le 2(\log k + 1)$.*

Proof We assume that $k \ge 3$, as the result is trivial otherwise. Using the tail bound for Gaussians (Mills's ratio for Gaussians, which is tighter than the standard sub-Gaussian bound) that $\mathbb{P}(W \ge t) \le \frac{1}{\sqrt{2\pi}t} e^{-t^2/2}$ for $t \ge 0$ and that $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \ge t) dt$ for a nonnegative random variable Z, we obtain that for any t_0,

$$\begin{aligned} \mathbb{E}[\max_j W_j^2] &= \int_0^\infty \mathbb{P}(\max_j W_j^2 \ge t) dt \le t_0 + \int_{t_0}^\infty \mathbb{P}(\max_j W_j^2 \ge t) dt \\ &\le t_0 + 2k \int_{t_0}^\infty \mathbb{P}(W_1 \ge \sqrt{t}) dt \le t_0 + \frac{2k}{\sqrt{2\pi}} \int_{t_0}^\infty e^{-t/2} dt = t_0 + \frac{4k}{\sqrt{2\pi}} e^{-t_0/2}. \end{aligned}$$

Setting $t_0 = 2\log(4k/\sqrt{2\pi})$ gives $\mathbb{E}[\max_j W_j^2] \le 2\log k + \log \frac{4}{\sqrt{2\pi}} + 1$. □

□

5.4 Bibliography and further reading

PAC-Bayes techniques originated with work of David McAllester [135, 136, 137], and we remark on his excellently readable tutorial [138]. The particular approaches we take to our proofs in Section 5.2 follow Catoni [44] and McAllester [137]. The PAC-Bayesian bounds we present, that simultaneously for *any* distribution π on $\mathcal{F}$, if $F \sim \pi$ then

$$\mathbb{E}[(P_n F - PF)^2 \mid X_1^n] \lesssim \frac{1}{n}\left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta}\right]$$

with probability at least $1 - \delta$ suggest that we can optimize them by choosing π carefully. For example, in the context of learning a statistical model parameterized by $\theta \in \Theta$ with losses $\ell(\theta; x, y)$, it is natural to attempt to find π minimizing

$$\mathbb{E}_\pi[P_n \ell(\theta; X, Y) \mid P_n] + C\sqrt{\frac{1}{n} D_{\rm kl}\left(\pi \| \pi_0\right)}$$

in π, where the expectation is taken over $\theta \sim \pi$. If this quantity has optimal value $\epsilon_n^\star$, then one is immediately guaranteed that for the population P, we have $\mathbb{E}_\pi[P\ell(\theta; X, y)] \le \epsilon_n^\star + C\sqrt{\log \frac{1}{\delta}}/\sqrt{n}$. Langford and Caruana [126] take this approach, and Dziugaite and Roy [79] use it to give (the first) non-trivial bounds for deep learning models.

The questions of interactive data analysis begin at least several decades ago, perhaps most profoundly highlighted positively by Tukey's *Exploratory Data Analysis* [168]. Problems of scientific replicability have, conversely, highlighted many of the challenges of reusing data or peeking, even innocently, at samples before performing statistical analyses [113, 86, 91]. Our approach to formalizing these ideas, and making rigorous limiting information leakage, draws from a more recent strain of work in the theoretical computer science literature, with major contributions from Dwork, Feldman, Hardt, Pitassi, Reingold, and Roth and Bassily, Nissim, Smith, Steinke, Stemmer, and Ullman [78, 76, 77, 20, 21]. Our particular treatment most closely follows Feldman and Steinke [82]. The problems these techniques target also arise frequently in high-dimensional statistics, where one often wishes to estimate uncertainty and perform inference *after* selecting a model. While we do not touch on these problems, a few references in this direction include [25, 166, 109].

5.5 Exercises

Exercise 5.1 (Duality in Donsker-Varadhan)**:** Here, we give a converse result to Theorem 5.1.1, showing that for any function $h : \mathcal{X} \to \mathbb{R}$,

$$\log \mathbb{E}_Q[e^{h(X)}] = \sup_P \left\{\mathbb{E}_P[h(X)] - D_{\rm kl}\left(P \| Q\right)\right\}, \tag{5.5.1}$$

where the supremum is taken over probability measures. If Q has a density, the supremum may be taken over probability measures having a density.

(a) Show the equality (5.5.1) in the case that $\mathcal{X}$ is discrete by directly computing the supremum. (That is, let $|\mathcal{X}| = k$, and identify probability measures P and Q with vectors $p, q \in \mathbb{R}_+^k$.)

(b) Let Q have density q. Assume that $\mathbb{E}_Q[e^{h(X)}] < \infty$ and let

$$Z_h(x) = \exp(h(x))/\mathbb{E}_Q[\exp(h(X))],$$

so $\mathbb{E}_Q[Z_h(X)] = 1$. Let P have density $p(x) = Z_h(x)q(x)$. Show that

$$\log \mathbb{E}_Q[e^{h(X)}] = \mathbb{E}_P[h(X)] - D_{\rm kl}\left(P \| Q\right).$$

Why does this imply equality (5.5.1) in this case?

(c) If $\mathbb{E}_Q[e^{h(X)}] = +\infty$, then monotone convergence implies that $\lim_{B \uparrow \infty} \mathbb{E}_Q[e^{\min\{B, h(X)\}}] = +\infty$. Conclude (5.5.1).

Exercise 5.2 (An alternative PAC-Bayes bound)**:** Let $f : \Theta \times \mathcal{X} \to \mathbb{R}$, and let π_0 be a density on $\theta \in \Theta$. Use the dual form (5.5.1) of the variational representation of the KL-divergence show that with probability at least $1 - \delta$ over the draw of $X_1^n \stackrel{\rm iid}{\sim} P$,

$$\int P_n f(\theta, X)\pi(\theta)d\theta \le \int \log \mathbb{E}_P \left[\exp(f(\theta, X))\right] \pi(\theta)d\theta + \frac{D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta}}{n}$$

simultaneously for all distributions π on Θ, where the expectation $\mathbb{E}_P$ is over $X \sim P$.

Exercise 5.3 (A mean estimator with sub-Gaussian concentration for a heavy-tailed distribution [45])**:** In this question, we use a PAC-Bayes bound to construct an estimator of the mean $\mathbb{E}[X]$ of a distribution with sub-Gaussian-like concentration that depends *only* on the second moments $\Sigma = \mathbb{E}[XX^\top]$ of the random vector X (not on any additional dimension-dependent quantitites) while only assuming that $\mathbb{E}[\|X\|^2] < \infty$. Let ψ be an odd function (i.e., $\psi(-t) = -\psi(t)$) satisfying

$$-\log(1 - t + t^2) \le \psi(t) \le \log(1 + t + t^2).$$

The function $\psi(t) = \min\{1, \max\{-1, t\}\}$ (the truncation of t to the range $[-1, 1]$) is such a function. Let π_θ be the normal distribution $\mathsf{N}(\theta, \sigma^2 I)$ and π_0 be $\mathsf{N}(0, \sigma^2 I)$.

(a) Let $\lambda > 0$. Use Exercise 5.2 to show that with probability at least $1 - \delta$, for all $\theta \in \mathbb{R}^d$

$$\frac{1}{\lambda} \int P_n \psi(\lambda \langle \theta', X\rangle)\pi_\theta(\theta')d\theta' \le \langle \theta, \mathbb{E}[X]\rangle + \lambda \left(\theta^\top \Sigma \theta + \sigma^2 \operatorname{tr}(\Sigma)\right) + \frac{\|\theta\|_2^2 / 2\sigma^2 + \log \frac{1}{\delta}}{n\lambda}.$$

(b) For $\lambda > 0$, define the "directional mean" estimator

$$E_n(\theta, \lambda) = \frac{1}{\lambda} \int P_n \psi(\lambda \langle \theta', X\rangle)\pi_\theta(\theta')d\theta'.$$

Give a choice of $\lambda > 0$ such that with probability $1 - \delta$,

$$\sup_{\theta \in \mathbb{S}^{d-1}} |E_n(\theta, \lambda) - \langle \theta, \mathbb{E}[X]\rangle| \le \frac{2}{\sqrt{n}} \sqrt{\left(\frac{1}{2\sigma^2} + \log \frac{1}{\delta}\right) \left(\|\Sigma\|_{\rm op} + \sigma^2 \operatorname{tr}(\Sigma)\right)},$$

where $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d \mid \|u\|_2 = 1\}$ is the unit sphere.

(c) Justify the following statement: choosing the vector $\widehat{\mu}_n$ minimizing

$$\sup_{\theta \in \mathbb{S}^{d-1}} |E_n(\theta, \lambda) - \langle \theta, \mu\rangle|$$

in μ guarantees that with probability at least $1 - \delta$,

$$\|\widehat{\mu}_n - \mathbb{E}[X]\|_2 \le \frac{4}{\sqrt{n}} \sqrt{\left(\frac{1}{2\sigma^2} + \log \frac{1}{\delta}\right) \left(\|\Sigma\|_{\rm op} + \sigma^2 \operatorname{tr}(\Sigma)\right)}.$$

(d) Give a choice of the prior/posterior variance σ^2 so that

$$\|\widehat{\mu}_n - \mathbb{E}[X]\|_2 \le \frac{4}{\sqrt{n}} \sqrt{\operatorname{tr}(\Sigma) + 2 \|\Sigma\|_{\rm op} \log \frac{1}{\delta}}$$

with probability at least $1 - \delta$.

Exercise 5.4 (Large-margin PAC-Bayes bounds for multiclass problems)**:** Consider the following multiclass prediction scenario. Data comes in pairs $(x, y) \in b\mathbb{B}_2^d \times [k]$ where $\mathbb{B}_2^d = \{v \in \mathbb{R}^d \mid \|v\|_2 \le 1\}$ denotes the ℓ_2-ball and $[k] = \{1, \ldots, k\}$. We make predictions using predictors $\theta_1, \ldots, \theta_k \in \mathbb{R}^d$, where the prediction of y on an example x is

$$\widehat{y}(x) := \operatorname*{argmax}_{i \le k} \langle \theta_i, x \rangle.$$

We suffer an error whenever $\widehat{y}(x) \neq y$, and the *margin* of our classifier on pair (x, y) is

$$\langle \theta_y, x \rangle - \max_{i \neq y} \langle \theta_i, x \rangle = \min_{i \neq y} \langle \theta_y - \theta_i, x \rangle.$$

If $\langle \theta_y, x \rangle > \langle \theta_i, x \rangle$ for all $i \neq y$, the margin is then positive (and the prediction is correct).

(a) Develop an analogue of the bounds in Section 5.2.2 in this k-class multiclass setting. To do so, you should (i) define the analogue of the margin-based loss ℓ_γ, (ii) show how Gaussian perturbations leave it similar, and (iii) prove an analogue of the bound in Section 5.2.2. You should assume one of the two conditions

$$\text{(C1)} \;\; \|\theta_i\|_2 \le r \text{ for all } i \qquad \text{(C2)} \;\; \sum_{i=1}^k \|\theta_i\|_2^2 \le kr^2$$

on your classification vectors θ_i. Specify which condition you choose.

(b) Describe a minimization procedure—just a few lines suffice—that uses convex optimization to find a (reasonably) large-margin multiclass classifier.

Exercise 5.5 (A variance-based information bound)**:** Let $\Phi = \{\phi_t\}_{t \in \mathcal{T}}$ be a collection of functions $\phi_t : \mathcal{X} \to \mathbb{R}$, where each ϕ_t satisfies the Bernstein condition (4.1.7) with parameters $\sigma^2(\phi_t)$ and b, that is, $|\mathbb{E}[(\phi_t(X) - P\phi_t(X))^k]| \le \frac{k!}{2}\sigma^2(\phi_t)b^{k-2}$ for all $k \ge 3$ and $\mathrm{Var}(\phi_t(X)) = \sigma^2(\phi_t)$. Let $T \in \mathcal{T}$ be any random variable, which may depend on an observed sample X_1^n. Show that for all $C > 0$ and $|\lambda| \le \frac{C}{2b}$, then

$$\left| \mathbb{E}\left[\frac{P_n\phi_T - P\phi_T}{\max\{C, \sigma(\phi_T)\}} \right] \right| \le \frac{1}{n|\lambda|} I(T; X_1^n) + |\lambda|.$$

Exercise 5.6 (An information bound on variance)**:** Let $\Phi = \{\phi_t\}_{t \in \mathcal{T}}$ be a collection of functions $\phi_t : \mathcal{X} \to \mathbb{R}$, where each $\phi_t : \mathcal{X} \to [-1, 1]$. Let $\sigma^2(\phi_t) = \mathrm{Var}(\phi_t(X))$. Let $s_n^2(\phi) = P_n\phi^2 - (P_n\phi)^2$ be the sample variance of ϕ. Show that for all $C > 0$ and $0 \le \lambda \le C/4$, then

$$\mathbb{E}\left[\frac{s_n^2(\phi_T)}{\max\{C, \sigma^2(\phi_T)\}} \right] \le \frac{1}{n\lambda} I(T; X_1^n) + 2.$$

The $\max\{C, \sigma^2(\phi_T)\}$ term is there to help avoid division by 0. *Hint:* If $0 \le x \le 1$, then $e^x \le 1 + 2x$, and if $X \in [0, 1]$, then $\mathbb{E}[e^X] \le 1 + 2\mathbb{E}[X] \le e^{2\mathbb{E}[X]}$. Use this to argue that $\mathbb{E}[e^{\lambda n P_n(\phi - P\phi)^2 / \max\{C, \sigma^2\}}] \le e^{2\lambda n}$ for any $\phi : \mathcal{X} \to [-1, 1]$ with $\mathrm{Var}(\phi) \le \sigma^2$, then apply the Donsker-Varadhan theorem.

Exercise 5.7: Consider the following scenario: let $\phi : \mathcal{X} \to [-1, 1]$ and let $\alpha > 0$, $\tau > 0$. Let $\mu = P_n\phi$ and $s^2 = P_n\phi^2 - \mu^2$. Define $\sigma^2 = \max\{\alpha s^2, \tau^2\}$, and assume that $\tau^2 \ge \frac{5\alpha}{n}$.

(a) Show that the mechanism with answer A_k defined by

$$A := P_n\phi + Z \quad \text{for} \quad Z \sim \mathsf{N}(0, \sigma^2)$$

is ε-KL-stable (Definition 5.1), where for a numerical constant $C < \infty$,

$$\varepsilon \le C \cdot \frac{s^2}{n^2\sigma^2} \cdot \left(1 + \frac{\alpha^2}{\sigma^2}\right).$$

(b) Show that if $\alpha^2 \le C'\tau^2$ for a numerical constant $C' < \infty$, then we can take $\varepsilon \le O(1)\frac{1}{n^2\alpha}$.

Hint: Use exercise 2.14, and consider the "alternative" mechanisms of sampling from

$$\mathsf{N}(\mu_{-i}, \sigma_{-i}^2) \;\text{ where }\; \sigma_{-i}^2 = \max\{\alpha s_{-i}^2, \tau^2\}$$

for

$$\mu_{-i} = \frac{1}{n-1}\sum_{j \ne i} \phi(X_j) \;\text{ and }\; s_{-i}^2 = \frac{1}{n-1}\sum_{j \ne i} \phi(X_j)^2 - \mu_{-i}^2.$$

Input: Sample $X_1^n \in \mathcal{X}^n$ drawn i.i.d. P, collection $\{\phi_t\}_{t\in\mathcal{T}}$ of possible queries $\phi_t : \mathcal{X} \to [-1, 1]$, parameters $\alpha > 0$ and $\tau > 0$

Repeat: for $k = 1, 2, \ldots$

i. Analyst chooses index $T_k \in \mathcal{T}$ and query $\phi := \phi_{T_k}$

ii. Set $s_k^2 := P_n\phi^2 - (P_n\phi)^2$ and $\sigma_k^2 := \max\{\alpha s_k^2, \tau^2\}$

iii. Mechanism draws independent $Z_k \sim \mathsf{N}(0, \sigma_k^2)$ and responds with answer

$$A_k := P_n\phi + Z_k = \frac{1}{n}\sum_{i=1}^n \phi(X_i) + Z_k.$$

Figure 5.3: Sequential Gaussian noise mechanism with variance sensitivity.

Exercise 5.8 (A general variance-dependent bound on interactive queries)**:** Consider the algorithm in Fig. 5.3. Let $\sigma^2(\phi_t) = \mathrm{Var}(\phi_t(X))$ be the variance of ϕ_t.

(a) Show that for $b > 0$ and for all $0 \le \lambda \le \frac{b}{2}$,

$$\mathbb{E}\left[\max_{j\le k} \frac{|A_j - P\phi_{T_j}|}{\max\{b, \sigma(\phi_{T_j})\}}\right] \le \frac{1}{n\lambda} I(X_1^n; T_1^k) + \lambda + \sqrt{2\log(ke)}\sqrt{\frac{4\alpha}{nb^2} I(X_1^n; T_1^k) + 2\alpha + \frac{\tau^2}{b^2}}.$$

(If you do not have quite the right constants, that's fine.)

(b) Using the result of Question 5.7, show that with appropriate choices for the parameters $\alpha, b, \tau^2, \lambda$ that for a numerical constant $C < \infty$

$$\mathbb{E}\left[\max_{j\le k} \frac{|A_j - P\phi_{T_j}|}{\max\{(k\log k)^{1/4}/\sqrt{n}, \sigma(\phi_{T_j})\}}\right] \le C\frac{(k\log k)^{1/4}}{\sqrt{n}}.$$

You may assume that k, n are large if necessary.

(c) Interpret the result from part (b). How does this improve over Theorem 5.3.11?

Chapter 6

Advanced techniques in concentration inequalities

6.1 Entropy and concentration inequalities

In the previous sections, we saw how moment generating functions and related techniques could be used to give bounds on the probability of deviation for fairly simple quantities, such as sums of random variables. In many situations, however, it is desirable to give guarantees for more complex functions. As one example, suppose that we draw a matrix $X \in \mathbb{R}^{m \times n}$, where the entries of X are bounded independent random variables. The operator norm of X, $\|X\| := \sup_{u,v}\{u^\top X v : \|u\|_2 = \|v\|_2 = 1\}$, is one measure of the size of X. We would like to give upper bounds on the probability that $\|X\| \geq \mathbb{E}[\|X\|] + t$ for $t \geq 0$, which the tools of the preceding sections do not address well because of the complicated dependencies on $\|X\|$.

In this section, we will develop techniques to give control over such complex functions. In particular, throughout we let $Z = f(X_1, \ldots, X_n)$ be some function of a sample of independent random variables X_i; we would like to know if Z is concentrated around its mean. We will use deep connections between information theoretic quantities and deviation probabilities to investigate these connections.

First, we give a definition.

Definition 6.1. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. The* ϕ-entropy *of a random variable X is*

$$\mathbb{H}_\phi(X) := \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]), \tag{6.1.1}$$

assuming the relevant expectations exist.

A first example of the ϕ-entropy is the variance:

Example 6.1.1 (Variance as ϕ-entropy)**:** Let $\phi(t) = t^2$. Then $\mathbb{H}_\phi(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathrm{Var}(X)$. $\diamond$

This example is suggestive of the fact that ϕ-entropies may help us to control deviations of random variables from their means. More generally, we have by Jensen's inequality that $\mathbb{H}_\phi(X) \geq 0$ for any convex ϕ; moreover, if ϕ is strictly convex and X is non-constant, then $\mathbb{H}_\phi(X) > 0$. The rough intuition we consider throughout this section is as follows: if a random variable X is tightly concentrated around its mean, then we should have $X \approx \mathbb{E}[X]$ "most" of the time, and so $\mathbb{H}_\phi(X)$ should be small. The goal of this section is to make this claim rigorous.

6.1.1 The Herbst argument

Perhaps unsurprisingly given the focus of these lecture notes, we focus on a specific ϕ, using $\phi(t) = t \log t$, which gives the entropy on which we focus:

$$\mathbb{H}(Z) := \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z], \tag{6.1.2}$$

defined whenever $Z \geq 0$ with probability 1. As our particular focus throughout this chapter, we consider the moment generating function and associated transformation $X \mapsto e^{\lambda X}$. If we know the moment generating function $\varphi_X(\lambda) := \mathbb{E}[e^{\lambda X}]$, then $\varphi_X'(\lambda) = \mathbb{E}[Xe^{\lambda X}]$, and so

$$\mathbb{H}(e^{\lambda X}) = \lambda \varphi_X'(\lambda) - \varphi_X(\lambda) \log \varphi_X(\lambda).$$

This suggests—in a somewhat roundabout way we make precise—that control of the entropy $\mathbb{H}(e^{\lambda X})$ should be sufficient for controlling the moment generating function of X.

The Herbst argument makes this rigorous.

Proposition 6.1.2. *Let X be a random variable and assume that there exists a constant $\sigma^2 < \infty$ such that*

$$\mathbb{H}(e^{\lambda X}) \leq \frac{\lambda^2 \sigma^2}{2} \varphi_X(\lambda). \tag{6.1.3}$$

for all $\lambda \in \mathbb{R}$ (respectively, $\lambda \in \mathbb{R}_+$) where $\varphi_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ denotes the moment generating function of X. Then

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$$

for all $\lambda \in \mathbb{R}$ (respectively, $\lambda \in \mathbb{R}_+$).

Proof Let $\varphi = \varphi_X$ for shorthand. The proof procedes by an integration argument, where we show that $\log \varphi(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$. First, note that

$$\varphi'(\lambda) = \mathbb{E}[Xe^{\lambda X}],$$

so that inequality (6.1.3) is equivalent to

$$\lambda \varphi'(\lambda) - \varphi(\lambda) \log \varphi(\lambda) = \mathbb{H}(e^{\lambda X}) \leq \frac{\lambda^2 \sigma^2}{2} \varphi(\lambda),$$

and dividing both sides by $\lambda^2 \varphi(\lambda)$ yields the equivalent statement

$$\frac{\varphi'(\lambda)}{\lambda \varphi(\lambda)} - \frac{1}{\lambda^2} \log \varphi(\lambda) \leq \frac{\sigma^2}{2}.$$

But by inspection, we have

$$\frac{\partial}{\partial \lambda} \frac{1}{\lambda} \log \varphi(\lambda) = \frac{\varphi'(\lambda)}{\lambda \varphi(\lambda)} - \frac{1}{\lambda^2} \log \varphi(\lambda).$$

Moreover, we have that

$$\lim_{\lambda \to 0} \frac{\log \varphi(\lambda)}{\lambda} = \lim_{\lambda \to 0} \frac{\log \varphi(\lambda) - \log \varphi(0)}{\lambda} = \frac{\varphi'(0)}{\varphi(0)} = \mathbb{E}[X].$$

Integrating from 0 to any λ_0, we thus obtain

$$\frac{1}{\lambda_0}\log\varphi(\lambda_0) - \mathbb{E}[X] = \int_0^{\lambda_0}\left[\frac{\partial}{\partial\lambda}\frac{1}{\lambda}\log\varphi(\lambda)\right]d\lambda \le \int_0^{\lambda_0}\frac{\sigma^2}{2}d\lambda = \frac{\sigma^2\lambda_0}{2}.$$

Multiplying each side by λ_0 gives

$$\log\mathbb{E}[e^{\lambda_0(X-\mathbb{E}[X])}] = \log\mathbb{E}[e^{\lambda_0 X}] - \lambda_0\mathbb{E}[X] \le \frac{\sigma^2\lambda_0^2}{2},$$

as desired. □

It is possible to give a similar argument for sub-exponential random variables, which allows us to derive Bernstein-type bounds, of the form of Corollary 4.1.18, but using the entropy method. In particular, in the exercises, we show the following result.

Proposition 6.1.3. *Assume that there exist positive constants b and σ such that*

$$\mathbb{H}(e^{\lambda X}) \le \lambda^2\left[b\varphi_X'(\lambda) + \varphi_X(\lambda)(\sigma^2 - b\mathbb{E}[X])\right] \tag{6.1.4a}$$

for all $\lambda \in [0, 1/b)$. Then X satisfies the sub-exponential bound

$$\log\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le \frac{\sigma^2\lambda^2}{[1-b\lambda]_+} \tag{6.1.4b}$$

for all $\lambda \ge 0$.

An immediate consequence of this proposition is that any random variable satisfying the entropy bound (6.1.4a) is $(2\sigma^2, 2b)$-sub-exponential. As another immediate consequence, we obtain the concentration guarantee

$$\mathbb{P}(X \ge \mathbb{E}[X] + t) \le \exp\left(-\frac{1}{4}\min\left\{\frac{t^2}{\sigma^2}, \frac{t}{b}\right\}\right)$$

as in Proposition 4.1.16.

6.1.2 Tensorizing the entropy

A benefit of the moment generating function approach we took in the prequel is the excellent behavior of the moment generating function for sums. In particular, the fact that $\varphi_{X_1+\cdots+X_n}(\lambda) = \prod_{i=1}^n \varphi_{X_i}(\lambda)$ allowed us to derive sharper concentration inequalities, and we were only required to work with *marginal* distributions of the X_i, computing only the moment generating functions of individual random variables rather than characteristics of the entire sum. One advantage of the entropy-based tools we develop is that they allow similar tensorization—based on the chain rule identities of Chapter 2 for entropy, mutual information, and KL-divergence—for substantially more complex functions. Our approach here mirrors that of Boucheron, Lugosi, and Massart [34].

With that in mind, we now present a series of inequalities that will allow us to take this approach. For shorthand throughout this section, we let

$$X_{\setminus i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

be the collection of all variables except X_i. Our first result is a consequence of the chain rule for entropy and is known as Han's inequality.

Proposition 6.1.4 (Han's inequality)**.** *Let $X_1, \ldots, X_n$ be discrete random variables. Then*

$$H(X_1^n) \le \frac{1}{n-1} \sum_{i=1}^n H(X_{\setminus i}).$$

Proof The proof is a consequence of the chain rule for entropy and that conditioning reduces entropy. We have

$$H(X_1^n) = H(X_i \mid X_{\setminus i}) + H(X_{\setminus i}) \le H(X_i \mid X_1^{i-1}) + H(X_{\setminus i}).$$

Writing this inequality for each $i = 1, \ldots, n$, we obtain

$$nH(X_1^n) \le \sum_{i=1}^n H(X_{\setminus i}) + \sum_{i=1}^n H(X_i \mid X_1^{i-1}) = \sum_{i=1}^n H(X_{\setminus i}) + H(X_1^n),$$

and subtracting $H(X_1^n)$ from both sides gives the result. □

We also require a divergence version of Han's inequality, which will allow us to relate the entropy $\mathbb{H}$ of a random variable to divergences and other information-theoretic quantities. Let $\mathcal{X}$ be an arbitrary space, and let Q be a distribution over $\mathcal{X}^n$ and $P = P_1 \times \cdots \times P_n$ be a product distribution on the same space. For $A \subset \mathcal{X}^{n-1}$, define the marginal densities

$$Q^{(i)}(A) := Q(X_{\setminus i} \in A) \quad \text{and} \quad P^{(i)}(A) = P(X_{\setminus i} \in A).$$

We then obtain the tensorization-type Han's inequality for relative entropies.

Proposition 6.1.5. *With the above definitions,*

$$D_{\rm kl}\left(Q \| P\right) \le \sum_{i=1}^n \left[D_{\rm kl}\left(Q \| P\right) - D_{\rm kl}\left(Q^{(i)} \| P^{(i)}\right)\right].$$

Proof We have seen earlier in the notes (recall the definition (2.2.1) of the KL divergence as a supremum over all quantizers and the surrounding discussion) that it is no loss of generality to assume that $\mathcal{X}$ is discrete. Thus, noting that the probability mass functions

$$q^{(i)}(x_{\setminus i}) = \sum_x q(x_1^{i-1}, x, x_{i+1}^n) \;\; \text{and} \;\; p^{(i)}(x_{\setminus i}) = \prod_{j \ne i} p_j(x_j),$$

we have that Han's inequality (Proposition 6.1.4) is equivalent to

$$(n-1) \sum_{x_1^n} q(x_1^n) \log q(x_1^n) \ge \sum_{i=1}^n \sum_{x_{\setminus i}} q^{(i)}(x_{\setminus i}) \log q^{(i)}(x_{\setminus i}).$$

Now, by subtracting $q(x_1^n) \log p(x_1^n)$ from both sides of the preceding display, we obtain

$$\begin{aligned}(n-1) D_{\rm kl}\left(Q \| P\right) &= (n-1) \sum_{x_1^n} q(x_1^n) \log q(x_1^n) - (n-1) \sum_{x_1^n} q(x_1^n) \log p(x_1^n) \\ &\ge \sum_{i=1}^n \sum_{x_{\setminus i}} q^{(i)}(x_{\setminus i}) \log q^{(i)}(x_{\setminus i}) - (n-1) \sum_{x_1^n} q(x_1^n) \log p(x_1^n).\end{aligned}$$

We expand the final term. Indeed, by the product nature of the distributions p, we have

$$(n-1)\sum_{x_1^n} q(x_1^n)\log p(x_1^n) = (n-1)\sum_{x_1^n} q(x_1^n)\sum_{i=1}^n \log p_i(x_i)$$
$$= \sum_{i=1}^n \sum_{x_1^n} q(x_1^n) \underbrace{\sum_{j\neq i} \log p_i(x_i)}_{=\log p^{(i)}(x_{\setminus i})} = \sum_{i=1}^n \sum_{x_{\setminus i}} q^{(i)}(x_{\setminus i}) \log p^{(i)}(x_{\setminus i}).$$

Noting that

$$\sum_{x_{\setminus i}} q^{(i)}(x_{\setminus i})\log q^{(i)}(x_{\setminus i}) - \sum_{x_{\setminus i}} q^{(i)}(x_{\setminus i})\log p^{(i)}(x_{\setminus i}) = D_{\rm kl}\left(Q^{(i)} \| P^{(i)}\right)$$

and rearranging gives the desired result. □

Finally, we will prove the main result of this subsection: a tensorization identity for the entropy $\mathbb{H}(Y)$ for an arbitrary random variable Y that is a function of n independent random variables. For this result, we use a technique known as *tilting*, in combination with the two variants of Han's inequality we have shown, to obtain the result. The tilting technique is one used to transform problems of random variables into one of distributions, allowing us to bring the tools of information and entropy to bear more directly. This technique is a common one, and used frequently in large deviation theory, statistics, for heavy-tailed data, amont other areas. More concretely, let $Y = f(X_1, \ldots, X_n)$ for some non-negative function f. Then we may always define a tilted density

$$q(x_1, \ldots, x_n) := \frac{f(x_1, \ldots, x_n)p(x_1, \ldots, x_n)}{\mathbb{E}_P[f(X_1, \ldots, X_n)]} \tag{6.1.5}$$

which, by inspection, satisfies $\int q(x_1^n) = 1$ and $q \geq 0$. In our context, if $f \approx$ constant under the distribution P, then we should have $f(x_1^n)p(x_1^n) \approx cp(x_1^n)$ and so $D_{\rm kl}\left(Q\|P\right)$ should be small; we can make this rigorous via the following tensorization theorem.

Theorem 6.1.6. *Let $X_1, \ldots, X_n$ be independent random variables and $Y = f(X_1^n)$, where f is a non-negative function. Define $\mathbb{H}(Y \mid X_{\setminus i}) = \mathbb{E}[Y \log Y \mid X_{\setminus i}]$. Then*

$$\mathbb{H}(Y) \leq \mathbb{E}\left[\sum_{i=1}^n \mathbb{H}(Y \mid X_{\setminus i})\right]. \tag{6.1.6}$$

Proof Inequality (6.1.6) holds for Y if and only if holds identically for cY for any $c > 0$, so we assume without loss of generality that $\mathbb{E}_P[Y] = 1$. We thus obtain that $\mathbb{H}(Y) = \mathbb{E}[Y \log Y] = \mathbb{E}[\phi(Y)]$, where assign $\phi(t) = t\log t$. Let P have density p with respect to a base measure μ. Then by defining the tilted distribution (density) $q(x_1^n) = f(x_1^n)p(x_1^n)$, we have $Q(\mathcal{X}^n) = 1$, and moreover, we have

$$D_{\rm kl}\left(Q\|P\right) = \int q(x_1^n)\log\frac{q(x_1^n)}{p(x_1^n)}d\mu(x_1^n) = \int f(x_1^n)p(x_1^n)\log f(x_1^n)d\mu(x_1^n) = \mathbb{E}_P[Y\log Y] = \mathbb{H}(Y).$$

Similarly, if $\phi(t) = t \log t$, then

$$\begin{aligned}
&D_{\rm kl}\left(Q^{(i)} \| P^{(i)}\right) \\
&= \int_{\mathcal{X}^{n-1}} \left(\int f(x_1^{i-1}, x, x_{i+1}^n) p_i(x) d\mu(x) \right) \log \frac{p^{(i)}(x_{\setminus i}) \int f(x_1^{i-1}, x, x_{i+1}^n) p_i(x) d\mu(x)}{p^{(i)}(x_{\setminus i})} p^{(i)}(x_{\setminus i}) d\mu(x_{\setminus i}) \\
&= \int_{\mathcal{X}^{n-1}} \mathbb{E}[Y \mid x_{\setminus i}] \log \mathbb{E}[Y \mid x_{\setminus i}] p^{(i)}(x_{\setminus i}) d\mu(x_{\setminus i}) \\
&= \mathbb{E}[\phi(\mathbb{E}[Y \mid X_{\setminus i}])].
\end{aligned}$$

The tower property of expectations then yields that

$$\mathbb{E}[\phi(Y)] - \mathbb{E}[\phi(\mathbb{E}[Y \mid X_{\setminus i}])] = \mathbb{E}[\mathbb{E}[\phi(Y) \mid X_{\setminus i}] - \phi(\mathbb{E}[Y \mid X_{\setminus i}])] = \mathbb{E}[\mathbb{H}(Y \mid X_{\setminus i})].$$

Using Han's inequality for relative entropies (Proposition 6.1.4) then immediately gives

$$\mathbb{H}(Y) = D_{\rm kl}\left(Q \| P\right) \le \sum_{i=1}^n \left[D_{\rm kl}\left(Q \| P\right) - D_{\rm kl}\left(Q^{(i)} \| P^{(i)}\right) \right] = \sum_{i=1}^n \mathbb{E}[\mathbb{H}(Y \mid X_{\setminus i})],$$

which is our desired result. □

Theorem 6.1.6 shows that if we can show that individually the conditional entropies $\mathbb{H}(Y \mid X_{\setminus i})$ are not too large, then the Herbst argument (Proposition 6.1.2 or its variant Proposition 6.1.3) allows us to provide strong concentration inequalities for general random variables Y.

Examples and consequences

We now show how to use some of the preceding results to derive strong concentration inequalities, showing as well how we may give convergence guarantees for a variety of procedures using these techniques.

We begin with our most straightforward example, which is the bounded differences inequality. In particular, we consider an arbitrary function f of n independent random variables, and we assume that for all $x_{1:n} = (x_1, \ldots, x_n)$, we have the bounded differences condition:

$$\sup_{x \in \mathcal{X}, x' \in \mathcal{X}} \left| f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x', x_{i+1}, \ldots, x_n) \right| \le c_i \quad \text{for all } x_{\setminus i}. \tag{6.1.7}$$

Then we have the following result.

Proposition 6.1.7 (Bounded differences)**.** *Assume that f satisfies the bounded differences condition (6.1.7), where $\frac{1}{4}\sum_{i=1}^n c_i^2 \le \sigma^2$. Let X_i be independent. Then $Y = f(X_1, \ldots, X_n)$ is σ^2-sub-Gaussian.*

Proof We use a similar integration argument to the Herbst argument of Proposition 6.1.2, and we apply the tensorization inequality (6.1.6). First, let U be an arbitrary random variable taking values in $[a, b]$. We claim that if $\varphi_U(\lambda) = \mathbb{E}[e^{\lambda U}]$ and $\psi(\lambda) = \log \varphi_U(\lambda)$ is its cumulant generating function, then

$$\frac{\mathbb{H}(e^{\lambda U})}{\mathbb{E}[e^{\lambda U}]} \le \frac{\lambda^2 (b-a)^2}{8}. \tag{6.1.8}$$

To see this, note that

$$\frac{\partial}{\partial\lambda}[\lambda\psi'(\lambda) - \psi(\lambda)] = \psi''(\lambda), \quad \text{so} \quad \lambda\psi'(\lambda) - \psi(\lambda) = \int_0^\lambda t\psi''(t)dt \le \frac{\lambda^2(b-a)^2}{8},$$

where we have used the homework exercise **XXXX** (recall Hoeffding's Lemma, Example 4.1.6), to argue that $\psi''(t) \le \frac{(b-a)^2}{4}$ for all t. Recalling that

$$\mathbb{H}(e^{\lambda U}) = \lambda\varphi_U'(\lambda) - \varphi_U(\lambda)\psi(\lambda) = \left[\lambda\psi'(\lambda) - \psi(\lambda)\right]\varphi_U(\lambda)$$

gives inequality (6.1.8).

Now we apply the tensorization identity. Let $Z = e^{\lambda Y}$. Then we have

$$\mathbb{H}(Z) \le \mathbb{E}\left[\sum_{i=1}^n \mathbb{H}(Z \mid X_{\setminus i})\right] \le \mathbb{E}\left[\sum_{i=1}^n \frac{c_i^2\lambda^2}{8}\mathbb{E}[e^{\lambda Z} \mid X_{\setminus i}]\right] = \sum_{i=1}^n \frac{c_i^2\lambda^2}{8}\mathbb{E}[e^{\lambda Z}].$$

Applying the Herbst argument gives the final result. □

As an immediate consequence of this inequality, we obtain the following dimension independent concentration inequality.

Example 6.1.8: Let $X_1, \ldots, X_n$ be independent vectors in $\mathbb{R}^d$, where d is arbitrary, and assume that $\|X_i\|_2 \le c_i$ with probability 1. (This could be taken to be a general Hilbert space with no loss of generality.) We claim that if we define

$$\sigma^2 := \sum_{i=1}^n c_i^2, \quad \text{then} \quad \mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\|_2 \ge t\right) \le \exp\left(-2\frac{[t - \sqrt{\sigma}]_+^2}{\sigma^2}\right).$$

Indeed, we have that $Y = \|\sum_{i=1}^n X_i\|_2$ satisfies the bounded differences inequality with parameters c_i, and so

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\|_2 \ge t\right) &= \mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\|_2 - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\|_2 \ge t - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\|_2\right) \\ &\le \exp\left(-2\frac{[t - \mathbb{E}\|\sum_{i=1}^n X_i\|_2]_+^2}{\sum_{i=1}^n c_i^2}\right).\end{aligned}$$

Noting that $\mathbb{E}[\|\sum_{i=1}^n X_i\|_2] \le \sqrt{\mathbb{E}[\|\sum_{i=1}^n X_i\|_2^2]} = \sqrt{\sum_{i=1}^n \mathbb{E}[\|X_i\|_2^2]}$ gives the result. ◇

6.1.3 Concentration of convex functions

We provide a second theorem on the concentration properties of a family of functions that are quite useful, for which other concentration techniques do not appear to give results. In particular, we say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *separately convex* if for each $i \in \{1, \ldots, n\}$ and all $x_{\setminus i} \in \mathbb{R}^{n-1}$ (or the domain of f), we have that

$$x \mapsto f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$$

is convex. We also recall that a function is L-Lipschitz if $|f(x) - f(y)| \le \|x - y\|_2$ for all $x, y \in \mathbb{R}^n$; any L-Lipschitz function is almost everywhere differentiable, and is L-Lipschitz if and only if $\|\nabla f(x)\|_2 \le L$ for (almost) all x. With these preliminaries in place, we have the following result.

Theorem 6.1.9. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [a, b]$ for all i. Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is separately convex and L-Lipschitz with respect to the $\|\cdot\|_2$ norm. Then*

$$\mathbb{E}[\exp(\lambda(f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]))] \le \exp\left(\lambda^2 (b-a)^2 L^2\right) \quad \text{for all } \lambda \ge 0.$$

We defer the proof of the theorem temporarily, giving two example applications. The first is to the matrix concentration problem that motivates the beginning of this section.

Example 6.1.10: Let $X \in \mathbb{R}^{m \times n}$ be a matrix with independent entries, where $X_{ij} \in [-1, 1]$ for all i, j, and let $|||\cdot|||$ denote the operator norm on matrices, that is, $|||A||| = \sup_{u,v}\{u^\top A v : \|u\|_2 \le 1, \|v\|_2 \le 1\}$. Then Theorem 6.1.9 implies

$$\mathbb{P}(|||X||| \ge \mathbb{E}[|||X|||] + t) \le \exp\left(-\frac{t^2}{16}\right)$$

for all $t \ge 0$. Indeed, we first observe that

$$|\, |||X||| - |||Y||| \,| \le |||X - Y||| \le \|X - Y\|_{\rm Fr},$$

where $\|\cdot\|_{\rm Fr}$ denotes the Frobenius norm of a matrix. Thus the matrix operator norm is 1-Lipschitz. Therefore, we have by Theorem 6.1.9 and the Chernoff bound technique that

$$\mathbb{P}(|||X||| \ge \mathbb{E}[|||X|||] + t) \le \exp(4\lambda^2 - \lambda t)$$

for all $\lambda \ge 0$. Taking $\lambda = t/8$ gives the desired result. $\diamond$

As a second example, we consider *Rademacher complexity*. These types of results are important for giving generalization bounds in a variety of statistical algorithms, and form the basis of a variety of concentration and convergence results. We defer further motivation of these ideas to subsequent chapters, just mentioning here that we can provide strong concentration guarantees for Rademacher complexity or Rademacher chaos.

Example 6.1.11: Let $\mathcal{A} \subset \mathbb{R}^n$ be any collection of vectors. The the *Rademacher complexity* of the class $\mathcal{A}$ is

$$R_n(\mathcal{A}) := \mathbb{E}\left[\sup_{a \in \mathcal{A}} \sum_{i=1}^n a_i \varepsilon_i\right], \tag{6.1.9}$$

where ε_i are i.i.d. Rademacher (sign) variables. Let $\widehat{R}_n(\mathcal{A}) = \sup_{a \in \mathcal{A}} \sum_{i=1}^n a_i \varepsilon_i$ denote the empirical version of this quantity. We claim that

$$\mathbb{P}(\widehat{R}_n(\mathcal{A}) \ge R_n(\mathcal{A}) + t) \le \exp\left(-\frac{t^2}{16 \operatorname{diam}(\mathcal{A})^2}\right),$$

where $\operatorname{diam}(\mathcal{A}) := \sup_{a \in \mathcal{A}} \|a\|_2$. Indeed, we have that $\varepsilon \mapsto \sup_{a \in \mathcal{A}} a^\top \varepsilon$ is a convex function, as it is the maximum of a family of linear functions. Moreover, it is Lipschitz, with Lipschitz constant bounded by $\sup_{a \in \mathcal{A}} \|a\|_2$. Applying Theorem 6.1.9 as in Example 6.1.10 gives the result. $\diamond$

Proof of Theorem 6.1.9 The proof relies on our earlier tensorization identity and a symmetrization lemma.

Lemma 6.1.12. *Let $X, Y \stackrel{\text{iid}}{\sim} P$ be independent. Then for any function $g : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{H}(e^{\lambda g(X)}) \le \lambda^2 \mathbb{E}[(g(X) - g(Y))^2 e^{\lambda g(X)} \mathbf{1}\{g(X) \ge g(Y)\}] \quad \text{for } \lambda \ge 0.$$

Moreover, if g is convex, then

$$\mathbb{H}(e^{\lambda g(X)}) \le \lambda^2 \mathbb{E}[(X - Y)^2 (g'(X))^2 e^{\lambda g(X)}] \quad \text{for } \lambda \ge 0.$$

Proof For the first result, we use the convexity of the exponential in an essential way. In particular, we have

$$\begin{aligned}\mathbb{H}(e^{\lambda g(X)}) &= \mathbb{E}[\lambda g(X) e^{\lambda g(X)}] - \mathbb{E}[e^{\lambda g(X)}] \log \mathbb{E}[e^{\lambda g(Y)}] \\ &\le \mathbb{E}[\lambda g(X) e^{\lambda g(X)}] - \mathbb{E}[e^{\lambda g(X)} \lambda g(Y)],\end{aligned}$$

because log is concave and $e^x \ge 0$. Using symmetry, that is, that $g(X) - g(Y)$ has the same distribution as $g(Y) - g(X)$, we then find

$$\mathbb{H}(e^{\lambda g(X)}) \le \frac{1}{2} \mathbb{E}[\lambda (g(X) - g(Y))(e^{\lambda g(X)} - e^{\lambda g(Y)})] = \mathbb{E}[\lambda (g(X) - g(Y))(e^{\lambda g(X)} - e^{\lambda g(Y)}) \mathbf{1}\{g(X) \ge g(Y)\}].$$

Now we use the classical first order convexity inequality—that a convex function f satisfies $f(t) \ge f(s) + f'(s)(t - s)$ for all t and s, Theorem B.3.3 in the appendices—which gives that $e^t \ge e^s + e^s(t - s)$ for all s and t. Rewriting, we have $e^s - e^t \le e^s(s - t)$, and whenever $s \ge t$, we have $(s - t)(e^s - e^t) \le e^s (s - t)^2$. Replacing s and t with $\lambda g(X)$ and $\lambda g(Y)$, respectively, we obtain

$$\lambda (g(X) - g(Y))(e^{\lambda g(X)} - e^{\lambda g(Y)}) \mathbf{1}\{g(X) \ge g(Y)\} \le \lambda^2 (g(X) - g(Y))^2 e^{\lambda g(X)} \mathbf{1}\{g(X) \ge g(Y)\}.$$

This gives the first inequality of the lemma.

To obtain the second inequality, note that if g is convex, then whenever $g(x) - g(y) \ge 0$, we have $g(y) \ge g(x) + g'(x)(y - x)$, or $g'(x)(x - y) \ge g(x) - g(y) \ge 0$. In particular,

$$(g(X) - g(Y))^2 \mathbf{1}\{g(X) \ge g(Y)\} \le (g'(X)(X - Y))^2,$$

which gives the second result. □

Returning to the main thread of the proof, we note that the separate convexity of f and the tensorization identity of Theorem 6.1.6 imply

$$\mathbb{H}(e^{\lambda f(X_{1:n})}) \le \mathbb{E}\left[\sum_{i=1}^n \mathbb{H}(e^{\lambda f(X_{1:n})} \mid X_{\setminus i})\right] \le \mathbb{E}\left[\sum_{i=1}^n \lambda^2 \mathbb{E}\left[(X_i - Y_i)^2 \left(\frac{\partial}{\partial x_i} f(X_{1:n})\right)^2 e^{\lambda f(X_{1:n})} \mid X_{\setminus i}\right]\right],$$

where Y_i are independent copies of the X_i. Now, we use that $(X_i - Y_i)^2 \le (b - a)^2$ and the definition of the partial derivative to obtain

$$\mathbb{H}(e^{\lambda f(X_{1:n})}) \le \lambda^2 (b - a)^2 \mathbb{E}[\|\nabla f(X_{1:n})\|_2^2 e^{\lambda f(X_{1:n}))}].$$

Noting that $\|\nabla f(X)\|_2^2 \le L^2$, and applying the Herbst argument, gives the result. □

Exercise 6.1 (A discrete isoperimetric inequality)**:** Let $A \subset \mathbb{Z}^d$ be a finite subset of the d-dimensional integers. Let the projection mapping $\pi_j : \mathbb{Z}^d \to \mathbb{Z}^{d-1}$ be defined by

$$\pi_j(z_1, \ldots, z_d) = (z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_d)$$

so that we "project out" the jth coordinate, and define the projected sets.

$$\begin{aligned} A_j = \pi_j(A) &= \{\pi_j(z) : z \in A\} \\ &= \left\{z \in \mathbb{Z}^{d-1} : \text{ there exists } z_\star \in \mathbb{Z} \text{ such that } (z_1, z_2, \ldots, z_{j-1}, z_\star, z_j, \ldots, z_{d-1}) \in A\right\}. \end{aligned}$$

Prove the Loomis-Whitney inequality, that is, that

$$\mathrm{card}(A) \le \left(\prod_{j=1}^{d} \mathrm{card}(A_j)\right)^{\frac{1}{d-1}}.$$

Chapter 7

Privacy and disclosure limitation

In this chapter, we continue to build on our ideas on stability in different scenarios, ranging from model fitting and concentration to interactive data analyses. Here, we show how stability ideas allow us to provide a new type of protection: the privacy of participants in studies. Until the mid-2000s, the major challenge in this direction had been a satisfactory definition of privacy, because collection of side information often results in unforeseen compromises of private information. The introduction of *differential privacy*—a type of stability in likelihood ratios for data releases from differing samples—alleviated these challenges, providing a firm foundation on which to build private estimators and other methodology. (Though it is possible to trace some of the definitions and major insights in privacy back at least to survey sampling literature in the 1960s.) Consequently, in this chapter we focus on privacy notions based on differential privacy and its cousins, developing the information-theoretic stability ideas helpful to understand the protections it is possible to provide.

7.1 Disclosure limitation, privacy, and definitions

We begin this chapter with a few cautionary tales and examples, which motivate the coming definitions of privacy that we consider. A natural belief might be that, given only certain summary statistics of a large dataset, individuals in the data are protected. Yet this appears, by and large, to be false. As an example, in 2008 Nils Homer and colleagues [107] showed that even releasing aggregated genetic frequency statistics (e.g., frequency of single nucleotide polymorphisms (SNP) in microarrays) can allow resolution of individuals within a database. Consequently, the US National Institutes of Health (NIH), the Wellcome Trust, and the Broad Institute removed genetic summaries from public access (along with imposing stricter requirements for private access) [161, 52].

Another hypothetical example may elucidate some of the additional challenges. Suppose that I release a dataset that consists of the frequent times that posts are made worldwide that denigrate government policies, but I am sure to remove all information such as IP addresses, usernames, or other metadata excepting the time of the post. This might seem *a priori* reasonably safe, but now suppose that an authoritarian government knows precisely when its citizens are online. Then by linking the two datasets, the government may be able to track those who post derogatory statements about their leaders.

Perhaps the strongest definition of privacy of databases and datasets is due to Dalenius [56], who suggests that "nothing about an individual should be learnable from the database that cannot be learned without access to the database." But quickly, one can see that it is essentially impossible to reconcile this idea with scientific advancement. Consider, for example, a situation where we

perform a study on smoking, and discover that smoking causes cancer. We publish the result, but now we have "compromised" the privacy of everyone who smokes who did not participate in the study: we know they are more likely to get cancer.

In each of these cases, the biggest challenge is one of side information: how can we be sure that, when releasing a particular statistic, dataset, or other quantity that no adversary will be able to infer sensitive data about participants in our study? We articulate three desiderata that—we believe—suffice for satisfactory definitions of privacy. In discussion of private releases of data, we require a bit of vocabulary. We term a (randomized) algorithm releasing data either a privacy *mechanism*, consistent with much of the literature in privacy, or a *channel*, mapping from the input sample to some output space, in keeping with our statistical and information-theoretic focus. In no particular order, we wish our privacy mechanism, which takes as input a sample $X_1^n \in \mathcal{X}^n$ and releases some Z to satisfy the following.

i. Given the output Z, even an adversary knowing everyone in the study (excepting one person) should not be able to test whether you belong to the study.

ii. If you participate in multiple "private" studies, there should be some graceful degradation in the privacy protections, rather than a catastrophic failure. As part of this, any definition should guarantee that further processing of the output Z of a private mechanism $X_1^n \to Z$, in the form of the Markov chain $X_1^n \to Z \to Y$, should not allow further compromise of privacy (that is, a data-processing inequality). Additional participation in "private" studies should continue to provide little additional information.

iii. The mechanism $X_1^n \to Z$ should be resilient to side information: even if someone knows something about you, he should learn little about you if you belong to X_1^n, and this should remain true even if the adversary later gleans more information about you.

The third desideratum is perhaps most elegantly phrased via a Bayesian perspective, where an adversary has some prior beliefs π on the membership of a dataset (these prior beliefs can then capture any side information the adversary has). The strongest adversary has a prior supported on two samples $\{x_1, \ldots, x_n\}$ and $\{x'_1, \ldots, x'_n\}$ differing in only a single element; a private mechanism would then guarantee the adversary's posterior beliefs (after the release $X_1^n \to Z$) should not change significantly.

Before continuing addressing these challenges, we take a brief detour to establish notation for the remainder of the chapter. It will be convenient to consider randomized procedures acting on samples themselves; a sample x_1^n is cleary isomorphic to the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i}$, and for two empirical distributions P_n and P'_n supported on $\{x_1, \ldots, x_n\}$ and $\{x'_1, \ldots, x'_n\}$, we evidently have

$$n \left\|P_n - P'_n\right\|_{\mathrm{TV}} = d_{\mathrm{ham}}(\{x_1, \ldots, x_n\}, \{x'_1, \ldots, x'_n\}),$$

and so we will identify samples with their empirical distributions. With this notational convenience in place, we then identify

$$\mathcal{P}_n = \left\{P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i} \mid x_i \in \mathcal{X}\right\}$$

as the set of all empirical distributions on n points in $\mathcal{X}$ and we also abuse notation in an obvious way to define $d_{\mathrm{ham}}(P_n, P'_n) := n \left\|P_n - P'_n\right\|_{\mathrm{TV}}$ as the number of differing observations in the samples P_n and P'_n represent. A mechanism M is then a (typically) randomized mapping $M : \mathcal{P}_n \to \mathcal{Z}$,

which we can identify with its induced Markov channel Q from $\mathcal{X}^n \to \mathcal{Z}$; we use the equivalent views as is convenient.

The challenges of side information motivate Dwork et al.'s definition of *differential privacy* [74]. The key in differential privacy is that the noisy channel releasing statistics provides guarantees of bounded likelihood ratios between neighboring samples, that is, samples differing in only a single entry.

Definition 7.1 (Differential privacy)**.** *Let $M : \mathcal{P}_n \to \mathcal{Z}$ be a randomized mapping. Then M is* ε-differentially private *if for all (measurable) sets $S \subset \mathcal{Z}$ and all $P_n, P_n' \in \mathcal{P}_n$ with $d_{\rm ham}(P_n, P_n') \le 1$,*

$$\frac{\mathbb{P}(M(P_n) \in S)}{\mathbb{P}(M(P_n') \in S)} \le e^\varepsilon. \tag{7.1.1}$$

The intuition and original motivation for this definition are that an individual has little incentive to participate (or not participate) in a study, as the individual's data has limited effect on the outcome.

The model (7.1.1) of differential privacy presumes that there is a trusted curator, such as a hospital, researcher, or corporation, who can collect all the data into one centralized location, and it is consequently known as the *centralized* model. A stronger model of privacy is the *local model*, in which data providers trust no one, not even the data collector, and privatize their individual data before the collector even sees it.

Definition 7.2 (Local differential privacy)**.** *A channel Q from $\mathcal{X}$ to $\mathcal{Z}$ is* ε-locally differentially private *if for all measurable $S \subset \mathcal{Z}$ and all $x, x' \in \mathcal{X}$,*

$$\frac{Q(Z \in S \mid x)}{Q(Z \in S \mid x')} \le e^\varepsilon. \tag{7.1.2}$$

It is clear that Definition 7.2 and the condition (7.1.2) are stronger than Definition 7.1: when samples $\{x_1, \ldots, x_n\}$ and $\{x_1', \ldots, x_n'\}$ differ in at most one observation, then the local model (7.1.2) guarantees that the densities

$$\frac{dQ(Z_1^n \mid \{x_i\})}{dQ(Z_1^n \mid \{x_i'\})} = \prod_{i=1}^n \frac{dQ(Z_i \mid x_i)}{dQ(Z_i \mid x_i')} \le e^\varepsilon,$$

where the inequality follows because only a single ratio may contain $x_i \neq x_i'$.

In the remainder of this introductory section, we provide a few of the basic mechanisms in use in differential privacy, then discuss its "semantics," that is, its connections to the three desiderata we outline above. In the coming sections, we revisit a few more advanced topics, in particular, the composition of multiple private mechanisms and a few weakenings of differential privacy, as well as more sophisticated examples.

7.1.1 Basic mechanisms

The basic mechanisms in either the local or centralized models of differential privacy use some type of noise addition to ensure privacy. We begin with the simplest and oldest mechanism, randomized response, for local privacy, due to Warner [173] in 1965.

Example 7.1.1 (Randomized response)**:** We wish to have a participant in a study answer a yes/no question about a sensitive topic (for example, drug use). That is, we would like to

estimate the proportion of the population with a characteristic (versus those without); call these groups 0 and 1. Rather than ask the participant to answer the question specifically, however, we give them a spinner with a face painted in two known areas, where the first corresponds to group 0 and has area $e^\varepsilon/(1+e^\varepsilon)$ and the second to group 1 and has area $1/(1+e^\varepsilon)$. Thus, when the participant spins the spinner, it lands in group 0 with probability $e^\varepsilon/(1+e^\varepsilon)$. Then we simply ask the participant, upon spinning the spinner, to answer "Yes" if he or she belongs to the indicated group, "No" otherwise.

Let us demonstrate that this randomized response mechanism provides ε-local differential privacy. Indeed, we have

$$\frac{Q(\text{Yes} \mid x = 0)}{Q(\text{Yes} \mid x = 1)} = e^{-\varepsilon} \quad \text{and} \quad \frac{Q(\text{No} \mid x = 0)}{Q(\text{No} \mid x = 1)} = e^{\varepsilon},$$

so that $Q(Z = z \mid x)/Q(Z = z \mid x') \in [e^{-\varepsilon}, e^\varepsilon]$ for all x, z. That is, the randomized response channel provides ε-local privacy. $\diamond$

The interesting question is, of course, whether we can still use this channel to estimate the proportion of the population with the sensitive characteristic. Indeed, we can. We can provide a somewhat more general analysis, however, which we now do so that we can give a complete example.

Example 7.1.2 (Randomized response, continued)**:** Suppose that we have an attribute of interest, x, taking the values $x \in \{1, \ldots, k\}$. Then we consider the channel (of Z drawn conditional on x)

$$Z = \begin{cases} x & \text{with probability } \frac{e^\varepsilon}{k-1+e^\varepsilon} \\ \mathsf{Uniform}([k] \setminus \{x\}) & \text{with probability } \frac{k-1}{k-1+e^\varepsilon}. \end{cases}$$

This (generalized) randomized response mechanism is evidently ε-locally private, satisfying Definition 7.2.

Let $p \in \mathbb{R}^k_+$, $p^T\mathbf{1} = 1$ indicate the true probabilities $p_i = \mathbb{P}(X = i)$. Then by inspection, we have

$$\mathbb{P}(Z = i) = p_i \frac{e^\varepsilon}{k-1+e^\varepsilon} + (1-p_i)\frac{1}{k-1+e^\varepsilon} = p_i \frac{e^\varepsilon - 1}{e^\varepsilon + k - 1} + \frac{1}{e^\varepsilon + k - 1}.$$

Thus, letting $\widehat{c}_n \in \mathbb{R}^k_+$ denote the empirical proportion of the Z observations in a sample of size n, we have

$$\widehat{p}_n := \frac{e^\varepsilon + k - 1}{e^\varepsilon - 1}\left(\widehat{c}_n - \frac{1}{e^\varepsilon + k - 1}\mathbf{1}\right)$$

satisfies $\mathbb{E}[\widehat{p}_n] = p$, and we also have

$$\mathbb{E}\left[\|\widehat{p}_n - p\|_2^2\right] = \left(\frac{e^\varepsilon + k - 1}{e^\varepsilon - 1}\right)^2 \mathbb{E}\left[\|\widehat{c}_n - \mathbb{E}[\widehat{c}_n]\|_2^2\right] = \frac{1}{n}\left(\frac{e^\varepsilon + k - 1}{e^\varepsilon - 1}\right)^2 \sum_{j=1}^k \mathbb{P}(Z = j)(1 - \mathbb{P}(Z = j)).$$

As $\sum_j \mathbb{P}(Z = j) = 1$, we always have the bound $\mathbb{E}[\|\widehat{p}_n - p\|_2^2] \le \frac{1}{n}(\frac{e^\varepsilon + k - 1}{e^\varepsilon - 1})^2$.

We may consider two regimes for simplicity: when $\varepsilon \le 1$ and when $\varepsilon \ge \log k$. In the former case—the high privacy regime—we have $\frac{1}{k} \lesssim \mathbb{P}(Z = i) \lesssim \frac{1}{k}$, so that the mean ℓ_2 squared error scales as $\frac{1}{n}\frac{k^2}{\varepsilon^2}$. When $\varepsilon \ge \log k$ is large, by contrast, we see that the error scales at worst as $\frac{1}{n}$, which is the "non-private" mean squared error. $\diamond$

While randomized response is essentially the standard mechanism in locally private settings, in centralized privacy, the "standard" mechanism is Laplace noise addition because of its exponential tails. In this case, we require a few additional definitions. Suppose that we wish to release some d-dimensional function $f(P_n)$ of the sample distribution P_n (equivalently, the associated sample X_1^n), where f takes values in $\mathbb{R}^d$. In the case that f is Lipschitz with respect to the Hamming metric—that is, the counting metric on $\mathcal{X}^n$—it is relatively straightforward to develop private mechanisms. To better reflect the nomenclature in the privacy literature and easier use in our future development, for $p \in [1, \infty]$ we define the *global sensitivity* of f by

$$\mathrm{GS}_p(f) := \sup_{P_n, P_n' \in \mathcal{P}_n} \left\{ \left\| f(P_n) - f(P_n') \right\|_p \mid d_{\mathrm{ham}}(P_n, P_n') \le 1 \right\}.$$

This is simply the Lipschitz constant of f with respect to the Hamming metric. The global sensitivity is a convenient metric, because it allows simple noise addition strategies.

Example 7.1.3 (Laplace mechanisms)**:** Recall the Laplace distribution, parameterized by a shape parameter β, which has density on $\mathbb{R}$ defined by

$$p(w) = \frac{1}{2\beta} \exp(-|w|/\beta),$$

and the analogous d-dimensional variant, which has density

$$p(w) = \frac{1}{(2\beta)^2} \exp(- \left\| w \right\|_1 / \beta).$$

If $W \sim \mathsf{Laplace}(\beta)$, $W \in \mathbb{R}$, then $\mathbb{E}[W] = 0$ by symmetry, while $\mathbb{E}[W^2] = \frac{1}{\beta} \int_0^\infty w^2 e^{-w/\beta} = 2\beta^2$.

Suppose that $f : \mathcal{P}_n \to \mathbb{R}^d$ has finite global sensitivity for the ℓ_1-norm,

$$\mathrm{GS}_1(f) = \sup \left\{ \left\| f(P_n) - f(P_n') \right\|_1 \mid d_{\mathrm{ham}}(P_n, P_n') \le 1, P_n, P_n' \in \mathcal{P}_n \right\}.$$

Letting $L = \mathrm{GS}_1(f)$ be the Lipschitz constant for simplicity, if we consider the mechanism defined by the addition of $W \in \mathbb{R}^d$ with independent $\mathsf{Laplace}(L/\varepsilon)$ coordinates,

$$Z := f(P_n) + W, \quad W_j \stackrel{\mathrm{iid}}{\sim} \mathsf{Laplace}(L/\varepsilon), \tag{7.1.3}$$

we have that Z is ε-differentially private. Indeed, for samples P_n, P_n' differing in at most a single example, Z has density ratio

$$\frac{q(z \mid P_n)}{q(z \mid P_n')} = \exp\left(-\frac{\varepsilon}{L} \left\| f(P_n) - z \right\|_1 + \frac{\varepsilon}{L} \left\| f(P_n') - z \right\|_1 \right) \le \exp\left(\frac{\varepsilon}{L} \left\| f(P_n) - f(P_n') \right\|_1 \right) \le \exp(\varepsilon)$$

by the triangle inequality and that f is L-Lipschitz with respect to the Hamming metric. Thus Z is ε-differentially private. Moreover, we have

$$\mathbb{E}[\left\| Z - f(P_n) \right\|_2^2] = \frac{2d\mathrm{GS}_1(f)^2}{\varepsilon^2},$$

so that if L is small, we may report the value of f accurately. $\diamond$

The most common instances and applications of the Laplace mechanism are in estimation of means and histograms. Let us demonstrate more carefully worked examples in these two cases.

Example 7.1.4 (Private one-dimensional mean estimation)**:** Suppose that we have variables X_i taking values in $[-b, b]$ for some $b < \infty$, and wish to estimate $\mathbb{E}[X]$. A natural function to release is then $f(X_1^n) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. This has Lipschitz constant $2b/n$ with respect to the Hamming metric, because for any two samples $x, x' \in [-b, b]^n$ differing in only entry i, we have

$$|f(x) - f(x')| = \frac{1}{n}|x_i - x_i'| \le \frac{2b}{n}$$

because $x_i \in [-b, b]$. Thus the Laplace mechanism (7.1.3) with the choice variance $W \sim \mathsf{Laplace}(2b/(n\varepsilon))$ yields

$$\mathbb{E}[(Z - \mathbb{E}[X])^2] = \mathbb{E}[(\overline{X}_n - \mathbb{E}[X])^2] + \mathbb{E}[(Z - \overline{X}_n)^2] = \frac{1}{n}\mathrm{Var}(X) + \frac{8b^2}{n^2\varepsilon^2} \le \frac{b^2}{n} + \frac{8b^2}{n^2\varepsilon^2}.$$

We can privately release means with little penalty so long as $\varepsilon \gg n^{-1/2}$. $\diamond$

Example 7.1.5 (Private histogram (multinomial) release)**:** Suppose that we wish to estimate a multinomial distribution, or put differently, a histogram. That is, we have observations $X \in \{1, \ldots, k\}$, where k may be large, and wish to estimate $p_j := \mathbb{P}(X = j)$ for $j = 1, \ldots, k$. For a given sample x_1^n, the empirical count vector $\widehat{p}_n$ with coordinates $\widehat{p}_{n,j} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = j\}$ satisfies

$$\mathrm{GS}_1(\widehat{p}_n) = \frac{2}{n}$$

because swapping a single example x_i for x_i' may change the counts for at most two coordinates j, j' by 1. Consequently, the Laplace noise addition mechanism

$$Z = \widehat{p}_n + W, \quad W_j \stackrel{\mathrm{iid}}{\sim} \mathsf{Laplace}\left(\frac{2}{n\varepsilon}\right)$$

satisfies

$$\mathbb{E}[\|Z - \widehat{p}_n\|_2^2] = \frac{8k}{n^2\varepsilon^2}$$

and consequently

$$\mathbb{E}[\|Z - p\|_2^2] = \frac{8k}{n^2\varepsilon^2} + \frac{1}{n}\sum_{j=1}^k p_j(1 - p_j) \le \frac{8k}{n^2\varepsilon^2} + \frac{1}{n}.$$

This example shows one of the challenges of differentially private mechanisms: even in the case where the quantity of interest is quite stable (insensitive to changes in the underlying sample, or has small Lipschitz constant), it may be the case that the resulting mechanism adds noise that introduces some dimension-dependent scaling. In this case, the conditions on privacy levels acceptable for good estimation—in that the rate of convergence is no different from the non-private case, which achieves $\mathbb{E}[\|\widehat{p}_n - p\|_2^2] = \frac{1}{n}\sum_{j=1}^k p_j(1 - p_j) \le \frac{1}{n}$ are that $\varepsilon \gg \frac{k}{n}$. Thus, in the case that the histogram has a large number of bins, the naive noise addition strategy cannot provide as much protection without sacrificing efficiency.

If instead of ℓ_2-error we consider ℓ_∞ error, it is possible to provide somewhat more satisfying results in this case. Indeed, we know that $\mathbb{P}(\|W\|_\infty \ge t) \le k\exp(-t/b)$ for $W_j \stackrel{\mathrm{iid}}{\sim} \mathsf{Laplace}(b)$, so that in the mechanism above we have

$$\mathbb{P}(\|Z - \widehat{p}_n\|_\infty \ge t) \le k\exp\left(-\frac{tn\varepsilon}{2}\right) \quad \text{all } t \ge 0,$$

so using that each coordinate of $\widehat{p}_n$ is 1-sub-Gaussian, we have

$$\mathbb{E}[\|Z - p\|_\infty] \le \mathbb{E}[\|\widehat{p}_n - p\|_\infty] + \mathbb{E}[\|W\|_\infty] \le \sqrt{\frac{2\log k}{n}} + \inf_{t \ge 0} \left\{ t + \frac{2k}{n\varepsilon} \exp\left(-\frac{tn\varepsilon}{2}\right) \right\}$$
$$\le \sqrt{\frac{2\log k}{n}} + \frac{2\log k}{n\varepsilon} + \frac{2}{n\varepsilon}.$$

In this case, then, whenever $\varepsilon \gg (n/\log k)^{-1/2}$, we obtain rate of convergence at least $\sqrt{2\log k/n}$, which is a bit loose (as we have not controlled the variance of $\widehat{p}_n$), but somewhat more satisfying than the k-dependent penalty above. $\diamond$

7.1.2 Resilience to side information, Bayesian perspectives, and data processing

One of the major challenges in the definition of privacy is to protect against side information, especially because in the future, information about you may be compromised, allowing various linkage attacks. With this in mind, we return to our three desiderata. First, we note the following simple fact: if Z is a differentially private view of a sample X_1^n (or associated empirical distribution P_n), then any downstream functions Y are also differentially private. That is, if we have the Markov chain $P_n \to Z \to Y$, then for any $P_n, P_n' \in \mathcal{P}_n$ with $d_{\rm ham}(P_n, P_n') \le 1$, we have for any set A that

$$\frac{\mathbb{P}(Y \in A \mid x)}{\mathbb{P}(Y \in A \mid x')} = \frac{\int P(Y \in A \mid z) q(z \mid P_n) d\mu(z)}{\int P(Y \in A \mid z) q(z \mid P_n') d\mu(z)} \le e^\varepsilon \frac{\int P(Y \in A \mid z) q(z \mid P_n') d\mu(z)}{\int P(Y \in A \mid z) q(z \mid P_n') d\mu(z)} = e^\varepsilon.$$

That is, any type of post-processing cannot reduce privacy.

With this simple idea out of the way, let us focus on our testing-based desideratum. In this case, we consider a testing scenario, where an adversary wishes to test two hypotheses against one another, where the hypotheses are

$$H_0 : X_1^n = x_1^n \quad \text{vs.} \quad H_1 : X_1^n = (x_1^{i-1}, x_i', x_{i+1}^n),$$

so that the samples under H_0 and H_1 differ only in the ith observation $X_i \in \{x_i, x_i'\}$. Now, for a channel taking inputs from $\mathcal{X}^n$ and outputting $Z \in \mathcal{Z}$, we define *ε-conditional hypothesis testing* privacy by saying that

$$Q(\Psi(Z) = 1 \mid H_0, Z \in A) + Q(\Psi(Z) = 0 \mid H_1, Z \in A) \ge 1 - \varepsilon \tag{7.1.4}$$

for all sets $A \subset \mathcal{Z}$ satisfying $Q(A \mid H_0) > 0$ and $Q(A \mid H_1) > 0$. That is, roughly, no matter *what* value Z takes on, the probability of error in a test of whether H_0 or H_1 is true—even with knowledge of $x_j, j \ne i$—is high. We then have the following proposition.

Proposition 7.1.6. *Assume the channel Q is ε-differentially private. Then Q is also $\bar{\varepsilon} = 1 - e^{-2\varepsilon} \le 2\varepsilon$-conditional hypothesis testing private.*

Proof Let Ψ be any test of H_0 versus H_1, and let $B = \{z \mid \Psi(z) = 1\}$ be the acceptance region of the test. Then

$$Q(B \mid H_0, Z \in A) + Q(B^c \mid H_1, Z \in A) = \frac{Q(A, B \mid H_0)}{Q(A \mid H_0)} + \frac{Q(A, B^c \mid H_1)}{Q(A \mid H_1)}$$
$$\ge e^{-2\varepsilon} \frac{Q(A, B \mid H_1)}{Q(A \mid H_1)} + \frac{Q(A, B^c \mid H_1)}{Q(A \mid H_1)}$$
$$\ge e^{-2\varepsilon} \frac{Q(A, B \mid H_1) + Q(A, B^c \mid H_1)}{Q(A \mid H_1)},$$

where the first inequality uses ε-differential privacy. Then we simply note that $Q(A, B \mid H_1) + Q(A, B^c \mid H_1) = Q(A \mid H_1)$. □

So we see that (roughly), even conditional on the output of the channel, we still cannot test whether the initial dataset was x or x' whenever x, x' differ in only a single observation.

An alternative perspective is to consider a Bayesian one, which allows us to more carefully consider side information. In this case, we consider the following thought experiment. An adversary has a set of prior beliefs π on $\mathcal{X}^n$, and we consider the adversary's posterior $\pi(\cdot \mid Z)$ induced by observing the output Z of some mechanism M. In this case, *Bayes factors*, which measure how much prior and posterior distributions differ after observations, provide one immediate perspective.

Proposition 7.1.7. *A mechanism $M : \mathcal{P}_n \to \mathcal{Z}$ is ε-differentially private if and only if for any prior distribution π on $\mathcal{P}_n$ and any observation $z \in \mathcal{Z}$, the posterior odds satisfy*

$$\frac{\pi(P_n \mid z)}{\pi(P_n' \mid z)} \le e^{\varepsilon}$$

for all $P_n, P_n' \in \mathcal{P}_n$ with $d_{\rm ham}(P_n, P_n') \le 1$.

Proof Let q be the associated density of $Z = M(\cdot)$ (conditional or marginal). We have $\pi(P_n \mid z) = q(z \mid P_n)\pi(P_n)/q(z)$. Then

$$\frac{\pi(P_n \mid z)}{\pi(P_n' \mid z)} = \frac{q(z \mid P_n)\pi(P_n)}{q(z \mid P_n')\pi(P_n')} \le e^{\varepsilon}\frac{\pi(P_n)}{\pi(P_n')}$$

for all z, P_n, P_n' if and only if M is ε-differentially private. □

Thus we see that private channels mean that prior and posterior odds between two neighboring samples cannot change substantially, no matter what the observation Z actually is.

For an an alternative view, we consider a somewhat restricted family of prior distributions, where we now take the view of a sample $x_1^n \in \mathcal{X}^n$. There is some annoyance in this calculation in that the *order* of the sample may be important, but it at least gets toward some semantic interpretation of differential privacy. We consider the adversary's beliefs on whether a particular value x belongs to the sample, but more precisely, we consider whether $X_i = x$. We assume that the prior density π on $\mathcal{X}^n$ satisfies

$$\pi(x_1^n) = \pi_{\setminus i}(x_{\setminus i})\pi_i(x_i), \tag{7.1.5}$$

where $x_{\setminus i} = (x_1^{i-1}, x_{i+1}^n) \in \mathcal{X}^{n-1}$. That is, the adversary's beliefs about person i in the dataset are independent of his beliefs about the other members of the dataset. (We assume that π is a density with respect to a measure μ on $\mathcal{X}^{n-1} \times \mathcal{X}$, where $d\mu(s, x) = d\mu(s)d\mu(x)$.) Under the condition (7.1.5), we have the following proposition.

Proposition 7.1.8. *Let Q be an ε-differentially private channel and let π be any prior distribution satisfying condition (7.1.5). Then for any z, the posterior density π_i on X_i satisfies*

$$e^{-\varepsilon}\pi_i(x) \le \pi_i(x \mid Z = z) \le e^{\varepsilon}\pi_i(x).$$

Proof We abuse notation and for a sample $s \in \mathcal{X}^{n-1}$, where $s = (x_1^{i-1}, x_{i+1}^n)$, we let $s \oplus_i x = (x_1^{i-1}, x, x_{i+1}^n)$. Letting μ be the base measure on $\mathcal{X}^{n-1} \times \mathcal{X}$ with respect to which π is a density and $q(\cdot \mid x_1^n)$ be the density of the channel Q, we have

$$\begin{aligned}
\pi_i(x \mid Z = z) &= \frac{\int_{s \in \mathcal{X}^{n-1}} q(z \mid s \oplus_i x) \pi(s \oplus_i x) d\mu(s)}{\int_{s \in \mathcal{X}^{n-1}} \int_{x' \in \mathcal{X}} q(z \mid s \oplus_i x') \pi(s \oplus_i x') d\mu(s, x')} \\
&\stackrel{(\star)}{\le} e^{\varepsilon} \frac{\int_{s \in \mathcal{X}^{n-1}} q(z \mid s \oplus_i x) \pi(s \oplus_i x) d\mu(s)}{\int_{s \in \mathcal{X}^{n-1}} \int_{x' \in \mathcal{X}} q(z \mid s \oplus_i x) \pi(s \oplus_i x') d\mu(s) d\mu(x')} \\
&= e^{\varepsilon} \frac{\int_{s \in \mathcal{X}^{n-1}} q(z \mid s \oplus_i x) \pi_{\setminus i}(s) d\mu(s) \pi_i(x)}{\int_{s \in \mathcal{X}^{n-1}} q(z \mid s \oplus_i x) \pi_{\setminus i}(s) d\mu(s) \int_{x' \in \mathcal{X}} \pi_i(x') d\mu(x')} \\
&= e^{\varepsilon} \pi_i(x),
\end{aligned}$$

where inequality $(\star)$ follows from ε-differential privacy. The lower bound is similar. □

Roughly, however, we see that Proposition 7.1.8 captures the idea that even if an adversary has substantial prior knowledge—in the form of a prior distribution π on the ith value X_i and everything else in the sample—the posterior cannot change much.

7.2 Weakenings of differential privacy

One challenge with the definition of differential privacy is that it can sometimes require the addition of more noise to a desired statistic than is practical for real use. Moreover, the privacy considerations interact in different ways with geometry: as we saw in Example 7.1.5, the Laplace mechanism adds noise that introduces dimension-dependent scaling, which we discuss more in Example 7.2.9. Consequently, it is of interest to develop weaker notions that—at least hopefully—still provide appropriate and satisfactory privacy protections. To that end, we develop two additional types of privacy that allow the development of more sophisticated and lower-noise mechanisms than standard differential privacy; their protections are necessarily somewhat weaker but are typically satisfactory.

We begin with a definition that allows (very rare) catostrophic privacy breaches—as long as the probability of this event is extremely small (say, 10^{-20}), these may be acceptable.

Definition 7.3. *Let $\varepsilon, \delta \ge 0$. A mechanism $M : \mathcal{P}_n \to \mathcal{Z}$ is* (ε, δ)-differentially private *if for all (measurable) sets $S \subset \mathcal{Z}$ and all neighboring samples P_n, P_n',*

$$\mathbb{P}(M(P_n) \in S) \le e^{\varepsilon} \mathbb{P}(M(P_n') \in S) + \delta. \tag{7.2.1}$$

One typically thinks of δ in the definition above as satisfying $\delta = \delta_n$, where $\delta_n \ll n^{-k}$ for any $k \in \mathbb{N}$. (That is, δ decays super-polynomially to zero.) Some practitioners contend that all real-world differentially private algorithms are in fact (ε, δ)-differentially private: while one may use cryptographically secure random number generators, there is some possibility (call this δ) that a cryptographic key may leak, or an encoding may be broken, in the future, making any mechanism (ε, δ)-private at best for some $\delta > 0$.

An alternative definition of privacy is based on Rényi divergences between distributions. These are essentially simply monotonically transformed f divergences (recall Chapter 2.2), though their structure is somewhat more amenable to analysis, especially in our contexts. With that in mind, we define

Definition 7.4. *Let P and Q be distributions on a space $\mathcal{X}$ with densities p and q (with respect to a measure μ). For $\alpha \in [1, \infty]$, the* Rényi-α-divergence *between P and Q is*

$$D_\alpha(P\|Q) := \frac{1}{\alpha - 1} \log \int \left(\frac{p(x)}{q(x)}\right)^\alpha q(x) d\mu(x).$$

Here, the values $\alpha \in \{1, \infty\}$ are defined in terms of their respective limits.

Rényi divergences satisfy $\exp((\alpha - 1)D_\alpha(P\|Q)) = 1 + D_f(P\|Q)$, i.e., $D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log(1 + D_f(P\|Q))$, for the f-divergence defined by $f(t) = t^\alpha - 1$, so that they inherit a number of the properties of such divergences. We enumerate a few here for later reference.

Proposition 7.2.1 (Basic facts on Rényi divergence)**.** *Rényi divergences satisfy the following.*

i. The divergence $D_\alpha(P\|Q)$ is non-decreasing in α.

ii. $\lim_{\alpha \downarrow 1} D_\alpha(P\|Q) = D_{\rm kl}(P\|Q)$ *and* $\lim_{\alpha \uparrow \infty} D_\alpha(P\|Q) = \sup\{t \mid Q(p(X)/q(X) \geq t) > 0\}$.

iii. Let $K(\cdot \mid x)$ be a Markov kernel from $\mathcal{X} \to \mathcal{Z}$ as in Proposition 2.2.13, and let K_P and K_Q be the induced marginals of P and Q under K, respectively. Then $D_\alpha(K_P\|K_Q) \leq D_\alpha(P\|Q)$.

We leave the proof of this proposition as Exercise 7.1, noting that property i is a consequence of Hölder's inequality, property ii is by L'Hopital's rule, and property iii is an immediate consequence of Proposition 2.2.13. Rényi divergences also tensorize nicely—generalizing the tensorization properties of KL-divergence and information of Chapter 2 (recall the chain rule (2.1.6) for KL-divergence)—and we return to this later. As a preview, however, these tensorization properties allow us to prove that the composition of multiple private data releases remains appropriately private.

With these preliminaries in place, we can then provide

Definition 7.5 (Rényi-differential privacy)**.** *Let $\varepsilon \geq 0$ and $\alpha \in [1, \infty]$. A channel Q from $\mathcal{P}_n$ to output space $\mathcal{Z}$ is* (ε, α)-Rényi private *if for all neighboring samples $P_n, P_n' \in \mathcal{P}_n$,*

$$D_\alpha\big(Q(\cdot \mid P_n)\|Q(\cdot \mid P_n')\big) \leq \varepsilon. \tag{7.2.2}$$

Clearly, any ε-differentially private channel is also (ε, α)-Rényi private for any $\alpha \geq 1$; as we soon see, we can provide tighter guarantees than this.

7.2.1 Basic mechanisms

We now describe a few of the basic mechanisms that provide guarantees of (ε, δ)-differential privacy and (ε, α)-Rényi privacy. The advantage for these settings is that they allow mechanisms that more naturally handle vectors in ℓ_2, and smoothness with respect to Euclidean norms, than with respect to ℓ_1, which is most natural for pure ε-differential privacy. A starting point is the following example, which we will leverage frequently.

Example 7.2.2 (Rényi divergence between Gaussian distributions)**:** Consider normal distributions $\mathsf{N}(\mu_0, \Sigma)$ and $\mathsf{N}(\mu_1, \Sigma)$. Then

$$D_\alpha(\mathsf{N}(\mu_0, \Sigma)\|\mathsf{N}(\mu_1, \Sigma)) = \frac{\alpha}{2}(\mu_0 - \mu_1)^T \Sigma^{-1}(\mu_0 - \mu_1). \tag{7.2.3}$$

To see this equality, we compute the appriate integral of the densities. Let p and q be the densities of $\mathsf{N}(\mu_0, \Sigma)$ and $\mathsf{N}(\mu_1, \Sigma)$, respectively. Then letting $\mathbb{E}_{\mu_1}$ denote expectation over $X \sim \mathsf{N}(\mu_1, \Sigma)$, we have

$$\begin{aligned}\int \left(\frac{p(x)}{q(x)}\right)^\alpha q(x)dx &= \mathbb{E}_{\mu_1}\left[\exp\left(-\frac{\alpha}{2}(X-\mu_0)^T\Sigma^{-1}(X-\mu_0) + \frac{\alpha}{2}(X-\mu_1)^T\Sigma^{-1}(X-\mu_1)\right)\right] \\ &\stackrel{(i)}{=} \mathbb{E}_{\mu_1}\left[\exp\left(-\frac{\alpha}{2}(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1) + \alpha(\mu_0-\mu_1)^T\Sigma^{-1}(X-\mu_1)\right)\right] \\ &\stackrel{(ii)}{=} \exp\left(-\frac{\alpha}{2}(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1) + \frac{\alpha^2}{2}(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1)\right),\end{aligned}$$

where equality (i) is simply using that $(x-a)^2 - (x-b)^2 = (a-b)^2 + 2(b-a)(x-b)$ and equality (ii) follows because $(\mu_0-\mu_1)^T\Sigma^{-1}(X-\mu_1) \sim \mathsf{N}(0, (\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0))$ under $X \sim \mathsf{N}(\mu_1, \Sigma)$. Noting that $-\alpha + \alpha^2 = \alpha(\alpha-1)$ and taking logarithms gives the result. $\diamond$

Example 7.2.2 is the key to developing different privacy-preserving schemes under Rényi privacy. Let us reconsider Example 7.1.3, except that instead of assuming the function f of interest is smooth with respect to ℓ_1 norm, we use the ℓ_2-norm.

Example 7.2.3 (Gaussian mechanisms)**:** Suppose that $f : \mathcal{P}_n \to \mathbb{R}^d$ has Lipschitz constant L with respect to the ℓ_2-norm (for the Hamming metric $d_{\rm ham}$), that is, global ℓ_2-sensitivity

$$\mathrm{GS}_2(f) = \sup\left\{\left\|f(P_n) - f(P_n')\right\|_2 \mid d_{\rm ham}(P_n, P_n') \le 1\right\} \le L.$$

Then, for any variance $\sigma^2 > 0$, we have that the mechanism

$$Z = f(P_n) + W, \quad W \sim \mathsf{N}(0, \sigma^2 I)$$

satisfies

$$D_\alpha\big(\mathsf{N}(f(P_n), \sigma^2) \| \mathsf{N}(f(P_n'), \sigma^2)\big) = \frac{\alpha}{2\sigma^2}\left\|f(P_n) - f(P_n')\right\|_2^2 \le \frac{\alpha}{2\sigma^2}L^2$$

for neighboring samples P_n, P_n'. Thus, if we have Lipschitz constant L and desire (ε, α)-Rényi privacy, we may take $\sigma^2 = \frac{L^2\alpha}{2\varepsilon}$, and then the mechanism

$$Z = f(P_n) + W \quad W \sim \mathsf{N}\left(0, \frac{L^2\alpha}{2\varepsilon}I\right) \tag{7.2.4}$$

satisfies (ε, α)-Rényi privacy. $\diamond$

Certain special cases can make this more concrete. Indeed, suppose we wish to estimate a mean $\mathbb{E}[X]$ where $X_i \stackrel{\rm iid}{\sim} P$ for some distribution P such that $\|X_i\|_2 \le r$ with probability 1 for some radius.

Example 7.2.4 (Bounded mean estimation with Gaussian mechanisms)**:** Letting $f(X_1^n) = \overline{X}_n$ be the sample mean, where X_i satisfy $\|X_i\|_2 \le r$ as above, we see immediately that

$$\mathrm{GS}_2(f) = \frac{2r}{n}.$$

In this case, the Gaussian mechanism (7.2.4) with $L = \frac{2r}{n}$ yields

$$\mathbb{E}\left[\left\|Z - \overline{X}_n\right\|_2^2\right] = \mathbb{E}[\|W\|_2^2] = \frac{2dr^2\alpha}{n^2\varepsilon}.$$

Then we have

$$\mathbb{E}[\|Z - \mathbb{E}[X]\|_2^2] = \mathbb{E}[\|\overline{X}_n - \mathbb{E}[X]\|_2^2] + \mathbb{E}[\|Z - \overline{X}_n\|_2^2] \le \frac{r^2}{n} + \frac{2dr^2\alpha}{n^2\varepsilon}.$$

It is not immediately apparent how to compare this quantity to the case for the Laplace mechanism in Example 7.1.3, but we will return to this shortly once we have developed connections between the various privacy notions we have developed. $\diamond$

7.2.2 Connections between privacy measures

An important consideration in our development of privacy definitions and mechanisms is to understand the relationships between the definitions, and when a channel Q satisfying one of the definitions satisfies one of our other definitions. Thus, we collect a few different consequences of our definitions, which help to show the various definitions are stronger or weaker than others.

First, we argue that ε-differential privacy implies stronger values of Rényi-differential privacy.

Proposition 7.2.5. *Let $\varepsilon \ge 0$ and let P and Q be distributions such that $e^{-\varepsilon} \le P(A)/Q(A) \le e^{\varepsilon}$ for all measurable sets A. Then for any $\alpha \in [1, \infty]$,*

$$D_\alpha(P\|Q) \le \min\left\{\frac{3\alpha}{2}\varepsilon^2, \varepsilon\right\}.$$

As an immediate corollary, we have

Corollary 7.2.6. *Let $\varepsilon \ge 0$ and assume that Q is ε-differentially private. Then for any $\alpha \ge 1$, Q is $(\min\{\frac{3\alpha}{2}\varepsilon^2, \varepsilon\}, \alpha)$-Rényi private.*

Before proving the proposition, let us see its implications for Example 7.2.4 versus estimation under ε-differential privacy. Let $\varepsilon \le 1$, so that roughly to have "similar" privacy, we require that our Rényi private channels satisfy $D_\alpha(Q(\cdot \mid x)\|Q(\cdot \mid x')) \le \varepsilon^2$. The ℓ_1-sensitivity of the mean satisfies $\|\overline{x}_n - \overline{x'}_n\|_1 \le \sqrt{d}\|\overline{x}_n - \overline{x'}_n\|_2 \le 2\sqrt{d}r/n$ for neighboring samples. Then the Laplace mechanism (7.1.3) satisfies

$$\mathbb{E}[\|Z_{\rm Laplace} - \mathbb{E}[X]\|_2^2] = \mathbb{E}[\left\|\overline{X}_n - \mathbb{E}[X]\right\|_2^2] + \frac{8r^2}{n^2\varepsilon^2} \cdot d^2,$$

while the Gaussian mechanism under (ε^2, α)-Rényi privacy will yield

$$\mathbb{E}[\|Z_{\rm Gauss} - \mathbb{E}[X]\|_2^2] = \mathbb{E}[\left\|\overline{X}_n - \mathbb{E}[X]\right\|_2^2] + \frac{2r^2}{n^2\varepsilon^2} \cdot d\alpha.$$

This is evidently better than the Laplace mechanism whenever $\alpha < d$.

Proof of Proposition 7.2.5 We asume that P and Q have densities p and q with respect to a base measure μ, which is no loss of generality, whence the ratio condition implies that $e^{-\varepsilon} \le p/q \le e^{\varepsilon}$ and $D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\int (p/q)^\alpha q d\mu$. We prove the result assuming that $\alpha \in (1, \infty)$, as continuity gives the result for $\alpha \in \{1, \infty\}$.

First, it is clear that $D_\alpha(P\|Q) \le \varepsilon$ always. For the other term in the minimum, let us assume that $\alpha \le 1 + \frac{1}{\varepsilon}$ and $\varepsilon \le 1$. If either of these fails, the result is trivial, because for $\alpha > 1 + \frac{1}{\varepsilon}$ we have $\frac{3}{2}\alpha\varepsilon^2 \ge \frac{3}{2}\varepsilon \ge \varepsilon$, and similarly $\varepsilon \ge 1$ implies $\frac{3}{2}\alpha\varepsilon^2 \ge \varepsilon$.

Now we perform a Taylor approximation of $t \mapsto (1+t)^\alpha$. By Taylor's theorem, we have for any $t > -1$ that

$$(1+t)^\alpha = 1 + \alpha t + \frac{\alpha(\alpha-1)}{2}(1+\widetilde{t})^{\alpha-2}t^2$$

for some $\widetilde{t} \in [0,t]$ (or $[t,0]$ if $t < 0$). In particular, if $1 + t \le c$, then $(1+t)^\alpha \le 1 + \alpha t + \frac{\alpha(\alpha-1)}{2}\max\{1, c^{\alpha-2}\}t^2$. Now, we compute the divergence: we have

$$\begin{aligned}
&\exp\left((\alpha-1)D_\alpha(P\|Q)\right) = \int \left(\frac{p(z)}{q(z)}\right)^\alpha q(z)d\mu(z) \\
&= \int \left(1 + \frac{p(z)}{q(z)} - 1\right)^\alpha q(z)d\mu(z) \\
&\le 1 + \alpha \int \left(\frac{p(z)}{q(z)} - 1\right) q(z)d\mu(z) + \frac{\alpha(\alpha-1)}{2}\max\{1, \exp(\varepsilon(\alpha-2))\} \int \left(\frac{p(z)}{q(z)} - 1\right)^2 q(z)d\mu(z) \\
&\le 1 + \frac{\alpha(\alpha-1)}{2} e^{\varepsilon[\alpha-2]_+} \cdot (e^\varepsilon - 1)^2.
\end{aligned}$$

Now, we know that $\alpha - 2 \le 1/\varepsilon - 1$ by assumption, so using that $\log(1+x) \le x$, we obtain

$$D_\alpha(P\|Q) \le \frac{\alpha}{2}(e^\varepsilon - 1)^2 \cdot \exp([1-\varepsilon]_+).$$

Finally, a numerical calculation yields that this quantity is at most $\frac{3\alpha}{2}\varepsilon^2$ for $\varepsilon \le 1$. □

We can also provide connections from (ε, α)-Rényi privacy to (ε, δ)-differential privacy, and then from there to ε-differential privacy. We begin by showing how to develop (ε, δ)-differential privacy out of Rényi privacy. Another way to think about this proposition is that whenever two distributions P and Q are close in Rényi divergence, then there is some limited "amplification" of probabilities that is possible in moving from one to the other.

Proposition 7.2.7. *Let P and Q satisfy $D_\alpha(P\|Q) \le \varepsilon$. Then for any set A,*

$$P(A) \le \exp\left(\frac{\alpha-1}{\alpha}\varepsilon\right) Q(A)^{\frac{\alpha-1}{\alpha}}.$$

Consequently, for any $\delta > 0$,

$$P(A) \le \min\left\{\exp\left(\varepsilon + \frac{1}{\alpha-1}\log\frac{1}{\delta}\right)Q(A), \delta\right\} \le \exp\left(\varepsilon + \frac{1}{\alpha-1}\log\frac{1}{\delta}\right)Q(A) + \delta.$$

As above, we have an immediate corollary to this result.

Corollary 7.2.8. *Assume that M is (ε, α)-Rényi private. Then it is also $(\varepsilon + \frac{1}{\alpha-1}\log\frac{1}{\delta}, \delta)$-differentially private for any $\delta > 0$.*

Before turning to the proof of the proposition, we show how it can provide prototypical (ε, δ)-private mechanisms via Gaussian noise addition.

Example 7.2.9 (Gaussian mechanisms, continued)**:** Consider Example 7.2.3, where $f : \mathcal{P}_n \to \mathbb{R}^d$ has ℓ_2-sensitivity L. Then by Example 7.2.2, the Gaussian mechanism $Z = f(P_n) + W$ for $W \sim \mathsf{N}(0, \sigma^2 I)$ is $(\frac{\alpha L^2}{2\sigma^2}, \alpha)$-Rényi private for all $\alpha \geq 1$. Combining this with Corollary 7.2.8, the Gaussian mechanism is also

$$\left(\frac{\alpha L^2}{2\sigma^2} + \frac{1}{\alpha - 1} \log \frac{1}{\delta}, \delta\right) \text{-differentially private}$$

for any $\delta > 0$ and $\alpha > 1$. Optimizing first over α by taking $\alpha = 1 + \sqrt{2\sigma^2 \log \delta^{-1}/L^2}$, we see that the channel is $(\frac{L^2}{2\sigma^2} + \sqrt{2L^2 \log \delta^{-1}/\sigma^2}, \delta)$-differentially private. Thus we have that the Gaussian mechanism

$$Z = f(P_n) + W, \quad W \sim \mathsf{N}(0, \sigma^2 I) \text{ for } \sigma^2 = L^2 \max\left\{\frac{8 \log \frac{1}{\delta}}{\varepsilon^2}, \frac{1}{\varepsilon}\right\} \tag{7.2.5}$$

is (ε, δ)-differentially private.

To continue with our ℓ_2-bounded mean-estimation in Example 7.2.4, let us assume that $\varepsilon < 8 \log \frac{1}{\delta}$, in which case the Gaussian mechanism (7.2.5) with $L^2 = r^2/n^2$ achieves (ε, δ)-differential privacy, and we have

$$\mathbb{E}[\|Z_{\rm Gauss} - \mathbb{E}[X]\|_2^2] = \mathbb{E}[\left\|\overline{X}_n - \mathbb{E}[X]\right\|_2^2] + O(1)\frac{r^2}{n^2\varepsilon^2} \cdot d \log \frac{1}{\delta}.$$

Comparing to the previous cases, we see an improvement over the Laplace mechanism whenever $\log \frac{1}{\delta} \ll d$, or that $\delta \gg e^{-d}$. $\diamond$

Proof of Proposition 7.2.7 We use the data processing inequality of Proposition 7.2.1.iii, which shows that

$$\varepsilon \geq D_\alpha(P\|Q) \geq \frac{1}{\alpha - 1} \log \left[\left(\frac{P(A)}{Q(A)}\right)^\alpha Q(A)\right].$$

Rearranging and taking exponentials, we immediately obtain the first claim of the proposition.

For the second, we require a bit more work. First, let us assume that $Q(A) > e^{-\varepsilon}\delta^{\frac{\alpha}{\alpha-1}}$. Then we have by the first claim of the proposition that

$$\begin{aligned} P(A) &\leq \exp\left(\frac{\alpha - 1}{\alpha}\varepsilon + \frac{1}{\alpha} \log \frac{1}{Q(A)}\right) Q(A) \\ &\leq \exp\left(\frac{\alpha - 1}{\alpha}\varepsilon + \frac{1}{\alpha}\varepsilon + \frac{1}{\alpha - 1} \log \frac{1}{\delta}\right) Q(A) = \exp\left(\varepsilon + \frac{1}{\alpha - 1} \log \frac{1}{\delta}\right) Q(A). \end{aligned}$$

On the other hand, when $Q(A) \leq e^{-\varepsilon}\delta^{\frac{\alpha}{\alpha-1}}$, then again using the first result of the proposition,

$$\begin{aligned} P(A) &\leq \exp\left(\frac{\alpha - 1}{\alpha}(\varepsilon + \log Q(A))\right) \\ &\leq \exp\left(\frac{\alpha - 1}{\alpha}\left(\varepsilon - \varepsilon + \frac{\alpha}{\alpha - 1} \log \delta\right)\right) = \delta. \end{aligned}$$

This gives the second claim of the proposition. $\square$

Finally, we develop our last set of connections, which show how we may relate (ε, δ)-private channels with ε-private channels. To provide this definition, we require one additional weakened notion of divergence, which relates (ε, δ)-differential privacy to Rényi-α-divergence with $\alpha = \infty$. We define

$$D_\infty^\delta(P\|Q) := \sup_{S \subset \mathcal{X}} \left\{ \log \frac{P(S) - \delta}{Q(S)} \mid P(S) > \delta \right\},$$

where the supremum is over measurable sets. Evidently equivalent to this definition is that $D_\infty^\delta(P\|Q) \le \varepsilon$ if and only if

$$P(S) \le e^\varepsilon Q(S) + \delta \quad \text{for all } S \subset \mathcal{X}.$$

Then we have the following lemma.

Lemma 7.2.10. *Let $\varepsilon > 0$ and $\delta \in (0, 1)$, and let P and Q be distributions on a space $\mathcal{X}$.*

(i) We have $D_\infty^\delta(P\|Q) \le \varepsilon$ if and only if there exists a probability distribution R on $\mathcal{X}$ such that $\|P - R\|_{\rm TV} \le \delta$ and $D_\infty(R\|Q) \le \varepsilon$.

(ii) We have $D_\infty^\delta(P\|Q) \le \varepsilon$ and $D_\infty^\delta(Q\|P) \le \varepsilon$ if and only if there exist distributions P_0 and Q_0 such that

$$\|P - P_0\|_{\rm TV} \le \frac{\delta}{1 + e^\varepsilon}, \quad \|Q - Q_0\|_{\rm TV} \le \frac{\delta}{1 + e^\varepsilon},$$

and

$$D_\infty(P_0\|Q_0) \le \varepsilon \quad \text{and} \quad D_\infty(Q_0\|P_0) \le \varepsilon.$$

The proof of the lemma is technical, so we defer it to Section 7.5.1. The key application of the lemma—which we shall see presently—is that (ε, δ)-differentially private algorithms compose in elegant ways.

7.2.3 Side information protections under weakened notions of privacy

We briefly discuss the side information protections these weaker notions of privacy protect. For both (ε, δ)-differential privacy and (ε, α)-Rényi privacy, we revisit the treatment in Proposition 7.1.7, considering Bayes factors and ratios of prior and posterior divergences, as these are natural formulations of side information in terms of an adversary's probabilistic beliefs. Our first analogue of Proposition 7.1.7, applies to the (ε, δ)-private case.

Proposition 7.2.11. *Let M be a (ε, δ)-differentially private mechanism. Then for any neighboring $P_n, P_n', P_n^{(0)} \in \mathcal{P}_n$, we have with probability at least $1 - \delta$ over the draw of $Z = M(P_n^{(0)})$, the posterior odds satisfy*

$$\frac{\pi(P_n \mid z)}{\pi(P_n' \mid z)} \le e^{3\varepsilon} \frac{\pi(P_n)}{\pi(P_n')}.$$

Deferring the proof momentarily, this result shows that as long as two samples x, x' are neighboring, then an adversary is extremely unlikely to be able to glean substantially distinguishing information between the samples. This is suggestive of a heuristic in differential privacy that if n is the sample size, then one should take $\delta \ll 1/n$ to limit the probability of disclosure: by a union bound, we see that for each individual $i \in \{1, \ldots, n\}$, we can simultaneously guarantee that the posterior odds for swapping individual i's data do not change much (with high probability).

Unsurprisingly at this point, we can also give posterior update bounds for Rényi differential privacy. Here, instead of giving high-probability bounds—though it is possible—we can show that moments of the odds ratio do not change significantly. Indeed, we have the following proposition:

Proposition 7.2.12. *Let M be a (ε, α)-Rényi private mechanism, where $\alpha \in (1, \infty)$. Then for any neighboring $P_n, P_n', P_n^{(0)} \in \mathcal{P}_n$, we have*

$$\mathbb{E}_0\left[\left(\frac{\pi(P_n \mid Z)}{\pi(P_n' \mid Z)}\right)^{\alpha-1}\right]^{\frac{1}{\alpha-1}} \le e^{\varepsilon}\frac{\pi(P_n)}{\pi(P_n')},$$

where $\mathbb{E}_0$ denotes expectation taken over $Z = M(P_n^{(0)})$.

Proposition 7.2.12 communicates a similar message to our previous results in this vein: even if we get information from the output of the private mechanism on some sample $x_0 \in \mathcal{X}^n$ near the samples (datasets) of interest x, x' that an adversary wishes to distinguish, it is impossible to update beliefs by much. The parameter α then controls the degree of difficulty of this "impossible" claim, which one can see by (for example) applying a Chebyshev-type bound to the posterior ratio and prior ratios.

We now turn to the promised proofs of Propositions 7.2.11 and 7.2.12. To prove the former, we require a definition.

Definition 7.6. *Distributions P and Q on a space $\mathcal{X}$ are* (ε, δ)-close *if for all measurable A*

$$P(A) \le e^{\varepsilon}Q(A) + \delta \quad \text{and} \quad Q(A) \le e^{\varepsilon}P(A) + \delta.$$

Letting p and q denote their densities (with respect to any shared base measure), they are (ε, δ)-pointwise close *if the set*

$$A := \{x \in \mathcal{X} : e^{-\varepsilon}q(x) \le p(x) \le e^{\varepsilon}q(x)\} = \{x \in \mathcal{X} : e^{-\varepsilon}p(x) \le q(x) \le e^{\varepsilon}p(x)\}$$

satisfies $P(A) \ge 1 - \delta$ and $Q(A) \ge 1 - \delta$.

The following lemma shows the strong relationship between closeness and approximate differential privacy.

Lemma 7.2.13. *If P and Q are (ε, δ)-close, then for any $\beta > 0$, the sets*

$$A_+ := \{x : p(x) > e^{(1+\beta)\varepsilon}q(x)\} \quad \text{and} \quad A_- := \{x : p(x) \le e^{-(1+\beta)\varepsilon}q(x)\}$$

satisfy

$$\max\{P(A_+), Q(A_-)\} \le \frac{e^{\beta\varepsilon}\delta}{e^{\beta\varepsilon} - 1}, \quad \max\{P(A_-), Q(A_+)\} \le \frac{e^{-\varepsilon}\delta}{e^{\beta\varepsilon} - 1}.$$

Conversely, if P and Q are (ε, δ)-pointwise close, then

$$P(A) \le e^{\varepsilon}Q(A) + \delta \quad \text{and} \quad Q(A) \le e^{\varepsilon}P(A) + \delta$$

for all sets A.

Proof Let $A = A_+ = \{x : p(x) > e^{(1+\beta)\varepsilon}q(x)\}$. Then

$$P(A) \le e^{\varepsilon}Q(A) + \delta \le e^{-\beta\varepsilon}P(A) + \delta,$$

so that $P(A) \le \frac{\delta}{1 - e^{-\beta\varepsilon}}$. Similarly,

$$Q(A) \le e^{-(1+\beta)\varepsilon}P(A) \le e^{-\beta\varepsilon}Q(A) + e^{-(1+\beta)\varepsilon}\delta,$$

so that $Q(A) \le e^{-(1+\beta)\varepsilon}\delta/(1-e^{-\beta\varepsilon}) = e^{-\varepsilon}\delta/(e^{\beta\varepsilon}-1)$. The set A_- satisfies the symmetric properties.

For the converse result, let $B = \{x : e^{-\varepsilon}q(x) \le p(x) \le e^{\varepsilon}q(x)\}$. Then for any set A we have

$$P(A) = P(A \cap B) + P(A \cap B^c) \le e^{\varepsilon}Q(A \cap B) + \delta \le e^{\varepsilon}Q(A) + \delta,$$

and the same inequalities yield $Q(A) \le e^{\varepsilon}P(A) + \delta$. □

That is, (ε, δ)-close distributions are $(2\varepsilon, \frac{e^{\varepsilon}+e^{-\varepsilon}}{e^{\varepsilon}-1}\delta)$-pointwise close, and (ε, δ)-pointewise close distributions are (ε, δ)-close.

A minor extension of this lemma (taking $\beta = 1$ and applying the lemma twice) yields the following result.

Lemma 7.2.14. *Let P_0, P_1, P_2 be distributions on a space $\mathcal{X}$, each (ε, δ)-close. Then for any i, j, k, $j \ne k$, the set*

$$A_{jk} := \left\{x \in \mathcal{X} : \log \frac{p_j(x)}{p_k(x)} > 3\varepsilon\right\} \quad \textit{satisfies} \quad P_i(A_{jk}) \le C\delta \max\{\varepsilon^{-1}, 1\}$$

for a numerical constant $C \le 2$.

With Lemma 7.2.14 in hand, we can prove Proposition 7.2.11:

Proof of Proposition 7.2.11 Let $P_n^{(0)} \in \mathcal{P}_n$ denote the "true" sample. Consider the three channels Q_0, Q_1, Q_2, which represent the induced distributions of $M(P_n^{(0)}), M(P_n)$, and $M(P_n')$, respectively. Then by Lemma 7.2.14, with probability at least $1 - 2\delta\max\{\varepsilon^{-1}, 1\}$, $Z \sim Q_0$ belongs to the set $A = \{z \in \mathcal{Z} \mid e^{-3\varepsilon}q_1(z) \le q_2(z) \le e^{3\varepsilon}q_1(z)\}$. Calculating the odds ratios immediately gives the result. □

Finally, we provide the proof of Proposition 7.2.12.

Proof of Proposition 7.2.12 Let $r = \alpha - 1$ for shorthand, and let $p = \frac{\alpha}{r} = \frac{\alpha}{\alpha-1} > 1$ and $p_* = \frac{p}{p-1} = \alpha$ be its conjugate. As in the proof of Proposition 7.2.11, let Q_0, Q_1, and Q_2 represent the distributions of $Z = M(P_n^{(0)})$, $Z = M(P_n)$, and $Z = M(P_n')$, respectively. We apply Hölder's inequality: letting q_i be the density of Q_i with respect to some base measure $d\mu$—which we know must exist by definition of Rényi differential privacy—we have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\pi(P_n \mid Z)}{\pi(P_n' \mid Z)}\right)^r\right] &= \int \left(\frac{q_1(z)\pi(P_n)}{q_2(z)\pi(P_n')}\right)^r q_0(z)d\mu \\
&= \left(\frac{\pi(P_n)}{\pi(P_n')}\right)^r \int \left(\frac{q_1(z)}{q_2(z)}\right)^r \frac{q_0(z)}{q_2(z)} q_2(z)d\mu \\
&\le \left(\frac{\pi(P_n)}{\pi(P_n')}\right)^r \left(\int \left(\frac{q_1(z)}{q_2(z)}\right)^{pr} q_2(z)d\mu\right)^{\frac{1}{p}} \left(\int \left(\frac{q_0(z)}{q_2(z)}\right)^{p_*} q_2(z)d\mu\right)^{\frac{1}{p_*}} \\
&= \left(\frac{\pi(P_n)}{\pi(P_n')}\right)^r \exp\left(\frac{(\alpha-1)^2}{\alpha} D_\alpha(Q_1\|Q_2) + \frac{\alpha-1}{\alpha} D_\alpha(Q_0\|Q_2)\right) \\
&\le \left(\frac{\pi(P_n)}{\pi(P_n')}\right)^r \exp\left(\frac{(\alpha-1)^2 + \alpha - 1}{\alpha}\varepsilon\right)
\end{aligned}$$

as $pr = \alpha$ and $p_* = \alpha$. Taking everything to the $1/(\alpha-1)$ power and gives the result. □

7.3 Composition and privacy based on divergence

One of the major challenges in privacy is to understand what happens when a user participates in multiple studies, each providing different privacy guarantees. In this case, we might like to understand and control privacy losses even when the mechanisms for information release may depend on one another. Conveniently, all Rényi divergences provide strong guarantees on composition, essentially for free, and these then allow us to prove strong results on the composition of multiple private mechanisms.

7.3.1 Composition of Rényi-private channels

A natural idea to address composition is to attempt to generalize our chain rules for KL-divergence and related ideas to Rényi divergences. Unfortunately, this plan of attack does not quite work, as there is no generally accepted definition of a conditional Rényi divergence, and associated chain rules do not sum naturally. In situations in which individual divergence of associated elements of a joint distribution have bounded Rényi divergence, however, we can provide some natural bounds.

Indeed, consider the following essentially arbitrary scheme for data generation: we have distributions P and Q on a space $\mathcal{Z}^n$, where $Z_1^n \sim P$ and $Z_1^n \sim Q$ may exhibit arbitrary dependence. If, however, we can bound the conditional Rényi divergence between $P(Z_i \mid Z_1^{i-1})$ and $Q(Z_i \mid Z_1^{i-1})$, we can provide some natural tensorization guarantees. To set notation, let $P_i(\cdot \mid z_1^{i-1})$ be the the (regular) conditional probability of Z_i conditional on $Z_1^{i-1} = z_1^{i-1}$ under P, and similarly for Q_i. We have the following theorem.

Theorem 7.3.1. *Let the conditions above hold, $\varepsilon_i < \infty$ for $i = 1, \ldots, n$, and $\alpha \in [1, \infty]$. Assume that conditional on z_1^{i-1}, we have $D_\alpha\big(P_i(\cdot \mid z_1^{i-1}) \| Q_i(\cdot \mid z_1^{i-1})\big) \le \varepsilon_i$. Then*

$$D_\alpha(P\|Q) \le \sum_{i=1}^n \varepsilon_i.$$

Proof We assume without loss of generality that the conditional distributions $P_i(\cdot \mid z_1^{i-1})$ and Q_i are absolutely continuous with respect to a base measure μ on $\mathcal{Z}$.[1] Then we have

$$\begin{aligned}
D_\alpha(P\|Q) &= \frac{1}{\alpha - 1} \log \int \prod_{i=1}^n \left(\frac{p_i(z_i \mid z_1^{i-1})}{q_i(z_i \mid z_1^{i-1})}\right)^\alpha q_i(z_i \mid z_1^{i-1}) d\mu^n(z_1^n) \\
&= \frac{1}{\alpha - 1} \log \int_{\mathcal{Z}_1^{n-1}} \left[\int \left(\frac{p_n(z_n \mid z_1^{n-1})}{q_n(z_n \mid z_1^{n-1})}\right)^\alpha q_n(z_n \mid z_1^{n-1}) d\mu(z_n)\right] \prod_{i=1}^{n-1} \left(\frac{p_i}{q_i}\right)^\alpha q_i d\mu^{n-1} \\
&\le \frac{1}{\alpha - 1} \log \int_{\mathcal{Z}_1^{n-1}} \exp((\alpha - 1)\varepsilon_n) \prod_{i=1}^{n-1} \left(\frac{p_i(z_i \mid z_1^{i-1})}{q_i(z_i \mid z_1^{i-1})}\right)^\alpha q_i(z_i \mid z_1^{i-1}) d\mu^{n-1}(z_1^{n-1}) \\
&= \varepsilon_n + D_\alpha\big(P_1^{n-1} \| Q_1^{n-1}\big).
\end{aligned}$$

Applying the obvious inductive argument then gives the result. □

[1] This is no loss of generality, as the general definition of f-divergences as suprema over finite partitions, or quantizations, of each X_i and Y_i separately, as in our discussion of KL-divergence in Chapter 2.2.2. Thus we may assume $\mathcal{Z}$ is discrete and μ is a counting measure.

7.3.2 Privacy games and composition

To understand arbitrary composition of private channels, let us consider a privacy "game," where an adversary may sequentially choose a dataset—in an arbitrary way—and then observes a private release Z_i of some mechanism applied to the dataset and the dataset with one entry (observation) modified. The adversary may then select a new dataset, and repeat the game. We then ask whether the resulting sequence of (private) observations Z_1^k remains private. Figure 7.1 captures this in an algorithmic form. Letting $Z_i^{(b)}$ denote the random observations under the bit $b \in \{0, 1\}$, whether

Input: Family of channels $\mathcal{Q}$ and bit $b \in \{0, 1\}$.
Repeat: for $k = 1, 2, \ldots$

i. Adversary chooses arbitrary space $\mathcal{X}$, $n \in \mathbb{N}$, and two datasets $x^{(0)}, x^{(1)} \in \mathcal{X}^n$ with $d_{\rm ham}(x^{(0)}, x^{(1)}) \le 1$.

ii. Adversary chooses private channel $Q_k \in \mathcal{Q}$.

iii. Adversary observes one sample $Z_k \sim Q_k(\cdot \mid x^{(b)})$.

Figure 7.1. The privacy game. In this game, the adversary may *not* directly observe the private $b \in \{0, 1\}$.

the distributions of $(Z_1^{(0)}, \ldots, Z_k^{(0)})$ and $(Z_1^{(1)}, \ldots, Z_k^{(1)})$ are substantially different. Note that, in the game in Fig. 7.1, the adversary may track everything, and even chooses the mechanisms Q_k.

Now, let $Z^{(0)} = (Z_1^{(0)}, \ldots, Z_k^{(0)})$ and $Z^{(1)} = (Z_1^{(1)}, \ldots, Z_k^{(1)})$ be the outputs of the privacy game above, and let their respective marginal distributions be $Q^{(0)}$ and $Q^{(1)}$. We then make the following definition.

Definition 7.7. *Let $\varepsilon \ge 0$, $\alpha \in [1, \infty]$, and $k \in \mathbb{N}$.*

(i) A collection $\mathcal{Q}$ of channels satisfies (ε, α)-Rényi privacy under k-fold adaptive composition *if, in the privacy game in Figure 7.1, the distributions $Q^{(0)}$ and $Q^{(1)}$ on $Z^{(0)}$ and $Z^{(1)}$, respectively, satisfy $D_\alpha\big(Q^{(0)} \| Q^{(1)}\big) \le \varepsilon$ and $D_\alpha\big(Q^{(1)} \| Q^{(0)}\big) \le \varepsilon$.*

(ii) Let $\delta > 0$. Then a collection $\mathcal{Q}$ of channels satisfies (ε, δ)-differential privacy under k-fold adaptive composition *if $D_\infty^\delta(Q^{(0)} \| Q^{(1)}) \le \varepsilon$ and $D_\infty^\delta(Q^{(1)} \| Q^{(0)}) \le \varepsilon$.*

By considering a special case centered around a particular individual in the game 7.1, we can gain some intuition for the definition. Indeed, suppose that an individual has some data x_0; in each round of the game the adversary generates two datasets, one containing x_0 and the other identical except that x_0 is removed. Then satisfying Definition 7.7 captures the intuition that an individual's privacy remains protected, even in the face of multiple (private) accesses of the individual's data.

As an immediate corollary to Theorem 7.3.1, we then have the following.

Corollary 7.3.2. *Assume that each channel in the game in Fig. 7.1 is (ε_i, α)-Rényi private. Then the arbitrary composition of k such channels remains $(\sum_{i=1}^k \varepsilon_i, \alpha)$-Rényi private.*

More sophisticated corollaries are possible once we start to use the connections between privacy measures we outline in Section 7.2.2. In this case, we can develop so-called *advanced composition* rules, which sometimes suggest that privacy degrades more slowly than might be expected under adaptive composition.

Corollary 7.3.3. *Assume that each channel in the game in Fig. 7.1 is ε-differentially private. Then the composition of k such channels is $k\varepsilon$-differentially private. Additionally, the composition of k such channels is*

$$\left(\frac{3k}{2}\varepsilon^2 + \sqrt{6k\log\frac{1}{\delta}} \cdot \varepsilon, \delta\right)$$

differentially private for all $\delta > 0$.

Proof The first claim is immediate: for $Q^{(0)}, Q^{(1)}$ as in Definition 7.7, we know that $D_\alpha\big(Q^{(0)} \| Q^{(1)}\big) \le k\varepsilon$ for all $\alpha \in [1, \infty]$ by Theorem 7.3.1 coupled with Proposition 7.2.5 (or Corollary 7.2.6).

For the second claim, we require a bit more work. Here, we use the bound $\frac{3\alpha}{2}\varepsilon^2$ in the Rényi privacy bound in Corollary 7.2.6. Then we have for any $\alpha \ge 1$ that

$$D_\alpha\Big(Q^{(0)} \| Q^{(1)}\Big) \le \frac{3k\alpha}{2}\varepsilon^2$$

by Theorem 7.3.1. Now we apply Proposition 7.2.7 and Corollary 7.2.8, which allow us to conclude (ε, δ)-differential privacy from Rényi privacy. Indeed, by the preceding display, setting $\eta = 1 + \alpha$, we have that the composition is $(\frac{3k}{2}\varepsilon^2 + \frac{3k\eta}{2}\varepsilon^2 + \frac{1}{\eta}\log\frac{1}{\delta}, \delta)$-differentially private for all $\eta > 0$ and $\delta > 0$. Optimizing over η gives the second result. □

We note in passing that it is possible to get slightly sharper results than those in Corollary 7.3.3; indeed, using ideas from Exercise 4.3 it is possible to achieve $(k\varepsilon(e^\varepsilon - 1) + \sqrt{2k\log\frac{1}{\delta}}\varepsilon, \delta)$-differential privacy under adaptive composition.

A more sophisticated result, which shows adaptive composition for (ε, δ)-differentially private channels, is also possible using Lemma 7.2.10.

Theorem 7.3.4. *Assume that each channel in the game in Fig. 7.1 is (ε, δ)-differentially private. Then the composition of k such channels is $(k\varepsilon, k\delta)$-differentially private. Additionally, they are*

$$\left(\frac{3k}{2}\varepsilon^2 + \sqrt{6k\log\frac{1}{\delta_0}} \cdot \varepsilon, \delta_0 + \frac{k\delta}{1 + e^\varepsilon}\right)$$

differentially private for all $\delta_0 > 0$.

Proof Consider the channels Q_i in Fig. 7.1. As each satisfies $D_\infty^\delta(Q_i(\cdot \mid x^{(0)}) \| Q_i(\cdot \mid x^{(1)})) \le \varepsilon$ and $D_\infty^\delta(Q_i(\cdot \mid x^{(1)}) \| Q_i(\cdot \mid x^{(0)})) \le \varepsilon$, Lemma 7.2.10 guarantees the existence (at each sequential step, which may depend on the preceding $i-1$ outputs) of probability measures $Q_i^{(0)}$ and $Q_i^{(1)}$ such that $D_\infty(Q_i^{(1-b)} \| Q_i^{(b)}) \le \varepsilon$, $\|Q_i^{(b)} - Q_i(\cdot \mid x^{(b)})\|_{\rm TV} \le \delta/(1 + e^\varepsilon)$ for $b \in \{0, 1\}$.

Note that by construction (and Theorem 7.3.1) we have $D_\alpha(Q_1^{(b)} \cdots Q_k^{(b)} \| Q_1^{(1-b)} \cdots Q_k^{(1-b)}) \le \min\{\frac{3k\alpha}{2}\varepsilon^2, k\varepsilon\}$, where $Q^{(b)}$ denotes the joint distribution on $Z_1, \ldots, Z_k$ under bit b. We also have by the triangle inequality that $\|Q_1^{(b)} \cdots Q_k^{(b)} - Q^{(b)}\|_{\rm TV} \le k\delta/(1 + e^\varepsilon)$ for $b \in \{0, 1\}$. (See Exercise 2.16.) As a consequence, we see as in the proof of Corollary 7.3.3 that the composition is $(\frac{3k}{2}\varepsilon^2 + \frac{3k\eta}{2}\varepsilon^2 + \frac{1}{\eta}\log\frac{1}{\delta_0}, \delta_0 + k\delta/(1 + e^\varepsilon))$-differentially private for all $\eta > 0$ and δ_0. Optimizing gives the result. □

As a consequence of these results, we see that whenever the privacy parameter $\varepsilon < 1$, it is possible to compose multiple privacy mechanisms together and have privacy penalty scaling only as the worse of $\sqrt{k}\varepsilon$ and $k\varepsilon^2$, which is substantially better than the "naive" bound of $k\varepsilon$. Of course, a challenge here—relatively unfrequently discussed in the privacy literature—is that when $\varepsilon \geq 1$, which is a frequent case for practical deployments of privacy, all of these bounds are much worse than a naive bound that k-fold composition of ε-differentially private algorithms is $k\varepsilon$-differentially private.

7.4 Additional mechanisms and privacy-preserving algorithms

Since the introduction of differential privacy, a substantial literature has grown providing mechanisms for different estimation, learning, and data release problems. Here, we describe a few of those beyond the basic noise addition schemes we have thus far developed, highlighting a few applications along the way. One major challenge with the naive approaches is that they rely on *global* sensitivity of the functions to be estimated, rather than local sensitivities—a worst case notion that sometimes forces privacy to add unnecessary noise. In Section 7.4.2, we give one potential approach to this problem, which we develop further in exercises and revisit in optimality guarantees in sequential chapters. Our view is necessarily somewhat narrow, but the results here can form a natural starting point for further work in this area.

7.4.1 The exponential mechanism

In many statistical, learning, and other problems, there is a natural notion of loss (or conversely, utility) in releasing a potentially noisy result of some computation. We abstract this by considering the input space $\mathcal{P}_n$ of samples of size n (that is, empirical distributions) and output space $\mathcal{Z}$ along with a loss function $\ell : \mathcal{P}_n \times \mathcal{Z} \to \mathbb{R}$, where $\ell(P_n, z)$ measures the loss of z on an input $P_n \in \mathcal{P}_n$. For example, if we wish to compute a function $f : \mathcal{P}_n \to \mathbb{R}$, a natural notion of loss is $\ell(P_n, z) = |f(P_n) - z|$ for $z \in \mathbb{R}$. As a more sophisticated and somewhat abstract formulation, suppose we wish to release a sample distribution $\widetilde{P}$ approximating an input sample $P_n \in \mathcal{P}_n$, where we wish $\widetilde{P}$ to be accurate for most statistical queries in some family, that is, $\frac{1}{n}\sum_{i=1}^n \phi(x_i) \approx \mathbb{E}_{\widetilde{P}}[\phi(X)]$ for all $\phi \in \Phi$. Then a natural loss is $\ell(P_n, \widetilde{P}) = \sup_{\phi\in\Phi} |\mathbb{E}_{P_n}\phi(X) - \mathbb{E}_{\widetilde{P}}[\phi(X)]|$.

In scenarios in which we have such a loss, the abstract *exponential mechanism* provides an attractive approach. We assume that for each $z \in \mathcal{Z}$, the loss $\ell(\cdot, z)$ has (global) sensitivity L, i.e., $|\ell(P_n, z) - \ell(P_n', z)| \leq L$ for all neighboring $P_n, P_n' \in \mathcal{P}_n$. We assume we have a base measure μ on $\mathcal{Z}$, and then define the exponential mechanism by

$$\mathbb{P}(M(P_n) \in A) = \frac{1}{\int \exp(-\frac{\varepsilon}{L}\ell(P_n, z))d\mu(z)} \int_A \exp\left(-\frac{\varepsilon}{L}\ell(P_n, z)\right) d\mu(z), \tag{7.4.1}$$

assuming $\int e^{-\frac{\varepsilon}{L}\ell(x,z)}d\mu(z)$ is finite for each $P_n \in \mathcal{P}_n$. (Typically, one assumes ℓ takes on values in $\mathbb{R}_+$ and μ is a finite measure, making the last assumption trivial.) That is, the exponential mechanism M releases $Z = M(P_n)$ with probability proportional to

$$\exp\left(-\frac{\varepsilon}{L}\ell(P_n, z)\right).$$

That the mechanism (7.4.1) is 2ε-differentially private is immediate: for any neighboring P_n, P_n',

we have

$$\frac{Q(A \mid P_n)}{Q(A \mid P_n')} = \frac{\int \exp(-\frac{\varepsilon}{L}\ell(P_n', z))d\mu(z) \int_A \exp(-\frac{\varepsilon}{L}\ell(P_n, z))d\mu(z)}{\int \exp(-\frac{\varepsilon}{L}\ell(P_n, z))d\mu(z) \int_A \exp(-\frac{\varepsilon}{L}\ell(P_n', z))d\mu(z)}$$
$$\le \sup_{z \in \mathcal{Z}} \left\{ \exp\left(\frac{\varepsilon}{L}[\ell(P_n, z) - \ell(P_n', z)]\right)\right\} \cdot \sup_{z \in A} \left\{ \exp\left(\frac{\varepsilon}{L}[\ell(P_n', z) - \ell(P_n, z)]\right)\right\} \le \exp(2\varepsilon).$$

As a first (somewhat trivial) example, we can recover the Laplace mechanism:

Example 7.4.1 (The Laplace mechanism)**:** We can recover Example 7.1.3 through the exponential mechanism. Indeed, suppose that we wish to release $f : \mathcal{P}_n \to \mathbb{R}^d$, where $\mathrm{GS}_1(f) \le L$. Then taking $z \in \mathbb{R}^d$, $\ell(P_n, z) = \|f(P_n) - z\|_1$, and μ to be the usual Lebesgue measure on $\mathbb{R}^d$, the exponential mechanism simply uses density

$$q(z \mid P_n) \propto \exp\left(-\frac{\varepsilon}{L}\|f(P_n) - z\|_1\right),$$

which is the Laplace mechanism. $\diamond$

One challenge with the exponential mechanism (7.4.1) is that it is somewhat abstract and is often hard to compute, as it requires evaluating an often high-dimensional integral to sample from. Yet it provides a nice abstract mechanism with strong privacy guarantees and, as we shall see, good utility guarantees. For the moment, we defer further examples and provide utility guarantees when $\mu(\mathcal{Z})$ is finite, giving bounds based on the measure of "bad" solutions. For notational convenience, we define the optimal value

$$\ell^\star(P_n) = \inf_{z \in \mathcal{Z}} \ell(P_n, z),$$

assuming tacitly that it is finite, and the sublevel sets

$$S_t := \{z \in \mathcal{Z} \mid \ell(P_n, z) \le \ell^\star(P_n) + t\}.$$

With these definitions, we have the following proposition.

Proposition 7.4.2. *Let $t \ge 0$. Then for the exponential mechanism (7.4.1), if $Z \sim Q(\cdot \mid P_n)$ then*

$$\ell(P_n, Z) \le \ell^\star(P_n) + 2t$$

with probability at least $1 - \exp\left(-\frac{\varepsilon t}{L} + \log\frac{\mu(\mathcal{Z})}{\mu(S_t)}\right)$.

Proof Assume without loss of generality (by scaling) that the global Lipschitzian (sensitivity) constant of ℓ is $L = 1$. Then for $Z \sim Q(\cdot \mid P_n)$, we have

$$\mathbb{P}\left(\ell(P_n, Z) \ge \ell^\star(P_n) + 2t\right) = \frac{\int_{S_{2t}^c} \exp(-\varepsilon\ell(P_n, z))d\mu(z)}{\int \exp(-\varepsilon\ell(P_n, z))d\mu(z)} = \frac{\int_{S_{2t}^c} \exp(-\varepsilon(\ell(P_n, z) - \ell^\star(P_n)))d\mu(z)}{\int \exp(-\varepsilon(\ell(P_n, z) - \ell^\star(P_n)))d\mu(z)}$$
$$\le \frac{\int_{S_{2t}^c} \exp(-2\varepsilon t)d\mu(z)}{\int_{S_t} \exp(-\varepsilon(\ell(P_n, z) - \ell^\star(P_n)))d\mu(z)} \le \exp(-\varepsilon t)\frac{\mu(S_{2t}^c)}{\mu(S_t)},$$

where the last inequality uses that $\ell(P_n, z) - \ell^\star(P_n) \le t$ on S_t. $\square$

We can provide a few simplifications of this result in different special cases. For example, if $\mathcal{Z}$ is finite with cardinality $\mathrm{card}(\mathcal{Z})$, then Proposition 7.4.2 implies that taking μ to be the counting measure on $\mathcal{Z}$ we have

Corollary 7.4.3. *In addition to the conditions in Proposition 7.4.2, assume that* $\mathrm{card}(\mathcal{Z})$ *is finite. Then for any* $u \in (0,1)$*, with probability at least* $1-u$*,*

$$\ell(P_n, Z) \le \ell^\star(P_n) + \frac{2L}{\varepsilon} \log \frac{\mathrm{card}(\mathcal{Z})}{u}.$$

That is, with extremely high probability, the loss of Z from the exponential mechanism is at most logarithmic in $\mathrm{card}(\mathcal{Z})$ and grows only linearly with the global sensitivity L.

A second corollary allows us to bound the expected loss of the exponential mechanism, assuming we have some control over the measure of the sublevel sets S_t.

Corollary 7.4.4. *Let* $t \ge 0$ *be the smallest scalar such that* $t \ge \frac{2L}{\varepsilon} \log \frac{\mu(\mathcal{Z})}{\mu(S_t)}$ *and* $t \ge \frac{L}{\varepsilon}$*. Then* Z *drawn from the exponential mechanism* (7.4.1) *satisfies*

$$\mathbb{E}[\ell(P_n, Z)] \le \ell^\star(P_n) + t + \frac{2L}{\varepsilon} \le \ell^\star(P_n) + 3t \le \ell^\star(P_n) + O(1)\frac{L}{\varepsilon} \log\left(1 + \frac{\mu(\mathcal{Z})}{\mu(S_t)}\right).$$

Proof We first recall that if $W \ge 0$ is a nonnegative random variable, then by a change of variables, $\mathbb{E}[W] = \int_0^\infty \mathbb{P}(W \ge t)dt$. Take $\ell(P_n, Z) - \ell^\star(P_n) \ge 0$ as our random variable, fix any $t_0 \ge 0$, and let $\rho = \log \frac{\mu(\mathcal{Z})}{\mu(S_{t_0})}$. Then by Proposition 7.4.2 we have

$$\begin{aligned}
\mathbb{E}[\ell(P_n, Z) - \ell^\star(P_n)] &\le t_0 + \int_{t_0}^\infty \mathbb{P}(\ell(P_n, Z) - \ell^\star(P_n) \ge t)dt \\
&= t_0 + 2\int_{t_0/2}^\infty \mathbb{P}(\ell(P_n, Z) - \ell^\star(P_n) \ge 2t)dt \\
&\le t_0 + 2\int_{t_0/2}^\infty \exp\left(-\frac{\varepsilon t}{L} + \log \frac{\mu(\mathcal{Z})}{\mu(S_t)}\right) dt \\
&\le t_0 + 2e^\rho \int_{t_0/2}^\infty \exp\left(-\frac{\varepsilon t}{L}\right) dt = t_0 + \frac{2L}{\varepsilon} \exp\left(\rho - \frac{\varepsilon t_0}{2L}\right).
\end{aligned}$$

Take t_0 as in the statement of the corollary to obtain the result. □

Corollary 7.4.4 may seem a bit circular: we require the ratio $\mu(\mathcal{Z})/\mu(S_t)$ to be controlled—but it is relatively straightforward to use it (and Proposition 7.4.2) with a bit of care and standard bounds on volumes.

Example 7.4.5 (Empirical risk minimization via the exponential mechanism)**:** We consider the empirical risk minimization problem, where we have losses $\ell : \Theta \times \mathcal{X} \to \mathbb{R}_+$, where $\Theta \subset \mathbb{R}^d$ is a parameter space of interest, and we wish to choose

$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ L(\theta, P_n) := \frac{1}{n} \sum_{i=1}^n \ell(\theta, x_i) \right\}$$

where $P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i}$. We make a few standard assumptions: first, for simplicity, that n is large enough that $\frac{n}{d} \ge \varepsilon$. We also assume that $\Theta \subset \mathbb{R}^d$ is an ℓ_2-ball of radius R, that $\theta \mapsto \ell(\theta, x_i)$ is M-Lipschitz for all x_i, and that $\ell(\theta, x_i) \in [0, 2MR]$ for all $\theta \in \Theta$. (Note that this last is no loss of generality, as $\ell(\theta, x_i) - \inf_{\theta \in \Theta} \ell(\theta, x_i) \le M \sup_{\theta, \theta' \in \Theta} \|\theta - \theta'\|_2 \le 2MR$.)

Take the empirical loss $L(\theta, P_n)$ as our criterion function for the exponential mechanism, which evidently satisfies $|L(\theta, P_n) - L(\theta, P_n')| \le \frac{2MR}{n}$ whenever $d_{\rm ham}(P_n, P_n') \le 1$, so that we release θ with density

$$q(\theta \mid x) \propto \exp\left(-\frac{n\varepsilon}{2MR} L(\theta, P_n)\right).$$

Let $\widehat{\theta}_n$ be the empirical minimizer as above; then by the Lipschitz continuity of ℓ, the sublevel set S_t evidently satisfies

$$S_t \supset \left\{\theta \in \Theta \mid \|\theta - \widehat{\theta}_n\|_2 \le \frac{t}{M}\right\}.$$

Then a volume calculation (with the factor of 2 necessary because we may have $\widehat{\theta}_n$ on the boundary of Θ) yields that for μ the Lebesgue measure,

$$\frac{\mu(S_t)}{\mu(\mathcal{Z})} \ge \left(\frac{t}{2MR}\right)^d.$$

As a consequence, by Corollary 7.4.4, whenever $t \ge O(1)\frac{MR}{n\varepsilon} \cdot d \log \frac{MR}{t}$, we have $\mathbb{E}[L(\theta, P_n) \mid P_n] \le L(\widehat{\theta}_n, P_n) + 3t$. The choice $t = O(1)\frac{MRd}{n\varepsilon}$ suffices whenever $\frac{\varepsilon}{d} \le 1$, so we obtain

$$\mathbb{E}[L(\theta, P_n)] \le L(\widehat{\theta}_n, P_n) + O(1)\frac{MRd}{n\varepsilon} \log \frac{n\varepsilon}{d},$$

whenever $\frac{d}{n\varepsilon} \le 1$. Notably, standard empirical risk minimization (recall Chapter 4.4) typically achieves rates of convergence roughly of $MR/\sqrt{n}$, so that the gap of the exponential mechanism is lower order whenever $\frac{d}{\sqrt{n}\varepsilon} \le 1$. $\diamond$

7.4.2 Local sensitivities and the inverse sensitivity mechanism

A particular choice of the exponential mechanism (7.4.1) can provide strong optimality guarantees for 1-dimensional quantities, and appears to be the "right" mechanism (in principle) when one wishes to estimate a scalar-valued functional $f(P_n)$. A better (in principle) algorithm than noise addition schemes using the global sensitivity $\mathrm{GS}(f) = \sup |f(P_n) - f(P_n')|$ is to use a *local* notion of sensitivity: we are only concerned with adding noise commensurate with the changes of f near $P_n \in \mathcal{P}_n$. With this in mind, define the *modulus of continuity* of f at P_n by

$$\omega_f(k; P_n) := \sup\left\{|f(P_n') - f(P_n)| \mid d_{\rm ham}(P_n, P_n') \le k\right\},$$

which measures the amount that changing k observations in P_n can change the function f. In the privacy literature, the particular choice $k = 1$ yields the *local sensitivity*

$$\mathrm{LS}(f, P_n) := \sup\left\{|f(P_n) - f(P_n')| \mid d_{\rm ham}(P_n', P_n) = 1\right\} = \omega_f(1; P_n). \tag{7.4.2}$$

A naive strategy, then, would be to release

$$Z = f(P_n) + \frac{\mathrm{LS}(f, P_n)}{\varepsilon} \cdot W \quad \text{for } W \sim \mathsf{Laplace}(1),$$

which is analogous to the Laplace mechanism (7.1.3), except that the noise scales with the local sensitivity of f at P_n. The issue, as the next example makes clear, is that the scale of this noise can compromise privacy.

Example 7.4.6 (The sensitivity of the sensitivity)**:** Consider estimating a median $f(P_n) = \mathrm{med}(P_n)$, where the data $x \in [0,1]$, where $n = 2m+1$ for simplicity, to make the median unique. If the sample consists of m points $x_i = 0$ and $m+1$ points $x_i = 1$, then the sensitivity $\omega_f(1, P_n) = 1$, the maximal value—we simply move one example from $x_i = 1$ to $x_i = 0$, changing the median from $\mathrm{med}(P_n) = 1$ to 0. On the other hand, on the sample P_n' with $m-1$ points $x_i = 0$ and $m+2$ points $x_i = 1$, the sensitivity $\omega_f(1, P_n') = 0$, because changing a single example cannot move the median from $f(P_n') = 1$. $\diamond$

Instead of using the inherently unstable quantity ω, then, we can instead use, essentially, its inverse: define the *inverse sensitivity*

$$d_f(t, P_n) := \inf \left\{ d_{\mathrm{ham}}(P_n', P_n) \mid f(P_n') = t \right\}, \tag{7.4.3}$$

where $d_f(t, P_n) = +\infty$ if no P_n' yields $f(P_n') = t$. So $d_f(t, P_n)$ counts the number of examples that must be changed in the sample P_n to move $f(P_n)$ to a target t, and by inspection, always satisfies

$$|d_f(t, P_n) - d_f(t, P_n')| \le d_{\mathrm{ham}}(P_n, P_n').$$

Then the *inverse sensitivity mechanism* releases a value t with probability density proportional to

$$q(t \mid P_n) \propto \exp\left(-\frac{\varepsilon}{2} d_f(t, P_n)\right). \tag{7.4.4}$$

Implicit in the definition (7.4.4) is a base measure μ, typically one of Lebesgue measure or counting measure on a discrete set. Then a quick calculation (or recognition that the density (7.4.4) is a particular instance of the exponential mechanism) gives the following proposition.

Proposition 7.4.7. *Let M be the inverse sensitivity mechanism with density (7.4.4). Then M is ε-differentially private.*

As in the general exponential mechanism (7.4.1), efficiently sampling from the density (7.4.4) can be challenging. Some cases admit easier reformulations.

Example 7.4.8 (Mean estimation with bounded data)**:** Suppose the data $x \in [a, b]$ are bounded and we wish to estimate the sample mean $f(P_n) = \mathbb{E}_{P_n}[X] = \overline{x}_n$, where $P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i}$. Changing a single observation can move the mean by at most $\frac{b-a}{n}$ (replace $x_i = a$ with $x_i' = b$). Thus, while discretization issues and that we may have $x_i \notin \{a, b\}$ make precisely computing d_f tedious, the approximation

$$d_{\mathrm{mean}}(t, P_n) = \left\lceil \frac{n|t - \overline{x}_n|}{b-a} \right\rceil,$$

where we define $d_{\mathrm{mean}}(t, P_n) = +\infty$ for $t \notin [a, b]$, is both Lipschitz (with respect to the Hamming metric) in the sample P_n, and approximates $d_f(t, P_n)$. (See Exercise 7.8 for a more general approach justifying this particular approximation.) The approximation

$$q(t \mid P_n) := \frac{\exp(-\frac{\varepsilon}{2} d_{\mathrm{mean}}(t, P_n))}{\int_a^b \exp(-\frac{\varepsilon}{2} d_{\mathrm{mean}}(s, P_n)) ds} \tag{7.4.5}$$

to the density (7.4.4) is thus ε-differentially private,

The density (7.4.5) yields a particular step-like density. Define the shells

$$S_k = \left\{ \left[\overline{x}_n - k\frac{b-a}{n}, \overline{x}_n - (k-1)\frac{b-a}{n} \right] \cup \left[\overline{x}_n + (k-1)\frac{b-a}{n}, \overline{x}_n + k\frac{b-a}{n} \right] \right\} \cap [a,b]$$

corresponding to the amount the mean may change if we modify k examples and let $\mathrm{Vol}(S_k)$ be volume (length) of the intervals making up S_k. To sample from the density (7.4.5), note that the denominator $C(P_n) := \int_a^b \exp(-\frac{\varepsilon}{2} d_{\mathrm{mean}}(s, P_n))ds = \sum_{k=1}^n \mathrm{Vol}(S_k) e^{-\frac{k\varepsilon}{2}}$. Then we draw an index $I \in [n]$ with probability $\mathbb{P}(I = k) = \mathrm{Vol}(S_k)e^{-\varepsilon k/2}/C(P_n)$, and then choose t uniformly at random within S_k. $\diamond$

Example 7.4.9 (Median estimation)**:** For the median, the inverse sensitivity takes a particularly clean form, making sampling from the density (7.4.4) fairly straightforward. In this case, for a sample $P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i}$, where $x_i \in \mathbb{R}$, we have

$$d_f(t, P_n) = \mathrm{card}\left\{ i \in [n] \mid x_i \in [f(P_n), t] \right\},$$

the number of examples between the median $f(P_n)$ and putative target t. If the data lie in a range $x \in [a,b]$, then the density q is relatively straightforward to compute. Similar to the approach to the stepped density in Example 7.4.8, divide $[a,b]$ into the intervals

$$S_k^- := [a_k^-, a_{k-1}^-] \quad \text{and} \quad S_k^+ := [a_{k-1}^+, a_k^+], \quad k = 1, \ldots, n/2,$$

where

$$a_k^- = \inf\left\{ f(P_n') \mid d_{\mathrm{ham}}(P_n', P_n) \le k \right\} \quad \text{and} \quad a_k^+ = \sup\left\{ f(P_n') \mid d_{\mathrm{ham}}(P_n', P_n) \le k \right\}.$$

That is, a_k^- is the smallest we can make the median by changing k examples and a_k^+ the largest, corresponding to the the $\frac{1}{2} - \frac{k}{n}$ and $\frac{1}{2} + \frac{k}{n}$ quantiles of the sample P_n, where the 0 quantile is a and 1 quantile is b. Then defining the normalization constant

$$C(P_n) := \int_a^b \exp\left(-\frac{\varepsilon}{2} d_f(t, P_n)\right) dt = \sum_{k=1}^n \mathrm{Vol}(S_k^- \cup S_k^+) \exp\left(-\frac{\varepsilon}{2}k\right)$$

(where the volume is simply interval length), we may sample from the density (7.4.4) by first drawing a random index $I \in \{1, \ldots, n\}$ with probability proportional to

$$\mathbb{P}(I = k \mid P_n) = \frac{\mathrm{Vol}(S_k^- \cup S_k^+)}{C(P_n)} \exp\left(-\frac{\varepsilon}{2}k\right),$$

then drawing t uniformly at random in the each of the intervals S_k^- or S_k^+ with probabilities $\mathrm{Vol}(S_k^-)/\mathrm{Vol}(S_k^- \cup S_k^+)$ or $\mathrm{Vol}(S_k^+)/\mathrm{Vol}(S_k^- \cup S_k^+)$, respectively. $\diamond$

The particular sampling strategies—where we construct concentric shells S_k around $f(P_n)$ and sample from these with geometrically decaying probabilities $e^{-k\varepsilon/2}$—point toward more general sampling stratgies and optimality guarantees for the inverse sensitivity mechanism. Define the "shells"

$$S_k := \{ f(P_n') \mid d_{\mathrm{ham}}(P_n, P_n') = k \}.$$

We focus on sampling from the density (7.4.4) in the case $t \in \mathbb{R}$, so sampling is equivalent to drawing an index $I \in [n]$ with probability

$$\mathbb{P}(I = k \mid P_n) = \frac{1}{C(P_n)} e^{-\frac{\varepsilon}{2}k} \quad \text{for} \quad C(P_n) := \sum_{k=1}^{n} \mathrm{Vol}(S_k) e^{-\frac{\varepsilon}{2}k}, \tag{7.4.6}$$

then choosing t uniformly at random in S_k.

Define the shorthand $\omega(k) = \omega_f(k, P_n)$. Then the values $t \in S_k$ all satisfy $|f(P_n) - t| \le \omega(k)$, and so the inverse sensitivity mechanism M guarantees

$$\mathbb{E}[|M(P_n) - f(P_n)|] \le \sum_{k=1}^{n} \mathbb{P}(M(P_n) \in S_k) \omega(k).$$

Now our calculations become heuristic, where we make an effort to give the rough flavor of results possible, and later apply the care necessary for tighter guarantees. Suppose that the interval lengths $\mathrm{Vol}(S_k)$ are of the same order for $k \lesssim \frac{1}{\varepsilon}$, and grow only polynomially quickly for $k \gg \frac{1}{\varepsilon}$. Then we have the heuristic bound $C(P_n) := \sum_{k=1}^{n} \mathrm{Vol}(S_k) e^{-k\varepsilon/2} \gtrsim \mathrm{Vol}(S_1) \sum_{k=1}^{n} e^{-k\varepsilon/2} \gtrsim \varepsilon^{-1} \mathrm{Vol}(S_1)$, while

$$\mathbb{E}[|M(P_n) - f(P_n)|] \le \sum_{k=1}^{n} \frac{\mathrm{Vol}(S_k) e^{-k\varepsilon/2}}{\sum_{i=1}^{n} \mathrm{Vol}(S_i) e^{-i\varepsilon/2}} \omega(k) \overset{\text{heuristic}}{\lesssim} \sum_{k=1}^{n} \varepsilon e^{-k\varepsilon/2} \omega(k) \lesssim \max_k e^{-k\varepsilon/2} \omega(k),$$

where the heuristic inequality is our bound on the normalizing constant $C(P_n)$, and the final bound follows because maxima are larger than (weighted) averages. Continuing the heuristic derivation, the final maximum has is exponentially small weight on $\omega(k)$ for $k \gg \frac{1}{\varepsilon}$. Thus—and again, this is highly non-rigorous—we expect roughly that

$$\mathbb{E}[|M(P_n) - f(P_n)|] \overset{\text{heuristic}}{\lesssim} \max_k e^{-k\varepsilon/2} \omega(k) \overset{\text{heuristic}}{\lesssim} \omega_f\left(\frac{c}{\varepsilon}, P_n\right), \tag{7.4.7}$$

where c is some numerical constant.

To gain some intuition for the claims of optimality we have made, let us revisit the equivalent definitions of privacy that repose on testing, as in Eq. (7.1.4) and Proposition 7.1.6. By the definition of differential privacy, the inverse sensitivity mechanism satisfies

$$\mathbb{P}(M(P_n) \in A) \le e^{k\varepsilon} \mathbb{P}(M(P_n') \in A)$$

for any samples P_n, P_n' satisfying $d_{\mathrm{ham}}(P_n, P_n') \le k$. So for $k \le \frac{1}{\varepsilon}$, we have

$$\mathbb{P}(M(P_n) \in A) \le \exp(1) \mathbb{P}(M(P_n') \in A),$$

and so no procedure exists that can test whether the sample is P_n or P_n' with probability of error less than e^{-2}, by Proposition 7.1.6. Thus, at a fundamental level, *no* procedure can reliably distinguish the outputs of $M(P_n)$ from those of $M(P_n')$ when P_n and P_n' differ in only $1/\varepsilon$ examples. Thus, we cannot expect to estimate $f(P_n)$ to accuracy better than $\omega_f(\frac{1}{\varepsilon}, P_n)$, and so for any ε-differentially private mechanism M and P_n, there exists $P_n' \in \mathcal{P}_n$ with $d_{\mathrm{ham}}(P_n, P_n') \le \frac{1}{\varepsilon}$ and for which

$$\max_{\widehat{P} \in \{P_n, P_n'\}} \mathbb{E}\left[|M(\widehat{P}) - f(\widehat{P})|\right] \gtrsim \omega_f\left(\frac{1}{\varepsilon}, P_n\right), \tag{7.4.8}$$

which the heuristic calculation (7.4.7) achieves.

To provide more rigorous guarantees requires restrictions on the functions f whose values we wish to release. The simplest is that the function $f : \mathcal{P}_n \to \mathbb{R}$ obey a natural ordering property, where larger changes in the sample distribution P_n beget larger changes in f.

Definition 7.8. *A function $f : \mathcal{P}_n \to \mathbb{R}$ is* sample monotone *if for each $s, t \in f(\mathcal{P}_n)$ satisfying $f(P_n) \le s \le t$ or $t \le s \le f(P_n)$, we have $d_f(s, P_n) \le d_f(t, P_n)$.*

So the mean and median (Examples 7.4.8 and 7.4.9) are both sample monotone. So, too, are appropriately continuous functions f. For this, we make the obvious identification of $f : \mathcal{P}_n \to \mathbb{R}$ with the induced function on $\mathcal{X}^n$ by defining $f_{\mathcal{X}}(x_1^n) := f(n^{-1} \sum_{i=1}^n \mathbf{1}_{x_i})$. Then we say $f : \mathcal{P}_n \to \mathbb{R}$ is continuous if the induced function $f_{\mathcal{X}}$ is.

Observation 7.4.10. *Let $f : \mathcal{P}_n \to \mathbb{R}$ be continuous and $\mathcal{X}$ convex. Then f is sample monotone.*

Proof Identify f with its induced function $f_{\mathcal{X}}$ for notational simplicity, and let $x \in \mathcal{X}^n$, $f(x) \le s \le t$, and $P_n = n^{-1} \sum_{i=1}^n \mathbf{1}_{x_i}$ be the empirical distribution associated with x. We show that $d_f(s, P_n) \le d_f(t, P_n)$. If $d_f(t, P_n) = +\infty$, then the desired inequality holds trivially. Otherwise, let $x' \in \mathcal{X}^n$ satisfy $f(x') = t$ and $d_{\rm ham}(x, x') = d_f(t, P_n)$. Then the function $g(\lambda) := f((1 - \lambda)x + \lambda x')$ is continuous in λ and satisfies $g(0) = f(x) \le g(1) = f(x') = t$. By the intermediate value theorem, there exists $\lambda_s \in [0, 1]$ with $g(\lambda_s) = s$, and as $\mathcal{X}$ is convex the vector $x_s = (1 - \lambda_s)x + \lambda_s x' \in \mathcal{X}^n$ satisfies $f(x_s) = g(\lambda_s) = s$. That x_s is a convex combination of x and x' then implies $d_f(s, P_n) \le d_{\rm ham}(x, x_s) \le d_{\rm ham}(x, x') = d_f(t, P_n)$. □

With Definition 7.8 in place, we can provide a few stronger guarantees for the inverse sensitivity mechanism. To avoid pathological sampling issues, one replaces the inverse sensitivity $d_f(t, P_n)$ with a "smoothed" version, where for $\rho \ge 0$ we define

$$d_{f,\rho}(t, P_n) := \inf \left\{ d_{\rm ham}(P_n, P_n') \mid |f(P_n') - t| \le \rho \right\}.$$

(Pathological cases include estimating the median where the sample P_n consists of a single point repeated n times, which would make the density (7.4.4) uniform.) Then instead of the density (7.4.4), we define the *continuous inverse sensitivity mechanism* $M_{\rm cont}$ to have density

$$q(t \mid P_n) = \frac{\exp(-\frac{\varepsilon}{2} d_{f,\rho}(t, P_n))}{\int \exp(-\frac{\varepsilon}{2} d_{f,\rho}(s, P_n)) ds}. \tag{7.4.9}$$

While the parameter ρ adds complexity, setting it to be very small (say, $\rho = \frac{1}{n^2}$) is a reasonable practical default.

The continuous inverse sensitivity enjoys fairly strong error guarantees, as the next two propositions demonstrate, providing two prototypical results. (Exercises 7.11 and 7.12 show how to prove the propositions.) The first proposition shows that the inverse sensivity mechanism is essentially never worse than the Laplace mechanism (7.1.3) when $\varepsilon \lesssim 1$.

Proposition 7.4.11. *Let f be sample monotone (Definition 7.8) and have finite global sensitivity $\mathrm{GS}(f) < \infty$. Then taking $\rho = 0$,*

$$\mathbb{E}\left[|M_{\rm cont}(P_n) - f(P_n)|\right] \le \frac{1}{1 - e^{-\varepsilon/2}} \mathrm{GS}(f).$$

As Example 7.1.3 shows, the standard Laplace mechanism M has error

$$\mathbb{E}\left[|M(P_n) - f(P_n)|\right] = \frac{\mathrm{GS}(f)}{\varepsilon},$$

the same scaling Proposition 7.4.11 guarantees, because $1 - e^{-\varepsilon/2} = \varepsilon/2 + O(\varepsilon^2)$.

For the next proposition, which provides a more nuanced guarantee, we require local sensitivities for samples P_n' near P_n, and so we define the largest local sensitivity within Hamming distance K of the sample P_n by

$$L(K) := \sup_{P_n' \in \mathcal{P}_n} \left\{ \mathrm{LS}(f, P_n') \mid d_{\mathrm{ham}}(P_n, P_n') \le K \right\} = \sup_{P_n' \in \mathcal{P}_n} \left\{ \omega_f(1, P_n') \mid d_{\mathrm{ham}}(P_n, P_n') \le K \right\},$$

where we recall the definition (7.4.2) of the local sensitivity of f. Then we have the following.

Proposition 7.4.12. *Let f be sample monotone (Definition 7.8) and have finite global sensitivity $\mathrm{GS}(f) < \infty$. Then for any $\rho \ge 0$ and $K_n = \left\lceil \frac{4 \log(2n \mathrm{GS}(f)/\rho)}{\varepsilon} \right\rceil$,*

$$\mathbb{E}\left[|M_{\mathrm{cont}}(P_n) - f(P_n)|\right] \le 2\rho + \frac{1}{1 - e^{-\varepsilon/2}} L(K_n).$$

Unpacking Proposition 7.4.12 a bit, let us make the default substitution $\rho = \frac{1}{n^2}$. Then because $1 - e^{-\varepsilon/2} = \varepsilon/2 + O(\varepsilon^2)$, for $\varepsilon \lesssim 1$ this yields

$$\mathbb{E}\left[|M_{\mathrm{cont}}(P_n) - f(P_n)|\right] \lesssim \frac{1}{\varepsilon} \sup_{P_n' \in \mathcal{P}_n} \left\{ \mathrm{LS}(f, P_n') \mid d_{\mathrm{ham}}(P_n', P_n) \le K_n \right\} + \frac{1}{n^2},$$

where $K_n = \frac{4 \log \mathrm{GS}(f) + 12 \log n}{\varepsilon} \lesssim \frac{1}{\varepsilon} \log n$ for large sample sizes n. Comparing this to the sketched lower bound (7.4.8), these quantities are of the same order whenever the moduli of continuity $\omega_f(k; P_n)$ are roughly additive and comparable near P_n, so that for $k \lesssim \frac{1}{\varepsilon}$ there is a chain $P_n^{(1)}, P_n^{(2)}, \ldots, P_n^{(k)}$ with $d_{\mathrm{ham}}(P_n^{(i)}, P_n^{(i+1)}) = 1$ and $\omega_f(k; P_n) \gtrsim \sum_{i=1}^k \mathrm{LS}(f, P_n^{(i)})$ and $\mathrm{LS}(f, P_n) \asymp \mathrm{LS}(f, P_n')$ for P_n' satisfying $d_{\mathrm{ham}}(P_n, P_n') \lesssim \frac{\log n}{\varepsilon}$. Under these conditions—which often require care to check, but which hold, for example, for mean estimation—we then obtain

$$\mathbb{E}\left[|M_{\mathrm{cont}}(P_n) - f(P_n)|\right] \lesssim \omega_f\left(\frac{1}{\varepsilon}, P_n\right) + \frac{1}{n^2}.$$

7.5 Deferred proofs

7.5.1 Proof of Lemma 7.2.10

We prove the first statement of the lemma first. Let us assume there exists R such that $\|P - R\|_{\mathrm{TV}} \le \delta$ and $D_\infty(R \| Q) \le \varepsilon$. Then for any set S we have

$$P(S) \le R(S) + \delta \le e^\varepsilon Q(S) + \delta, \quad \text{i.e.} \quad \log \frac{P(S) - \delta}{Q(S)} \le \varepsilon,$$

which is equivalent to $D_\infty^\delta(P \| Q) \le \varepsilon$. Now, let us assume that $D_\infty^\delta(P \| Q) \le \varepsilon$, whence we must construct the distribution R.

We assume w.l.o.g. that P and Q have densities p, q, and define the sets

$$S := \{x : p(x) > e^\varepsilon q(x)\} \quad \text{and} \quad T := \{x : p(x) < q(x)\}.$$

On these sets, we have $0 \le P(S) - e^\varepsilon Q(S) \le \delta$ by assumption, and we then define a distribution R with density that we partially specify via

$$\begin{aligned} x \in S &\Rightarrow r(x) := e^\varepsilon q(x) < p(x) \\ x \in (T \cup S)^c &\Rightarrow r(x) := p(x) \le e^\varepsilon q(x) \quad \text{and} \quad r(x) \ge q(x). \end{aligned}$$

Now, we note that $e^\varepsilon q(x) \ge p(x) \ge q(x)$ for $x \in (S \cup T)^c$, and thus

$$\begin{aligned} Q(S) + Q(S^c \cap T^c) &\le e^\varepsilon Q(S) + P(S^c \cap T^c) \\ &= R(S) + R(S^c \cap T^c) \\ &= e^\varepsilon Q(S) + P(S^c \cap T^c) < P(S) + P(S^c \cap T^c). \end{aligned} \tag{7.5.1}$$

In particular, when $x \in T$, we may take the density r so that $p(x) \le r(x) \le q(x)$, as

$$R(S) + R(S^c \cap T^c) + P(T) < 1 \quad \text{and} \quad R(S) + R(S^c \cap T^c) + Q(T) > 1$$

by the inequalities (7.5.1), and so that $R(\mathcal{X}) = 1$. With this, we evidently have $r(x) \le e^\varepsilon q(x)$ by construction, and because $S \subset T^c$, we have

$$R(T) - P(T) = P(T^c) - R(T^c) = P(S \cap T^c) - R(S \cap T^c) + P(S^c \cap T^c) - R(S^c \cap T^c) = P(S) - R(S),$$

where we have used that $r = p$ on $(T \cup S)^c$ by construction. Thus we find that

$$\begin{aligned} \|P - R\|_{\rm TV} &= \frac{1}{2} \int_S |r - p| + \frac{1}{2} \int_T |r - p| = \frac{1}{2}(P(S) - R(S)) + \frac{1}{2}(R(T) - P(T)) \\ &= P(S) - R(S) = P(S) - e^\varepsilon Q(S) \le \delta \end{aligned}$$

by assumption.

Now, we turn to the second statement of the lemma. We start with the easy direction, where we assume that P_0 and Q_0 satisfy $D_\infty(P_0 \| Q_0) \le \varepsilon$ and $D_\infty(Q_0 \| P_0) \le \varepsilon$ as well as $\|P - P_0\|_{\rm TV} \le \delta$ and $\|Q - Q_0\|_{\rm TV} \le \delta$. Then for any set S we have

$$P(S) \le P_0(S) + \frac{\delta}{1 + e^\varepsilon} \le e^\varepsilon Q_0(S) + \frac{\delta}{1 + e^\varepsilon} \le e^\varepsilon Q(S) + e^\varepsilon \delta + \frac{\delta}{1 + e^\varepsilon},$$

or $D_\infty^\delta(P \| Q) \le \varepsilon$. The other direction is similar.

We consider the converse direction, where we have both $D_\infty^\delta(P \| Q) \le \varepsilon$ and $D_\infty^\delta(Q \| P) \le \varepsilon$. Let us construct P_0 and Q_0 as in the statement of the lemma. Define the sets

$$S := \{x : p(x) > e^\varepsilon q(x)\} \quad \text{and} \quad S' := \{x : q(x) > e^\varepsilon p(x)\}$$

as well as the sets

$$T := \{x : e^\varepsilon q(x) \ge p(x) \ge q(x)\} \quad \text{and} \quad T' := \{x : e^{-\varepsilon} q(x) \le p(x) < q(x)\},$$

so that S, S', T, T' are all disjoint, and $\mathcal{X} = S \cup S' \cup T \cup T'$. We begin by constructing intermediate measures—which end up not being probabilities—P_1 and Q_1, which we modify slightly to actually construct P_0 and Q_0. We first construct densities similar to our construction above for part (i), setting

$$\begin{aligned} x \in S \Rightarrow p_1(x) &:= e^\varepsilon q_1(x), \quad q_1(x) := \frac{1}{1 + e^\varepsilon}(p(x) + q(x)) \\ x \in S' \Rightarrow q_1(x) &:= e^\varepsilon p_1(x), \quad p_1(x) := \frac{1}{1 + e^\varepsilon}(p(x) + q(x)). \end{aligned}$$

Now, define the two quantities

$$\alpha := P(S) - P_1(S) = P(S) - \frac{e^\varepsilon}{1 + e^\varepsilon}(P(S) + Q(S)) = \frac{P(S) - e^\varepsilon Q(S)}{1 + e^\varepsilon} \le \frac{\delta}{1 + e^\varepsilon}.$$

and similarly

$$\alpha' := Q(S') - Q_1(S') = \frac{Q(S') - e^\varepsilon P(S')}{1 + e^\varepsilon} \le \frac{\delta}{1 + e^\varepsilon}.$$

Note also that we have $P(S) - P_1(S) = Q_1(S) - Q(S)$ and $Q(S') - Q_1(S') = P_1(S') - P(S')$ by construction.

We assume w.l.o.g. that $\alpha \ge \alpha'$, so that if $\beta = \alpha - \alpha' \ge 0$, we have $\beta \le \frac{\delta}{1+e^\varepsilon}$, and we have the sandwiching

$$P_1(S) + P_1(S') + P(T \cup T') = P_1(S) + P_1(S') + 1 - P(S \cup S') = 1 - \beta < 1$$

because S and S' are disjoint and $T_< \cup T_> = (S \cup S')^c$, and similarly

$$Q_1(S) + Q_1(S') + Q(T \cup T') = Q_1(S) + Q_1(S') + 1 - Q(S \cup S') = 1 + \beta > 1.$$

Let $p_1 = p$ on the set $T \cup T'$ and similarly for $q_1 = q$. Then we have $P_1(\mathcal{X}) = 1 - \beta$, $Q_1(\mathcal{X}) = 1 + \beta$, and $|\log \frac{p_1}{q_1}| \le \varepsilon$.

Now, note that $S \cup T = \{x : q_1(x) \ge p_1(x)\}$, and we have

$$\begin{aligned} Q_1(S) + Q_1(T) - P_1(S) - P_1(T) &= Q_1(S) + Q(T) - P_1(S) - P(T) \\ &\ge Q_1(S) + Q_1(S') + Q(T) + Q(T') - P_1(S) - P_1(S') - P(T) - P(T') = 2\beta. \end{aligned}$$

Now, (roughly) we decrease the density q_1 to q_0 on $S \cup T$ and increase p_1 to p_0 on $S \cup T$, while still satisfying $q_0 \ge p_0$ on $S \cup T$. In particular, we may choose the densities $q_0 = q_1$ on $T' \cup S'$ and $p_0 = p_1$ on $T' \cup S'$, while choosing q_0, p_0 so that

$$p_1(x) \le p_0(x) \le q_0(x) \le q_1(x) \;\; \text{on } S \cup T,$$

where

$$P_0(S \cup T) = P_1(S \cup T) + \beta \quad \text{and} \quad Q_0(S \cup T) = Q_1(S \cup T) - \beta. \tag{7.5.2}$$

With these choices, we evidently obtain $Q_0(\mathcal{X}) = P_0(\mathcal{X}) = 1$ and that $D_\infty(P_0 \| Q_0) \le \varepsilon$ and $D_\infty(Q_0 \| P_0) \le \varepsilon$ by construction. It remains to consider the variation distances. As $p_0 = p$ on T', we have

$$\begin{aligned} \|P - P_0\|_{\rm TV} &= \frac{1}{2}\int_S |p - p_0| + \frac{1}{2}\int_{S'} |p - p_0| + \frac{1}{2}\int_T |p - p_0| \\ &= \frac{1}{2}(P(S) - P_0(S)) + \frac{1}{2}(P_0(S') - P(S)) + \frac{1}{2}(P_0(T) - P(T)) \\ &\le \frac{1}{2}\underbrace{(P(S) - P_1(S))}_{=\alpha} + \frac{1}{2}\underbrace{(P_0(S') - P(S))}_{=\alpha'} + \frac{1}{2}\underbrace{(P_0(T) - P(T))}_{\le \beta}, \end{aligned}$$

where the $P_0(T) - P(T) \le \beta$ claim follows becase $p_1(x) = p(x)$ on T and by the increasing construction yielding equality (7.5.2), we have $P_0(T) - P(T) = P_0(T) - P_1(T) = \beta + P_1(S) - P_0(S) \le \beta$. In particular, we have $\|P - P_0\|_{\rm TV} \le \frac{\alpha + \alpha'}{2} + \frac{\beta}{2} = \alpha \le \frac{\delta}{1+e^\varepsilon}$. The argument that $\|Q - Q_0\|_{\rm TV} \le \frac{\delta}{1+e^\varepsilon}$ is similar.

7.6 Bibliography

Given the broad focus of this book, our treatment of privacy is necessarily somewhat brief, and there is substantial depth to the subject that we do not cover.

The initial development of randomized response began with Warner [173], who proposed randomized response in survey sampling as a way to collect sensitive data. This elegant idea remained in use for many years, and a generalization to data release mechanisms with bounded likelihood ratios—essentially, the local differential privacy definition 7.2—is due to Evfimievski et al. [80] in 2003 in the databases community. Dwork, McSherry, Nissim, and Smith [74] and the subsequent work of Dwork et al. [73] defined differential privacy and its (ε, δ)-approximate relaxation. A small industry of research has built out of these papers, with numerous extensions and developments.

Exponential mechanism is McSherry and Talwar [139].

The book of Dwork and Roth [72] surveys much of the field, from the perspective of computer science, as of 2014. Lemma 7.2.10 is due to Dwork et al. [75], and our proof is based on theirs.

7.7 Exercises

Exercise 7.1: Prove Proposition 7.2.1.

Exercise 7.2: Prove Proposition 7.4.7.

Exercise 7.3 (Laplace mechanisms versus randomized response)**:** In this question, you will investigate using Laplace and randomized response mechanisms, as in Examples 7.1.3 and 7.1.1–7.1.2, to perform *locally* private estimation of a mean, and compare this with randomized-response based mechanisms.

We consider the following scenario: we have data $X_i \in [0, 1]$, drawn i.i.d., and wish to estimate the mean $\mathbb{E}[X]$ under local ε-differential privacy.

(a) The Laplace mechanism simply sets $Z_i = X_i + W_i$ for $W_i \stackrel{\text{iid}}{\sim} \mathsf{Laplace}(b)$ for some b. What choice of b guarantees ε-local differential privacy?

(b) For your choice of b, let $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$. Give $\mathbb{E}[(\overline{Z}_n - \mathbb{E}[X])^2]$.

(c) A randomized response mechanism for this case is the following: first, we randomly round X_i to $\{0, 1\}$, by setting

$$\widetilde{X}_i = \begin{cases} 1 & \text{with probability } X_i \\ 0 & \text{otherwise.} \end{cases}$$

Conditional on $\widetilde{X}_i = x$, we then set

$$Z_i = \begin{cases} x & \text{with probability } \frac{e^\varepsilon}{1+e^\varepsilon} \\ 1 - x & \text{with probability } \frac{1}{1+e^\varepsilon}. \end{cases}$$

What is $\mathbb{E}[Z_i]$?

(d) For the randomized response Z_i above, give constants a and b so that $aZ_i - b$ is unbiased for $\mathbb{E}[X]$, that is, $\mathbb{E}[aZ_i - b] = \mathbb{E}[X]$. Let $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n (aZ_i - b)$ be your mean estimator. What is $\mathbb{E}[(\widehat{\theta}_n - \mathbb{E}[X])^2]$? Does this converge to the mean-square error of the sample mean $\mathbb{E}[(\overline{X}_n - \mathbb{E}[X])^2] = \mathrm{Var}(X)/n$ as $\varepsilon \uparrow \infty$?

(e) Now, it is time to compare the simple randomized response estimator from part (d) with the Laplace mechanism from part (b). For each of the following distributions, generate samples of size $N = 10, 100, 1000, 10000$, and then for $T = 25$ tests, compute the two estimators, both with $\varepsilon = 1$. Then plot the mean-squared error and confidence intervals for each of the two methods as well as the sample mean without any privacy.

i. Uniform distribution: $X \sim \mathsf{Uniform}[0, 1]$, with $\mathbb{E}[X] = 1/2$.

ii. Bernoulli distribution: $X \sim \mathsf{Bernoulli}(p)$, where $p = .1$.

iii. Uniform distribution: $X \sim \mathsf{Uniform}[.49, .51]$, with $\mathbb{E}[X] = 1/2$.

Do you prefer the Laplace or randomized response mechanism? In one sentence, why?

Exercise 7.4 (A more sophisticated randomized response scheme)**:** Let us consider a more sophisticated randomized response scheme than that in Exercise 7.3. Define quantized values

$$b_0 = 0, \; b_1 = \frac{1}{k}, \; \ldots, b_{k-1} = \frac{k-1}{k}, \; b_k = 1. \tag{7.7.1}$$

Now consider a randomized response estimator that, when $X \in [b_j, b_{j+1}]$ first rounds X randomly to $\widetilde{X} \in \{b_j, b_{j+1}\}$ so that $\mathbb{E}[\widetilde{X} \mid X] = X$. Conditional on $\widetilde{X} = j$, we then set

$$Z = \begin{cases} j & \text{with probability } \frac{e^\varepsilon}{k+e^\varepsilon} \\ \mathsf{Uniform}(\{0, \ldots, k\} \setminus \{j\}) & \text{with probability } \frac{k}{k+e^\varepsilon}. \end{cases}$$

(a) Give a and b so that $\mathbb{E}[aZ - b] = \mathbb{E}[X]$.

(b) For your values of a and b above, let $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n (aZ_i - b)$. Give a (reasonably tight) bound on $\mathbb{E}[(\widehat{\theta}_n - \mathbb{E}[X])^2]$.

(c) For any given $\varepsilon > 0$, give (approximately) the k in the choice of the number of bins (7.7.1) that optimizes your bound, and (approximately) evaluate $\mathbb{E}[(\widehat{\theta}_n - \mathbb{E}[X])^2]$ with your choice of k. As $\varepsilon \uparrow \infty$, does this converge to $\mathrm{Var}(X)/n$?

Exercise 7.5 (Subsampling via divergence measures (Balle et al. [14]))**:** The *hockey stick* divergence functional, defined for $\alpha \geq 1$, is $\phi_\alpha(t) = [1 - \alpha t]_+$. It is straightforward to relate this to (ε, δ)-differential privacy via Definition 7.6: two distributions P and Q are (ε, δ)-close if and only their ϕ_{e^ε}-divergences are less than δ, i.e., if and only if

$$D_{\phi_{e^\varepsilon}}(P\|Q) \leq \delta \;\; \text{and} \;\; D_{\phi_{e^\varepsilon}}(Q\|P) \leq \delta.$$

(In your answer to this question, feel free to use $D_\alpha(P\|Q)$ as a shorthand for $D_{\phi_\alpha}(P\|Q)$.)

(a) Let P_0, P_1, Q_1 be any three distributions, and for some $q \in [0, 1]$ and $\alpha \geq 1$, define $P = (1-q)P_0 + qP_1$ and $Q = (1-q)P_0 + qQ_1$. Let $\alpha' = 1 + q(\alpha - 1) = (1-q) + q\alpha$ and $\theta = \alpha'/\alpha \leq 1$. Show that

$$D_{\phi_{\alpha'}}(P\|Q) = qD_{\phi_\alpha}\left((1-\theta)P_0 + \theta P_1 \| Q_1\right).$$

(b) Let $\varepsilon > 0$ and define $\varepsilon(q) = \log(1 + q(e^\varepsilon - 1))$. Show that

$$D_{\phi_{e^{\varepsilon(q)}}}(P\|Q) \leq q \max\left\{D_{\phi_{e^\varepsilon}}(P_0\|Q_1), D_{\phi_{e^\varepsilon}}(P_1\|Q_1)\right\}.$$

Exercise 7.6 (Subsampling and privacy amplification (Balle et al. [14])): Consider the following subsampling approach to privacy. Assume that we have a private (randomized) algorithm, represented by $\mathcal{A}$, that acts on samples of size m and guarantees (ε, δ)-differential privacy. The subsampling mechanism is then defined as follows: given a sample X_1^n of size $n > m$, choose a subsample $X_{\rm sub}$ of size m uniformly at random from X_1^n, and then release $Z = \mathcal{A}(X_{\rm sub})$.

(a) Use the results of parts (a) and (b) in Exercise 7.5 to show that Z is $(\varepsilon(q), \delta q)$-differentially private, where $q = m/n$ and $\varepsilon(q) = \log(1 + q(e^\varepsilon - 1))$.

(b) Show that if $\varepsilon \le 1$, then Z is $((e-1)q\varepsilon, q\delta)$-differentially private, and if $\varepsilon \le \frac{1}{2}$, then Z is $(2(\sqrt{e}-1)q\varepsilon, q\delta)$-differentially private. *Hint:* Argue that for any $T > 0$, one has $e^t - 1 \le (e^T - 1)\frac{t}{T}$ for all $t \in [0, T]$.

Exercise 7.7 (Concentration and privacy composition): In this question, we give an alternative to the privacy composition approaches we exploit in Section 7.3.2. Consider an identical scenario to that in Fig. 7.1, and begin by assuming that each channel Q_i is ε-differentially private with density q_i, and let $Q^{(b)}$ be shorthand for $Q(\cdot \mid x^{(b)})$. Define the log-likelihood ratio

$$L^{(b)}(Z_1^k) := \sum_{i=1}^k \log \frac{q_i^{(b)}(Z_i)}{q_i^{(1-b)}(Z_i)}.$$

(a) Let P, Q be any two distributions satisfying $D_\infty(P\|Q) \le \varepsilon$ and $D_\infty(Q\|P) \le \varepsilon$, i.e., that $\log \frac{P(A)}{Q(A)} \in [-\varepsilon, \varepsilon]$ for all sets A. Show that

$$D_{\rm kl}(P\|Q) \le \varepsilon(e^\varepsilon - 1).$$

(b) Let $Q^{(b)}$ denote the joint distribution of $Z_1, \ldots, Z_k$ when bit b holds in the privacy game in Fig. 7.1. Show that

$$\mathbb{E}_b[L^{(b)}(Z_1^k)] \le k\varepsilon(e^\varepsilon - 1)$$

where $\mathbb{E}_b$ denotes expectation under $Q^{(b)}$, and that for all $t \ge 0$,

$$Q^{(b)}\left(L^{(b)}(Z_1^k) \ge k\varepsilon(e^\varepsilon - 1) + t\right) \le \exp\left(-\frac{t^2}{2k\varepsilon^2}\right).$$

Conclude that for any $\delta \in (0,1)$, with probability at least $1 - \delta$ over $Z_1^k \sim Q^{(b)}$,

$$L^{(b)}(Z_1^k) \le k(e^\varepsilon - 1)\varepsilon + \sqrt{2k \log \frac{1}{\delta}} \cdot \varepsilon.$$

(c) Argue that for any (measurable) set A,

$$Q^{(b)}(Z_1^k \in A) \le e^{\varepsilon(k,\delta)} \cdot Q^{(1-b)}(Z_1^k \in A) + \delta$$

for all $\delta \in [0,1]$, where $\varepsilon(k,\delta) = k\varepsilon(e^\varepsilon - 1) + \sqrt{2k \log \frac{1}{\delta}} \cdot \varepsilon$.

(d) Conclude the following tighter variant of Corollary 7.3.3: if each channel in Fig. 7.1 is ε-differentially private, then the composition of k such channels is

$$\left(k\varepsilon(e^\varepsilon - 1) + \sqrt{2k \log \frac{1}{\delta}} \cdot \varepsilon, \delta\right)$$

differentially private for all $\delta > 0$.

As an aside, a completely similar derivation yields the following tighter analogue of Theorem 7.3.4: if each channel is (ε, δ)-differentially private, then their composition is

$$\left(k\varepsilon(e^\varepsilon - 1) + \sqrt{2k \log \frac{1}{\delta_0}} \cdot \varepsilon, \delta_0 + \frac{k\delta}{1 + e^\varepsilon}\right)$$

differentially private for all $\delta_0 > 0$.

Exercise 7.8 (One-dimensional minimization with inverse sensitivity)**:** Consider the private minimization of the one dimensional loss $\ell(\theta, x)$ (for $\theta \in \Theta \subset \mathbb{R}$), where we wish to estimate

$$\widehat{\theta}(P_n) \in \operatorname*{argmin}_\theta \{P_n \ell(\theta, X) := \frac{1}{n} \sum_{i=1}^n \ell(\theta, X_i)\},$$

where we recall the notation from Chapters 4 and 5. Assume that the loss ℓ is convex, differentiable in θ, and that it satisfies the Lipschitz-type guarantees that there exist constants $0 < L_0 \le L_1 < \infty$

$$[-L_0, L_0] \subset \{\ell'(\theta, x)\}_{x \in \mathcal{X}} \subset [-L_1, L_1] \tag{7.7.2}$$

for all $\theta \in \Theta$ and that $\{\ell'(\theta, x)\}_{x \in \mathcal{X}}$ is an interval. (That is, the set of potential derivatives $\ell'(\theta, x)$ as x varies includes $[-L_0, L_0]$, is convex, and $|\ell'(\theta, x)| \le L_1$ for all $\theta \in \Theta, x \in \mathcal{X}$.)

(a) Let the loss ℓ be the Huber loss $\ell(\theta, x) = h_u(\theta - x)$ for some fixed $u > 0$, where

$$h_u(t) = \begin{cases} \frac{1}{2u} t^2 & \text{if } |t| \le u \\ |t| + \frac{u}{2} & \text{if } |t| \ge u. \end{cases}$$

When $\mathcal{X} = \mathbb{R}$, show that ℓ satisfies the containment (7.7.2) with $L_0 = L_1 = 1$.

(b) Let the loss ℓ be the absolute value $\ell(\theta, x) = |\theta - x|$, where we abuse notation to call $\{\ell'(\theta, x)\}_{x=\theta} = [-1, 1]$ (the subdifferential). When $\mathcal{X} = \mathbb{R}$, show that ℓ satisfies the containment (7.7.2) with $L_0 = L_1 = 1$.

(c) Let $d_{\widehat{\theta}}$ be the inverse sensitivity (7.4.3) for the minimizer $\widehat{\theta}(P_n)$, which is the solution (in θ) to $P_n \ell'(\theta, X) = 0$. Assuming inequality (7.7.2) holds, show that

$$\left\lceil \frac{n|P_n \ell'(\theta, X)|}{2L_1} \right\rceil \le d_{\widehat{\theta}}(\theta, P_n) \le \left\lceil \frac{n|P_n \ell'(\theta, X)|}{L_0} \right\rceil.$$

(d) Show that the function

$$d(\theta, P_n) := \left\lceil \frac{n|P_n \ell'(\theta, X)|}{2L_1} \right\rceil$$

is 1-Lipschitz with respect to the Hamming metric in P_n.

The Lipschitz behavior of $d(\theta, P_n)$ in part (d) makes this a computationally attractive alternative to the pure inverse sensitivity (7.4.3) and associated mechanism with density (7.4.4).

Exercise 7.9 (Estimating means with inverse sensitivity mechanisms)**:** In this question, we compare behavior of mean estimation under differential privacy with the Laplace mechanism and the inverse sensitivity-type mechanism in Example 7.4.8. Let $\mathcal{X} = [-1, 1]$ be the data space and consider estimating the mean $\overline{x}_n$ of $x_1^n \in \mathcal{X}^n$.

(a) Implement the Laplace mechanism (7.1.3) for this problem. Fix $n = 200$ and repeat the following experiment 50 times. For $\varepsilon = .1, .5, 1, 2$, generate a sample $x_1^n \in \mathcal{X}^n$ (from whatever distribution you like), then estimate $\overline{x}_n$ using the Laplace mechanism. Give a table of the mean squared errors $(\overline{x}_n - M(x_1^n))^2$.

(b) Implement the inverse sensitivity mechanism using the approximation in Example 7.4.8. Repeat the experiment in part (a).

(c) Compare the results.

Exercise 7.10 (Estimating medians with the inverse sensitivity mechanism)**:** The data at `https://stats311.stanford.edu/data/salaries.txt` contains approximately 250,000 salaries from the University of California Schools between 2011 and 2014. Assuming that the maximum salary is $3 \cdot 10^6$ and minimum is 0 (so the data $x \in [0, 3 \cdot 10^6]$), implement the inverse sensitivity mechanism for the median as in Example 7.4.9. Repeat the following 20 times: for each of $\varepsilon = .0625, .125, .25, .5, 1, 2$, estimate the median using the inverse sensitivity mechanism with ε-differential privacy. Compute the mean absolute errors across the 20 experiments for each ε.

Exercise 7.11 (Shells and accuracy in inverse sensitivity)**:** Let $f : \mathcal{P}_n \to \mathbb{R}$ be sample monotone (Def. 7.8) and $\rho \geq 0$. Let $M = M_{\rm cont}$ be the continuous inverse sensitivity mechanism with density (7.4.9). Define the upper and lower shells

$$S_{k^+} = \{t > f(P_n) \mid d_{f,\rho}(t, P_n) = k\} \quad \text{and} \quad S_{k^-} = \{t < f(P_n) \mid d_{f,\rho}(t, P_n) = k\},$$

and the upper and lower moduli of continuity (values in the shells $S_{k^\pm}$)

$$\omega^+(k) := \sup\{t \in S_{k^+}\} - f(P_n) \quad \text{and} \quad \omega^-(k) := f(P_n) - \inf\{t \in S_{k^-}\}.$$

Let $S_0 = \{t \in \mathbb{R} \mid |f(P_n) - t| \leq \rho\}$.

(a) Justify the inequality

$$\begin{aligned}
&\mathbb{E}[|M(P_n) - f(P_n)|] \\
&\leq \mathbb{P}(M(P_n) \in S_0)\rho + \sum_{k=1}^n \mathbb{P}(M(P_n) \in S_{k^+})(\omega^+(k) + \rho) + \sum_{k=1}^n \mathbb{P}(M(P_n) \in S_{k^-})(\omega^-(k) + \rho).
\end{aligned}$$

(b) Bound $\mathbb{P}(M(P_n) \in S_{k^+})$ and $\mathbb{P}(M(P_n) \in S_{k^-})$, and using these bounds demonstrate that

$$\begin{aligned}
&\mathbb{E}\left[|M(P_n) - f(P_n)|\right] \\
&\leq \rho + \frac{\sum_{k=1}^n \omega^+(k) \cdot (\omega^+(k) - \omega^+(k-1)) e^{-k\varepsilon/2}}{\rho + \sum_{k=1}^n (\omega^+(k) - \omega^+(k-1)) e^{-k\varepsilon/2} + \sum_{k=1}^n (\omega^-(k) - \omega^-(k-1)) e^{-k\varepsilon/2}} \\
&\quad + \frac{\sum_{k=1}^n \omega^+(k) \cdot (\omega^-(k) - \omega^-(k-1)) e^{-k\varepsilon/2}}{\rho + \sum_{k=1}^n (\omega^-(k) - \omega^-(k-1)) e^{-k\varepsilon/2} + \sum_{k=1}^n (\omega^-(k) - \omega^-(k-1)) e^{-k\varepsilon/2}}
\end{aligned}$$

(c) Show that

$$\sum_{k=1}^n \left[(\omega^-(k) - \omega^-(k-1)) + \omega^+(k) - \omega^+(k-1)\right] e^{-k\varepsilon/2} \geq (1 - e^{-\varepsilon/2}) \sum_{k=1}^n \left(\omega^+(k) + \omega^-(k)\right) e^{-k\varepsilon/2}.$$

Exercise 7.12 (Accuracy of the inverse sensitivity mechanism)**:** In this question, we prove Propositions 7.4.11 and 7.4.12. Let the conditions and notation of Exercise 7.11 hold. Recall the definition

$$L(K) := \sup_{P_n' \in \mathcal{P}_n} \left\{ \mathrm{LS}(f, P_n') \mid d_{\mathrm{ham}}(P_n, P_n') \le K \right\}.$$

(a) Use Exercise 7.11.(b) and (c) to show that for any $K \in \mathbb{N}$,

$$\mathbb{E}\left[|M_{\mathrm{cont}}(P_n) - f(P_n)|\right] \le \rho + \frac{L(K)}{1 - e^{-\varepsilon/2}} \cdot \frac{\sum_{k=1}^{K} \left(\omega^+(k) + \omega^-(k)\right) e^{-k\varepsilon/2}}{\sum_{k=1}^{n} \left(\omega^+(k) + \omega^-(k)\right) e^{-k\varepsilon/2}} + \frac{\mathrm{GS}(f)}{\rho} \sum_{k=K+1}^{n} \left(\omega^+(k) + \omega^-(k)\right) e^{-k\varepsilon/2}.$$

(b) Choose values for ρ and K to show that $\mathbb{E}[|M_{\mathrm{cont}}(P_n) - f(P_n)|] \le \frac{1}{1-e^{-\varepsilon/2}} \mathrm{GS}(f)$, giving Proposition 7.4.11.

(c) Prove Proposition 7.4.12.

Exercise 7.13 (Subsampling and Rényi privacy)**:** We would like to estimate the mean $\mathbb{E}[X]$ of $X \sim P$, where $X \in B = \{x \in \mathbb{R}^d \mid \|x\|_2 \le 1\}$, the ℓ_2-ball in $\mathbb{R}^d$. We investigate the extent to which subsampling of a dataset can *improve* privacy by providing some additional anonymity. Consider the following mechanism for estimating (scaled) multiples of this mean: for a dataset $\{X_1, \ldots, X_n\}$, we let $S_i \in \{0, 1\}$ be i.i.d. $\mathsf{Bernoulli}(q)$, that is, $\mathbb{E}[S_i] = q$, and then consider the algorithm

$$Z = \sum_{i=1}^{n} X_i S_i + \sigma W, \quad W \sim \mathsf{N}(0, I_d). \tag{7.7.3}$$

In this question, we investigate the Rényi privacy properties of the subsampling (7.7.3). (Recall the Rényi divergence of Definition 7.4, $D_\alpha(P\|Q) = \frac{1}{\alpha-1} \log \int (p/q)^\alpha q$.)

We consider a slight variant of Rényi privacy, where we define data matrices X and X' to be adjacent if $X \in \mathbb{R}^{d \times n}$ and $X' \in \mathbb{R}^{d \times n-1}$ where X' is X with a single column removed. Then a mechanism is (ε, α)-Rényi private against single removals if and only if

$$D_\alpha\big(Q(\cdot \mid X) \| Q(\cdot \mid X')\big) \le \varepsilon \quad \text{and} \quad D_\alpha\big(Q(\cdot \mid X') \| Q(\cdot \mid X)\big) \le \varepsilon \tag{7.7.4}$$

for all neighboring X and X' consisting of samples of size n and $n-1$, respectively.

(a) Let $Q(\cdot \mid X)$ and $Q(\cdot \mid X')$ denote the channels for the mechanism (7.7.3) with data matrices $X = [x_1 \ \cdots \ x_{n-1} \ x]$ and $X' = [x_1 \ \cdots \ x_{n-1}] \in \mathbb{R}^{d \times n}$. Let P_μ denote the normal distribution $\mathsf{N}(\mu, \sigma^2 I)$ with mean μ and covariance $\sigma^2 I$ on $\mathbb{R}^d$. Show that for any $\alpha \in (1, \infty)$,

$$D_\alpha\big(Q(\cdot \mid X) \| Q(\cdot \mid X')\big) \le D_\alpha(qP_x + (1-q)P_0 \| P_0)$$

and

$$D_\alpha\big(Q(\cdot \mid X') \| Q(\cdot \mid X)\big) \le D_\alpha(P_0 \| qP_x + (1-q)P_0).$$

(b) Show that for the Rényi $\alpha = 2$-divergence,

$$D_2\left(qP_x + (1-q)P_0 \| P_0\right) \le \log\left(1 + q^2\left(\exp(\|x\|_2^2/\sigma^2) - 1\right)\right) \quad \text{and}$$
$$D_2\left(P_0 \| qP_x + (1-q)P_0\right) \le \log\left(1 + \frac{q^2}{1-q}\left(\exp(\|x\|_2^2/\sigma^2) - 1\right)\right).$$

(*Hint:* Example 7.2.2.)

Consider two mechanisms for computing a sample mean $\overline{X}_n$ of vectors, where $\|x_i\|_2 \le b$ for all i. The first is to repeat the following T times: for $t = 1, 2, \ldots, T$,

i. Draw $S \in \{0,1\}^n$ with $S_i \stackrel{\text{iid}}{\sim} \mathsf{Bernoulli}(q)$

ii. Set $Z_t = \frac{1}{nq}(XS + \sigma_{\text{sub}} W_t)$, where $W_t \stackrel{\text{iid}}{\sim} \mathsf{N}(0, I)$, as in (7.7.3).

Then set $Z_{\text{sub}} = \frac{1}{T}\sum_{t=1}^T Z_t$. The other mechanism is to simply set $Z_{\text{Gauss}} = \overline{X}_n + \sigma_{\text{Gauss}} W$ for $W \sim \mathsf{N}(0, I)$.

(c) What level of privacy does Z_{sub} have? That is, Z_{sub} is $(\varepsilon, 2)$-Rényi private (against single removals (7.7.4)). Give a tight upper bound on ε.

(d) What level of $(\varepsilon, 2)$-Rényi privacy does Z_{Gauss} provide?

(e) Fix $\varepsilon > 0$, and assume that each mechanism Z_{sub} and Z_{Gauss} have parameters chosen so that they are $(\varepsilon, 2)$-Rényi private. Optimize over $T, q, n, \sigma_{\text{sub}}$ in the subsampling mechanism and σ_{Gauss} in the Gaussian mechanism, and provide the sharpest bound you can on

$$\mathbb{E}[\left\|Z_{\text{sub}} - \overline{X}_n\right\|_2^2] \quad \text{and} \quad \mathbb{E}[\left\|Z_{\text{Gauss}} - \overline{X}_n\right\|_2^2].$$

You may assume $\|x_i\|_2 = b$ for all i. (In your derivation, to avoid annoying constants, you should replace $\log(1+t)$ with its upper bound, $\log(1+t) \le t$, which is fairly sharp for $t \approx 0$.)

Part II

Fundamental limits and optimality

JCD Comment: Put a brief commentary here. Some highlights:

i. Minimax lower bounds (both local and global) using Le Cam's, Fano's, and Assouad's methods. Worked out long example with nonparametric regression.

ii. Strong data processing inequalities, along with some bounds on them (constrained risk inequalities).

iii. Functionals for lower bounds perhaps

Chapter 8

Minimax lower bounds: the Le Cam, Fano, and Assouad methods

Understanding the fundamental limits of estimation and optimization procedures is important for a multitude of reasons. Indeed, developing bounds on the performance of procedures can give complementary insights. By exhibiting fundamental limits of performance (perhaps over restricted classes of estimators), it is possible to guarantee that an algorithm we have developed is optimal, so that searching for estimators with better statistical performance will have limited returns, though searching for estimators with better performance in other metrics may be interesting. Moreover, exhibiting refined lower bounds on the performance of estimators can also suggest avenues for developing alternative, new optimal estimators; lower bounds need not be a fully pessimistic exercise.

In this chapter, we define and then discuss techniques for lower-bounding the minimax risk, giving three standard techniques for deriving minimax lower bounds that have proven fruitful in a variety of estimation problems [177]. In addition to reviewing these standard techniques—the Le Cam, Fano, and Assouad methods—we present a few simplifications and extensions that may make them more "user friendly." Finally, the concluding sections of the chapter (Sections 8.6 and 8.7) present extensions of the ideas to *nonparametric problems*, where the effective number of parameters to estimate grows with the sample size n; this culminates with an essentially geometric treatment of information and divergence measures directly relating covering and packing numbers to estimation.

8.1 Basic framework and minimax risk

Our first step here is to establish the minimax framework we use. When we study classical estimation problems, we use a standard version of minimax risk; we will also show how minimax bounds can be used to study optimization problems, in which case we use a specialization of the general minimax risk that we call minimax *excess* risk (while minimax risk handles this case, it is important enough that we define additional notation).

Let us begin by defining the standard minimax risk, deferring temporarily our discussion of minimax excess risk. Throughout, we let $\mathcal{P}$ denote a class of distributions on a sample space $\mathcal{X}$, and let $\theta : \mathcal{P} \to \Theta$ denote a function defined on $\mathcal{P}$, that is, a mapping $P \mapsto \theta(P)$. The goal is to estimate the parameter $\theta(P)$ based on observations X_i drawn from the (unknown) distribution P. In certain cases, the parameter $\theta(P)$ uniquely determines the underlying distribution; for example, if we attempt to estimate a normal mean θ from the family $\mathcal{P} = \{\mathsf{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$ with

known variance σ^2, then $\theta(P) = \mathbb{E}_P[X]$ uniquely determines distributions in $\mathcal{P}$. In other scenarios, however, θ does not uniquely determine the distribution: for instance, we may be given a class of densities $\mathcal{P}$ on the unit interval $[0,1]$, and we wish to estimate $\theta(P) = \int_0^1 (p'(t))^2 dt$, where p is the density of P. Such problems arise, for example, in estimating the uniformity of the distribution of a species over an area (large $\theta(P)$ indicates an irregular distribution). In this case, θ does not parameterize P, so we take a slightly broader viewpoint of estimating functions of distributions in these notes.

The space Θ in which the parameter $\theta(P)$ takes values depends on the underlying statistical problem; as an example, if the goal is to estimate the univariate mean $\theta(P) = \mathbb{E}_P[X]$, we have $\Theta \subset \mathbb{R}$. To evaluate the quality of an estimator $\widehat{\theta}$, we let $\rho : \Theta \times \Theta \to \mathbb{R}_+$ denote a (semi)metric on the space Θ, which we use to measure the error of an estimator for the parameter θ, and let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function with $\Phi(0) = 0$ (for example, $\Phi(t) = t^2$).

For a distribution $P \in \mathcal{P}$, we assume we receive i.i.d. observations X_i drawn according to some P, and based on these $\{X_i\}$, the goal is to estimate the unknown parameter $\theta(P) \in \Theta$. For a given estimator $\widehat{\theta}$—a measurable function $\widehat{\theta} : \mathcal{X}^n \to \Theta$—we assess the quality of the estimate $\widehat{\theta}(X_1, \ldots, X_n)$ in terms of the risk

$$\mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X_1 \ldots, X_n), \theta(P))\big) \right].$$

For instance, for a univariate mean problem with $\rho(\theta, \theta') = |\theta - \theta'|$ and $\Phi(t) = t^2$, this risk is the mean-squared error. As the distribution P is varied, we obtain the *risk functional* for the problem, which gives the risk of any estimator $\widehat{\theta}$ for the family $\mathcal{P}$.

For any fixed distribution P, there is always a trivial estimator of $\theta(P)$: simply return $\theta(P)$, which will have minimal risk. Of course, this "estimator" is unlikely to be good in any real sense, and it is thus important to consider the risk functional not in a pointwise sense (as a function of individual P) but to take a more global view. One approach to this is Bayesian: we place a prior π on the set of possible distributions $\mathcal{P}$, viewing $\theta(P)$ as a random variable, and evaluate the risk of an estimator $\widehat{\theta}$ taken in expectation with respect to this prior on P. Another approach, first suggested by Wald [172], which is to choose the estimator $\widehat{\theta}$ minimizing the maximum risk

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X_1 \ldots, X_n), \theta(P))\big) \right].$$

An optimal estimator for this metric then gives the *minimax risk*, which is defined as

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta(P))\big) \right], \tag{8.1.1}$$

where we take the supremum (worst-case) over distributions $P \in \mathcal{P}$, and the infimum is taken over all estimators $\widehat{\theta}$. Here the notation $\theta(\mathcal{P})$ indicates that we consider parameters $\theta(P)$ for $P \in \mathcal{P}$ and distributions in $\mathcal{P}$.

In some scenarios, we study a specialized notion of risk appropriate for optimization problems (and statistical problems in which all we care about is prediction). In these settings, we assume there exists some loss function $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$, where for an observation $x \in \mathcal{X}$, the value $\ell(\theta; x)$ measures the instantaneous loss associated with using θ as a predictor. In this case, we define the risk

$$L_P(\theta) := \mathbb{E}_P[\ell(\theta; X)] = \int_{\mathcal{X}} \ell(\theta; x) dP(x) \tag{8.1.2}$$

as the expected loss of the vector θ. (See, e.g., Chapter 5 of the lectures by Shapiro, Dentcheva, and Ruszczyński [159], or work on stochastic approximation by Nemirovski et al. [143].)

Example 8.1.1 (Support vector machines)**:** In linear classification problems, we observe pairs $z = (x, y)$, where $y \in \{-1, 1\}$ and $x \in \mathbb{R}^d$, and the goal is to find a parameter $\theta \in \mathbb{R}^d$ so that $\sign(\langle \theta, x \rangle) = y$. A convex loss surrogate for this problem is the hinge loss $\ell(\theta; z) = [1 - y\langle \theta, x \rangle]_+$; minimizing the associated risk functional (8.1.2) over a set $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le r\}$ gives the support vector machine [51]. $\diamond$

Example 8.1.2 (Two-stage stochastic programming)**:** In operations research, one often wishes to allocate resources to a set of locations $\{1, \ldots, m\}$ before seeing demand for the resources. Suppose that the (unobserved) sample x consists of the pair $x = (C, v)$, where $C \in \mathbb{R}^{m \times m}$ corresponds to the prices of shipping a unit of material, so $c_{ij} \ge 0$ gives the cost of shipping from location i to j, and $v \in \mathbb{R}^m$ denotes the value (price paid for the good) at each location. Letting $\theta \in \mathbb{R}^m_+$ denote the amount of resources allocated to each location, we formulate the loss as

$$\ell(\theta; x) := \inf_{r \in \mathbb{R}^m, T \in \mathbb{R}^{m \times m}} \left\{ \sum_{i,j} c_{ij} T_{ij} - \sum_{i=1}^m v_i r_i \mid r_i = \theta_i + \sum_{j=1}^m T_{ji} - \sum_{j=1}^m T_{ij}, \; T_{ij} \ge 0, \; \sum_{j=1}^m T_{ij} \le \theta_i \right\}.$$

Here the variables T correspond to the goods transported to and from each location (so T_{ij} is goods shipped from i to j), and we wish to minimize the cost of our shipping and maximize the profit. By minimizing the risk (8.1.2) over a set $\Theta = \{\theta \in \mathbb{R}^m_+ : \sum_i \theta_i \le b\}$, we maximize our expected reward given a budget constraint b on the amount of allocated resources. $\diamond$

For a (potentially random) estimator $\widehat{\theta} : \mathcal{X}^n \to \Theta$ given access to a sample $X_1, \ldots, X_n$, we may define the associated maximum *excess risk* for the family $\mathcal{P}$ by

$$\sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[L_P(\widehat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} L(\theta) \right\},$$

where the expectation is taken over X_i and any randomness in the procedure $\widehat{\theta}$. This expression captures the difference between the (expected) risk performance of the procedure $\widehat{\theta}$ and the best possible risk, available if the distribution P were known ahead of time. The *minimax excess risk*, defined with respect to the loss ℓ, domain Θ, and family $\mathcal{P}$ of distributions, is then defined by the best possible maximum excess risk,

$$\mathfrak{M}_n(\Theta, \mathcal{P}, \ell) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[L_P(\widehat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} L_P(\theta) \right\}, \tag{8.1.3}$$

where the infimum is taken over all estimators $\widehat{\theta} : \mathcal{X}^n \to \Theta$ and the risk L_P is implicitly defined in terms of the loss ℓ. The techniques for providing lower bounds for the minimax risk (8.1.1) or the excess risk (8.1.3) are essentially identical; we focus for the remainder of this section on techniques for providing lower bounds on the minimax risk.

8.2 Preliminaries on methods for lower bounds

There are a variety of techniques for providing lower bounds on the minimax risk (8.1.1). Each of them transforms the maximum risk by lower bounding it via a Bayesian problem (e.g. [110, 127, 130]), then proving a lower bound on the performance of all possible estimators for the Bayesian problem (it is often the case that the worst case Bayesian problem is equivalent to the original

minimax problem [127]). In particular, let $\{P_v\} \subset \mathcal{P}$ be a collection of distributions in $\mathcal{P}$ indexed by v and π be any probability mass function over v. Then for any estimator $\widehat{\theta}$, the maximum risk has lower bound

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi(\rho(\widehat{\theta}(X_1^n), \theta(P)))\right] \geq \sum_v \pi(v) \mathbb{E}_{P_v} \left[\Phi(\rho(\widehat{\theta}(X_1^n), \theta(P_v)))\right].$$

While trivial, this lower bound serves as the departure point for each of the subsequent techniques for lower bounding the minimax risk.

8.2.1 From estimation to testing

A standard first step in proving minimax bounds is to "reduce" the estimation problem to a testing problem [177, 175, 167]. The idea is to show that the probability of error in testing problems lower bounds the estimation risk, and we can develop tools for the former. We use two types of testing problems: one a multiple hypothesis test and the second based on multiple binary hypothesis tests.

Given an index set $\mathcal{V}$ of finite cardinality, consider a family of distributions $\{P_v\}_{v \in \mathcal{V}}$ contained within $\mathcal{P}$. This family induces a collection of parameters $\{\theta(P_v)\}_{v \in \mathcal{V}}$; we call the family a 2δ-packing in the ρ-semimetric if

$$\rho(\theta(P_v), \theta(P_{v'})) \geq 2\delta \quad \text{for all } v \neq v'.$$

We use this family to define the *canonical hypothesis testing problem*:

- first, nature chooses V according to the uniform distribution over $\mathcal{V}$;
- second, conditioned on the choice $V = v$, the random sample $X = X_1^n = (X_1, \ldots, X_n)$ is drawn from the n-fold product distribution P_v^n.

Given the observed sample X, the goal is to determine the value of the underlying index v. We refer to any measurable mapping $\Psi : \mathcal{X}^n \to \mathcal{V}$ as a test function. Its associated error probability is $\mathbb{P}(\Psi(X_1^n) \neq V)$, where $\mathbb{P}$ denotes the joint distribution over the random index V and X. In particular, if we set $\overline{P} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v$ to be the mixture distribution, then the sample X is drawn (marginally) from $\overline{P}$, and our hypothesis testing problem is to determine the randomly chosen index V given a sample from this mixture $\overline{P}$.

With this setup, we obtain the classical reduction from estimation to testing.

Proposition 8.2.1. *The minimax error* (8.1.1) *has lower bound*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}(\Psi(X_1, \ldots, X_n) \neq V), \tag{8.2.1}$$

where the infimum ranges over all testing functions.

Proof To see this result, fix an arbitrary estimator $\widehat{\theta}$. Suppressing dependence on X throughout the derivation, first note that it is clear that for any fixed θ, we have

$$\mathbb{E}[\Phi(\rho(\widehat{\theta}, \theta))] \geq \mathbb{E}\left[\Phi(\delta) \mathbf{1}\left\{\rho(\widehat{\theta}, \theta) \geq \delta\right\}\right] = \Phi(\delta) \mathbb{P}(\rho(\widehat{\theta}, \theta) \geq \delta),$$

where the final inequality follows because Φ is non-decreasing. Now, let us define $\theta_v = \theta(P_v)$, so that $\rho(\theta_v, \theta_{v'}) \geq 2\delta$ for $v \neq v'$. By defining the testing function

$$\Psi(\widehat{\theta}) := \operatorname*{argmin}_{v \in \mathcal{V}} \{\rho(\widehat{\theta}, \theta_v)\},$$

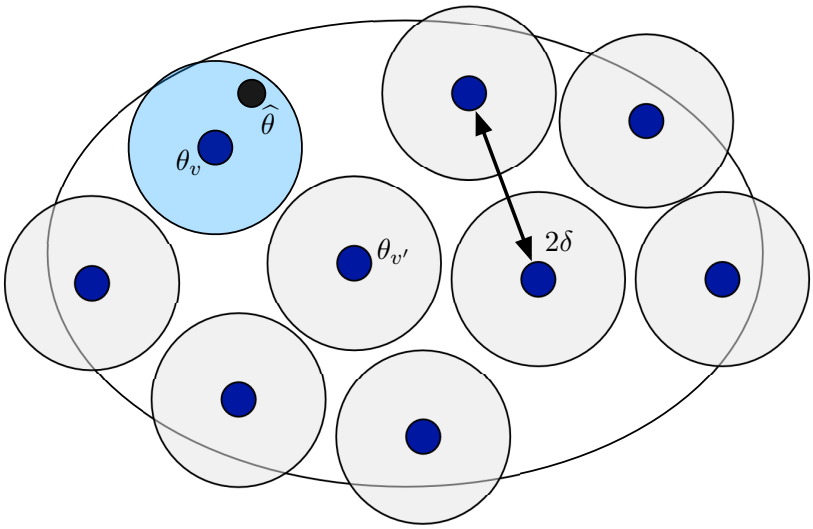

Figure 8.1. Example of a 2δ-packing of a set. The estimate $\widehat{\theta}$ is contained in at most one of the δ-balls around the points θ_v.

breaking ties arbitrarily, we have that $\rho(\widehat{\theta}, \theta_v) < \delta$ implies that $\Psi(\widehat{\theta}) = v$ because of the triangle inequality and 2δ-separation of the set $\{\theta_v\}_{v \in \mathcal{V}}$. Indeed, assume that $\rho(\widehat{\theta}, \theta_v) < \delta$; then for any $v' \neq v$, we have

$$\rho(\widehat{\theta}, \theta_{v'}) \geq \rho(\theta_v, \theta_{v'}) - \rho(\widehat{\theta}, \theta_v) > 2\delta - \delta = \delta.$$

The test must thus return v as claimed. Equivalently, for $v \in \mathcal{V}$, the inequality $\Psi(\widehat{\theta}) \neq v$ implies $\rho(\widehat{\theta}, \theta_v) \geq \delta$. (See Figure 8.1.) By averaging over $\mathcal{V}$, we find that

$$\sup_P \mathbb{P}(\rho(\widehat{\theta}, \theta(P)) \geq \delta) \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{P}(\rho(\widehat{\theta}, \theta(P_v)) \geq \delta \mid V = v) \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{P}(\Psi(\widehat{\theta}) \neq v \mid V = v).$$

Taking an infimum over all tests $\Psi : \mathcal{X}^n \to V$ gives inequality (8.2.1). □

The remaining challenge is to lower bound the probability of error in the underlying multi-way hypothesis testing problem, which we do by choosing the separation δ to trade off between the loss $\Phi(\delta)$ (large δ increases the loss) and the probability of error (small δ, and hence separation, makes the hypothesis test harder). Usually, one attempts to choose the largest separation δ that guarantees a constant probability of error. There are a variety of techniques for this, and we present three: Le Cam's method, Fano's method, and Assouad's method, including extensions of the latter two to enhance their applicability. Before continuing, however, we review some inequalities between divergence measures defined on probabilities, which will be essential for our development, and concepts related to packing sets (metric entropy, covering numbers, and packing).

8.2.2 Inequalities between divergences and product distributions

We now present a few inequalities, and their consequences when applied to product distributions, that will be quite useful for proving our lower bounds. The three divergences we relate are the total variation distance, Kullback-Leibler divergence, and Hellinger distance, all of which are instances

of f-divergences (recall Section 2.2.3). We first recall the definitions of the three when applied to distributions P, Q on a set $\mathcal{X}$, which we assume have densities p, q with respect to a base measure μ. Then we recall the total variation distance (2.2.6) is

$$\|P - Q\|_{\rm TV} := \sup_{A \subset \mathcal{X}} |P(A) - Q(A)| = \frac{1}{2} \int |p(x) - q(x)| d\mu(x),$$

which is the f-divergence $D_f(P\|Q)$ generated by $f(t) = \frac{1}{2}|t-1|$. The Hellinger distance (2.2.7) is

$$d_{\rm hel}(P, Q)^2 := \int (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x),$$

which is the f-divergence $D_f(P\|Q)$ generated by $f(t) = (\sqrt{t} - 1)^2$. We also recall the Kullback-Leibler (KL) divergence

$$D_{\rm kl}(P\|Q) := \int p(x) \log \frac{p(x)}{q(x)} d\mu(x), \tag{8.2.2}$$

which is the f-divergence $D_f(P\|Q)$ generated by $f(t) = t \log t$. As noted in Section 2.2.3, Proposition 2.2.8, these divergences have the following relationships.

Proposition (Proposition 2.2.8, restated)**.** *The total variation distance satisfies the following relationships:*

(a) For the Hellinger distance,

$$\frac{1}{2} d_{\rm hel}(P, Q)^2 \le \|P - Q\|_{\rm TV} \le d_{\rm hel}(P, Q) \sqrt{1 - d_{\rm hel}(P, Q)^2/4}.$$

(b) Pinsker's inequality: for any distributions P, Q,

$$\|P - Q\|_{\rm TV}^2 \le \frac{1}{2} D_{\rm kl}(P\|Q).$$

We now show how Proposition 2.2.8 is useful, because KL-divergence and Hellinger distance both are easier to manipulate on product distributions than is total variation. Specifically, consider the product distributions $P = P_1 \times \cdots \times P_n$ and $Q = Q_1 \times \cdots \times Q_n$. Then the KL-divergence satisfies the decoupling equality

$$D_{\rm kl}(P\|Q) = \sum_{i=1}^n D_{\rm kl}(P_i\|Q_i), \tag{8.2.3}$$

while the Hellinger distance satisfies

$$\begin{aligned} d_{\rm hel}(P, Q)^2 &= \int \left(\sqrt{p_1(x_1) \cdots p_n(x_n)} - \sqrt{q_1(x_1) \cdots q_n(x_n)}\right)^2 d\mu(x_1^n) \\ &= \int \left(\prod_{i=1}^n p_i(x_i) + \prod_{i=1}^n q_i(x_i) - 2\sqrt{p_1(x_1) \cdots p_n(x_n) q_1(x_n) \cdots q_n(x_n)}\right) d\mu(x_1^n) \\ &= 2 - 2 \prod_{i=1}^n \int \sqrt{p_i(x) q_i(x)} d\mu(x) = 2 - 2 \prod_{i=1}^n \left(1 - \frac{1}{2} d_{\rm hel}(P_i, Q_i)^2\right). \end{aligned} \tag{8.2.4}$$

In particular, we see that for product distributions P^n and Q^n, Proposition 2.2.8 implies that

$$\|P^n - Q^n\|_{\rm TV}^2 \le \frac{1}{2} D_{\rm kl}\left(P^n \| Q^n\right) = \frac{n}{2} D_{\rm kl}\left(P \| Q\right)$$

and

$$\|P^n - Q^n\|_{\rm TV} \le d_{\rm hel}(P^n, Q^n) \le \sqrt{2 - 2(1 - d_{\rm hel}(P, Q)^2)^n}.$$

As a consequence, if we can guarantee that $D_{\rm kl}\left(P \| Q\right) \le 1/n$ or $d_{\rm hel}(P, Q) \le 1/\sqrt{n}$, then we guarantee the strict inequality $\|P^n - Q^n\|_{\rm TV} \le 1 - c$ for a fixed constant $c > 0$, for any n. We will see how this type of guarantee can be used to prove minimax lower bounds in the following sections.

8.2.3 Metric entropy and packing numbers

The second part of proving our lower bounds involves the construction of the packing set in Section 8.2.1. The size of the space Θ of parameters associated with our estimation problem—and consequently, how many parameters we can pack into it—is strongly coupled with the difficulty of estimation. The tools we develop in Section 4.3.2 on metric entropies and covering and packing numbers therefore become central.

Probably the most central construction relies on volume bounds on packing and covering numbers, which we recall from Lemma 4.3.10: the covering and packing numbers of a norm ball $\mathbb{B}$ in its own norm $\|\cdot\|$ scale exponentially in the dimension. In particular, for any $\delta < 1$, there is a packing $\mathcal{V}$ of $\mathbb{B}$ such that $\|v - v'\| \ge \delta$ for all distinct $v, v' \in \mathcal{V}$ and $|\mathcal{V}| \ge (1/\delta)^d$, because we know $M(\delta, \mathbb{B}, \|\cdot\|) \ge N(\delta, \mathbb{B}, \|\cdot\|)$ as in Lemma 4.3.8. We thus state the following corollary for later use, which states that we can construct exponentially large packings of arbitrary norm-balls (in finite dimensions) where the points have constant distance from one another.

Corollary 8.2.2. *Let $\mathbb{B}^d = \{v \in \mathbb{R}^d \mid \|v\| \le 1\}$ be the unit ball for the norm $\|\cdot\|$. Then there exists $\mathcal{V} \subset \mathbb{B}^d$ with $|\mathcal{V}| \ge 2^d$ and $\|v - v'\| \ge \frac{1}{2}$ for each $v \ne v' \in \mathcal{V}$.*

Another common packing arises from coding theory, where the technique is to construct well-separated code-books ($\{0,1\}$-valued bit strings associated to individual symbols to be communicated) for communication. In showing our lower bounds, we show that even if a code-book is well-separated, it may still be hard to estimate. With that, we now demonstrate that there exist (exponentially) large packings of the d-dimensional hypercube of points that are $O(d)$-separated in the Hamming metric.

Lemma 8.2.3 (Gilbert-Varshamov bound)**.** *Let $d \ge 1$. There is a subset $\mathcal{V}$ of the d-dimensional hypercube $\mathcal{H}_d = \{-1, 1\}^d$ of size $|\mathcal{V}| \ge \exp(d/8)$ such that the ℓ_1-distance*

$$\left\|v - v'\right\|_1 = 2 \sum_{j=1}^{d} \mathbf{1}\left\{v_j \ne v'_j\right\} \ge \frac{d}{2}$$

for all $v \ne v'$ with $v, v' \in \mathcal{V}$.

Proof We use the proof of Guntuboyina [97]. Consider a maximal subset $\mathcal{V}$ of $\mathcal{H}_d = \{-1, 1\}^d$ satisfying

$$\left\|v - v'\right\|_1 \ge d/2 \;\; \text{for all distinct } v, v' \in \mathcal{V}. \tag{8.2.5}$$

That is, the addition of any vector $w \in \mathcal{H}_d, w \notin \mathcal{V}$ to $\mathcal{V}$ will break the constraint (8.2.5). This means that if we construct the closed balls $B(v, d/2) := \{w \in \mathcal{H}_d : \|v - w\|_1 \le d/2\}$, we must have

$$\bigcup_{v \in \mathcal{V}} B(v, d/2) = \mathcal{H}_d \quad \text{so} \quad |\mathcal{V}||B(0, d/2)| = \sum_{v \in \mathcal{V}} |B(v, d/2)| \ge 2^d. \tag{8.2.6}$$

We now upper bound the cardinality of $B(v, d/2)$ using the probabilistic method, which will imply the desired result. Let S_i, $i = 1, \ldots, d$, be i.i.d. Bernoulli $\{0, 1\}$-valued random variables. Then by their uniformity, for any $v \in \mathcal{H}_d$,

$$\begin{aligned} 2^{-d}|B(v, d/2)| &= \mathbb{P}(S_1 + S_2 + \ldots + S_d \le d/4) = \mathbb{P}(S_1 + S_2 + \ldots + S_d \ge 3d/4) \\ &\le \mathbb{E}\left[\exp(\lambda S_1 + \ldots + \lambda S_d)\right] \exp(-3\lambda d/4) \end{aligned}$$

for any $\lambda > 0$, by Markov's inequality (or the Chernoff bound). Since $\mathbb{E}[\exp(\lambda S_1)] = \frac{1}{2}(1 + e^\lambda)$, we obtain

$$2^{-d}|B(v, d/2)| \le \inf_{\lambda \ge 0} \left\{2^{-d}(1 + e^\lambda)^d \exp(-3\lambda d/4)\right\}$$

Choosing $\lambda = \log 3$, we have

$$|B(v, d/2)| \le 4^d \exp(-(3/4)d \log 3) = 3^{-3d/4} 4^d.$$

Recalling inequality (8.2.6), we have

$$|\mathcal{V}| 3^{-3d/4} 4^d \ge |\mathcal{V}||B(v, d/2)| \ge 2^d, \quad \text{or} \quad |\mathcal{V}| \ge \frac{3^{3d/4}}{2^d} = \exp\left(d\left[\frac{3}{4}\log 3 - \log 2\right]\right) \ge \exp(d/8),$$

as claimed. □

8.3 Le Cam's method

Le Cam's method, in its simplest form, provides lower bounds on the error in simple binary hypothesis testing testing problems. In this section, we explore this connection, showing the connection between hypothesis testing and total variation distance, and we then show how this can yield lower bounds on minimax error (or the optimal Bayes' risk) for simple—often one-dimensional—estimation problems.

In the first homework, we considered several representations of the total variation distance, including a question showing its relation to optimal testing. We begin again with this strand of thought, recalling the general testing problem discussed in Section 8.2.1. Suppose that we have a Bayesian hypothesis testing problem where V is chosen with equal probability to be 1 or 2, and given $V = v$, the sample X is drawn from the distribution P_v. Denoting by $\mathbb{P}$ the joint distribution of V and X, we have for any test $\Psi : \mathcal{X} \to \{1, 2\}$ that the probability of error is

$$\mathbb{P}(\Psi(X) \ne V) = \frac{1}{2} P_1(\Psi(X) \ne 1) + \frac{1}{2} P_2(\Psi(X) \ne 2).$$

Recalling Section 8.2.1, we note that Proposition 2.3.1 gives an exact representation of the testing error using total variation distance. In particular, we have

Proposition (Proposition 2.3.1, restated)**.** *For any distributions P_1 and P_2 on $\mathcal{X}$, we have*

$$\inf_{\Psi} \{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\} = 1 - \|P_1 - P_2\|_{\rm TV}, \tag{8.3.1}$$

where the infimum is taken over all tests $\Psi : \mathcal{X} \to \{1, 2\}$.

Returning to the setting in which we receive n i.i.d. observations $X_i \sim P$, when $V = 1$ with probability $\frac{1}{2}$ and 2 with probability $\frac{1}{2}$, we have

$$\inf_{\Psi} \mathbb{P}\left(\Psi(X_1, \ldots, X_n) \neq V\right) = \frac{1}{2} - \frac{1}{2}\left\|P_1^n - P_2^n\right\|_{\rm TV}. \tag{8.3.2}$$

The representations (8.3.1) and (8.3.2), in conjunction with our reduction of estimation to testing in Proposition 8.2.1, imply the following lower bound on minimax risk. For any family $\mathcal{P}$ of distributions for which there exists a pair $P_1, P_2 \in \mathcal{P}$ satisfying $\rho(\theta(P_1), \theta(P_2)) \geq 2\delta$, then the minimax risk after n observations has lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta)\left[\frac{1}{2} - \frac{1}{2}\left\|P_1^n - P_2^n\right\|_{\rm TV}\right]. \tag{8.3.3}$$

The lower bound (8.3.3) suggests the following strategy: we find distributions P_1 and P_2, which we choose as a function of δ, that guarantee $\|P_1^n - P_2^n\|_{\rm TV} \leq \frac{1}{2}$. In this case, so long as $\rho(\theta(P_1), \theta(P_2)) \geq 2\delta$, we have the lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta)\left[\frac{1}{2} - \frac{1}{2} \cdot \frac{1}{4}\right] = \frac{1}{4}\Phi(\delta).$$

We now give an example illustrating this idea.

Example 8.3.1 (Bernoulli mean estimation)**:** Consider the problem of estimating the mean $\theta \in [-1, 1]$ of a $\{\pm 1\}$-valued Bernoulli distribution under the squared error loss $(\theta - \widehat{\theta})^2$, where $X_i \in \{-1, 1\}$. In this case, by fixing some $\delta > 0$, we set $\mathcal{V} = \{-1, 1\}$, and we define P_v so that

$$P_v(X = 1) = \frac{1 + v\delta}{2} \quad \text{and} \quad P_v(X = -1) = \frac{1 - v\delta}{2},$$

whence we see that the mean $\theta(P_v) = \delta v$. Using the metric $\rho(\theta, \theta') = |\theta - \theta'|$ and loss $\Phi(\delta) = \delta^2$, we have separation 2δ of $\theta(P_{-1})$ and $\theta(P_1)$. Thus, via Le Cam's method (8.3.3), we have that

$$\mathfrak{M}_n(\mathsf{Bernoulli}([-1, 1]), (\cdot)^2) \geq \frac{1}{2}\delta^2\left(1 - \left\|P_{-1}^n - P_1^n\right\|_{\rm TV}\right).$$

We would thus like to upper bound $\|P_{-1}^n - P_1^n\|_{\rm TV}$ as a function of the separation δ and sample size n; here we use Pinsker's inequality (Proposition 2.2.8(a)) and the tensorization identity (8.2.3) that makes KL-divergence so useful. Indeed, we have

$$\left\|P_{-1}^n - P_1^n\right\|_{\rm TV}^2 \leq \frac{1}{2} D_{\rm kl}\left(P_{-1}^n \| P_1^n\right) = \frac{n}{2} D_{\rm kl}\left(P_{-1} \| P_1\right) = \frac{n}{2}\delta \log \frac{1 + \delta}{1 - \delta}.$$

Noting that $\delta \log \frac{1+\delta}{1-\delta} \leq 3\delta^2$ for $\delta \in [0, 1/2]$, we obtain that $\|P_{-1}^n - P_1^n\|_{\rm TV} \leq \delta\sqrt{3n/2}$ for $\delta \leq 1/2$. In particular, we can guarantee a high probability of error in the associated hypothesis testing problem (recall inequality (8.3.2)) by taking $\delta = 1/\sqrt{6n}$; this guarantees $\|P_{-1}^n - P_1^n\|_{\rm TV} \leq \frac{1}{2}$. We thus have the minimax lower bound

$$\mathfrak{M}_n(\mathsf{Bernoulli}([-1, 1]), (\cdot)^2) \geq \frac{1}{2}\delta^2\left(1 - \frac{1}{2}\right) = \frac{1}{24n}.$$

While the factor $1/24$ is smaller than necessary, this bound is optimal to within constant factors; the sample mean $(1/n)\sum_{i=1}^n X_i$ achieves mean-squared error $(1-\theta^2)/n$.
As an alternative proof, we may use the Hellinger distance and its associated decoupling identity (8.2.4). We sketch the idea, ignoring lower order terms when convenient. In this case, Proposition 2.2.7 implies

$$\|P_1^n - P_2^n\|_{\rm TV} \le \sqrt{2} d_{\rm hel}(P_1^n, P_2^n) = \sqrt{2 - 2(1 - d_{\rm hel}(P_1, P_2)^2)^n}.$$

Noting that

$$d_{\rm hel}(P_1, P_2)^2 = \left(\sqrt{\frac{1+\delta}{2}} - \sqrt{\frac{1-\delta}{2}}\right)^2 = 1 - 2\sqrt{\frac{1-\delta^2}{4}} = 1 - \sqrt{1-\delta^2} \approx \frac{1}{2}\delta^2,$$

and noting that $(1-\delta^2) \approx e^{-\delta^2}$, we have (up to lower order terms in δ) that $\|P_1^n - P_2^n\|_{\rm TV} \le \sqrt{2 - 2\exp(-\delta^2 n/2)}$. Choosing $\delta^2 = 1/(4n)$, we have $\sqrt{2 - 2\exp(-\delta^2 n/2)} \le 1/2$, thus giving the lower bound

$$\mathfrak{M}_n(\mathsf{Bernoulli}([-1,1]), (\cdot)^2) \text{ “}\ge\text{”} \frac{1}{2}\delta^2\left(1 - \frac{1}{2}\right) = \frac{1}{16n},$$

where the quotations indicate we have been fast and loose in the derivation. $\diamond$

This example shows the “usual” rate of convergence in parametric estimation problems, that is, that we can estimate a parameter θ at a rate (in squared error) scaling as $1/n$. The mean estimator above is, in some sense, the prototypical example of such regular problems. In some “irregular” scenarios—including estimating the support of a uniform random variable, which we study in the homework—faster rates are possible.

We also note in passing that their are substantially more complex versions of Le Cam’s method that can yield sharp results for a wider variety of problems, including some in nonparametric estimation [127, 177]. For our purposes, the simpler two-point perspective provided in this section will be sufficient.

JCD Comment: Talk about Euclidean structure with KL space and information geometry a bit here to suggest the KL approach later.

8.4 Fano’s method

Fano’s method, originally proposed by Has’minskii [100] for providing lower bounds in nonparametric estimation problems, gives a somewhat more general technique than Le Cam’s method, and it applies when the packing set $\mathcal{V}$ has cardinality larger than two. The method has played a central role in minimax theory, beginning with the pioneering work of Has’minskii and Ibragimov [100, 110]. More recent work following this initial push continues to the present day (e.g. [28, 177, 175, 29, 149, 97, 43]).

8.4.1 The classical (local) Fano method

We begin by stating Fano’s inequality, which provides a lower bound on the error in a multi-way hypothesis testing problem. Let V be a random variable taking values in a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| \ge 2$. If we let the function $h_2(p) = -p\log p - (1-p)\log(1-p)$ denote the entropy of the Bernoulli random variable with parameter p, Fano’s inequality (Proposition 2.3.3 from Chapter 2) takes the following form:

Proposition 8.4.1 (Fano inequality)**.** *For any Markov chain $V \to X \to \widehat{V}$, we have*

$$h_2(\mathbb{P}(\widehat{V} \neq V)) + \mathbb{P}(\widehat{V} \neq V) \log(|\mathcal{V}| - 1) \geq H(V \mid \widehat{V}). \tag{8.4.1}$$

Restating the results in Chapter 2, we also have the following convenient rewriting of Fano's inequality when V is uniform in $\mathcal{V}$ (recall Corollary 2.3.4).

Corollary 8.4.2. *Assume that V is uniform on $\mathcal{V}$. For any Markov chain $V \to X \to \widehat{V}$,*

$$\mathbb{P}(\widehat{V} \neq V) \geq 1 - \frac{I(V; X) + \log 2}{\log(|\mathcal{V}|)}. \tag{8.4.2}$$

In particular, Corollary 8.4.2 shows that we have

$$\inf_{\Psi} \mathbb{P}(\Psi(X) \neq V) \geq 1 - \frac{I(V; X) + \log 2}{\log |\mathcal{V}|},$$

where the infimum is taken over all testing procedures Ψ. By combining Corollary 8.4.2 with the reduction from estimation to testing in Proposition 8.2.1, we obtain the following result.

Proposition 8.4.3. *Let $\{\theta(P_v)\}_{v \in \mathcal{V}}$ be a 2δ-packing in the ρ-semimetric. Assume that V is uniform on the set $\mathcal{V}$, and conditional on $V = v$, we draw a sample $X \sim P_v$. Then the minimax risk has lower bound*

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta) \left(1 - \frac{I(V; X) + \log 2}{\log |\mathcal{V}|}\right).$$

To gain some intuition for Proposition 8.4.3, we think of the lower bound as a function of the separation $\delta > 0$. Roughly, as $\delta \downarrow 0$, the separation condition between the distributions P_v is relaxed and we expect the distributions P_v to be closer to one another. In this case—as will be made more explicity presently—the hypothesis testing problem of distinguishing the P_v becomes more challenging, and the information $I(V; X)$ shrinks. Thus, what we roughly attempt to do is to choose our packing $\theta(P_v)$ as a function of δ, and find the largest $\delta > 0$ making the mutual information small enough that

$$\frac{I(V; X) + \log 2}{\log |\mathcal{V}|} \leq \frac{1}{2}. \tag{8.4.3}$$

In this case, the minimax lower bound is at least $\Phi(\delta)/2$. We now explore techniques for achieving such results.

Mutual information and KL-divergence

Many techniques for upper bounding mutual information rely on its representation as the KL-divergence between multiple distributions. Indeed, given random variables V and X as in the preceding sections, if we let $P_{V,X}$ denote their joint distribution and P_V and P_X their marginals, then

$$I(V; X) = D_{\rm kl}\left(P_{X,V} \| P_X \times P_V\right),$$

where $P_X \times P_V$ denotes the distribution of (X, V) when the random variables are independent. By manipulating this definition, we can rewrite it into a form more convenient for our purposes.

Indeed, focusing on our setting of testing, let us assume that V is drawn from a prior distribution π (this may be a discrete or arbitrary distribution, though for simplicity we focus on the case when

π is discrete). Let P_v denote the distribution of X conditional on $V = v$, as in Proposition 8.4.3. Then marginally, we know that X is drawn from the mixture distribution

$$\overline{P} := \sum_v \pi(v) P_v.$$

With this definition of the mixture distribution, via algebraic manipulations, we have

$$I(V; X) = \sum_v \pi(v) D_{\rm kl}\left(P_v \| \overline{P}\right), \tag{8.4.4}$$

a representation that plays an important role in our subsequent derivations. To see equality (8.4.4), let μ be a base measure over $\mathcal{X}$ (assume w.l.o.g. that X has density $p(\cdot \mid v) = p_v(\cdot)$ conditional on $V = v$), and note that

$$I(V; X) = \sum_v \int_{\mathcal{X}} p(x \mid v)\pi(v) \log \frac{p(x \mid v)}{\sum_{v'} p(x \mid v')\pi(v')} d\mu(x) = \sum_v \pi(v) \int_{\mathcal{X}} p(x \mid v) \log \frac{p(x \mid v)}{\overline{p}(x)} d\mu(x).$$

Representation (8.4.4) makes it clear that if the distributions of the sample X conditional on V are all similar, then there is little information content. Returning to the discussion after Proposition 8.4.3, we have in this uniform setting that

$$\overline{P} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \quad \text{and} \quad I(V; X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\rm kl}\left(P_v \| \overline{P}\right).$$

The mutual information is small if the typical conditional distribution P_v is difficult to distinguish—has small KL-divergence—from $\overline{P}$.

The local Fano method

The local Fano method is based on a weakening of the mixture representation of mutual information (8.4.4), then giving a uniform upper bound on divergences between all pairs of the conditional distributions P_v and $P_{v'}$. (This method is known in the statistics literature as the "generalied Fano method," a poor name, as it is based on a weak upper bound on mutual information.) In particular (focusing on the case when V is uniform), the convexity of $-\log$ implies that

$$I(V; X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\rm kl}\left(P_v \| \overline{P}\right) \le \frac{1}{|\mathcal{V}|^2} \sum_{v, v'} D_{\rm kl}\left(P_v \| P_{v'}\right). \tag{8.4.5}$$

In the local Fano method approach, we construct a *local packing*. This local packing approach is based on constructing a family of distributions P_v for $v \in \mathcal{V}$ defining a 2δ-packing (recall Section 8.2.1), meaning that $\rho(\theta(P_v), \theta(P_{v'})) \ge 2\delta$ for all $v \ne v'$, but which additionally satisfy the uniform upper bound

$$D_{\rm kl}\left(P_v \| P_{v'}\right) \le \kappa^2 \delta^2 \quad \text{for all } v, v' \in \mathcal{V}, \tag{8.4.6}$$

where $\kappa > 0$ is a fixed problem-dependent constant. If we have the inequality (8.4.6), then so long as we can find a *local* packing $\mathcal{V}$ such that

$$\log |\mathcal{V}| \ge 2(\kappa^2 \delta^2 + \log 2),$$

we are guaranteed the testing error condition (8.4.3), and hence the minimax lower bound

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta).$$

The difficulty in this approach is constructing the packing set $\mathcal{V}$ that allows δ to be chosen to obtain sharp lower bounds, and we often require careful choices of the packing sets $\mathcal{V}$. (We will see how to reduce such difficulties in subsequent sections.)

Constructing local packings As mentioned above, the main difficulty in using Fano's method is in the construction of so-called "local" packings. In these problems, the idea is to construct a packing $\mathcal{V}$ of a fixed set (in a vector space, say $\mathbb{R}^d$) with constant radius and constant distance. Then we scale elements of the packing by $\delta > 0$, which leaves the cardinality $|\mathcal{V}|$ identical, but allows us to scale δ in the separation in the packing and the uniform divergence bound (8.4.6). In particular, Lemmas 8.2.3 and 4.3.10 show that we can construct exponentially large packings of certain sets with balls of a fixed radius.

We now illustrate these techniques via two examples.

Example 8.4.4 (Normal mean estimation)**:** Consider the d-dimensional normal location family $\mathcal{N}_d = \{\mathsf{N}(\theta, \sigma^2 I_{d\times d}) \mid \theta \in \mathbb{R}^d\}$; we wish to estimate the mean $\theta = \theta(P)$ of a given distribution $P \in \mathcal{N}_d$ in mean-squared error, that is, with loss $\|\widehat{\theta} - \theta\|_2^2$. Let $\mathcal{V}$ be a 1/2-packing of the unit ℓ_2-ball with cardinality at least 2^d, as guaranteed by Lemma 4.3.10. (We assume for simplicity that $d \geq 2$.)

Now we construct our local packing. Fix $\delta > 0$, and for each $v \in \mathcal{V}$, set $\theta_v = \delta v \in \mathbb{R}^d$. Then we have

$$\|\theta_v - \theta_{v'}\|_2 = \delta \left\|v - v'\right\|_2 \geq \frac{\delta}{2}$$

for each distinct pair $v, v' \in \mathcal{V}$, and moreover, we note that $\|\theta_v - \theta_{v'}\|_2 \leq \delta$ for such pairs as well. By applying the Fano minimax bound of Proposition 8.4.3, we see that (given n normal observations $X_i \stackrel{\text{iid}}{\sim} P$)

$$\mathfrak{M}_n(\theta(\mathcal{N}_d), \|\cdot\|_2^2) \geq \left(\frac{1}{2} \cdot \frac{\delta}{2}\right)^2 \left(1 - \frac{I(V; X_1^n) + \log 2}{\log |\mathcal{V}|}\right) = \frac{\delta^2}{16}\left(1 - \frac{I(V; X_1^n) + \log 2}{d \log 2}\right).$$

Now note that for any pair v, v', if P_v is the normal distribution $\mathsf{N}(\theta_v, \sigma^2 I_{d\times d})$ we have

$$D_{\rm kl}\left(P_v^n \| P_{v'}^n\right) = n \cdot D_{\rm kl}\left(\mathsf{N}(\delta v, \sigma^2 I_{d\times d}) \| \mathsf{N}(\delta v', \sigma^2 I_{d\times d})\right) = n \cdot \frac{\delta^2}{2\sigma^2}\left\|v - v'\right\|_2^2,$$

as the KL-divergence between two normal distributions with identical covariance is

$$D_{\rm kl}\left(\mathsf{N}(\theta_1, \Sigma) \| \mathsf{N}(\theta_2, \Sigma)\right) = \frac{1}{2}(\theta_1 - \theta_2)^\top \Sigma^{-1}(\theta_1 - \theta_2)$$

as in Example 2.1.7. As $\|v - v'\|_2 \leq 1$, we have the KL-divergence bound (8.4.6) with $\kappa^2 = n/2\sigma^2$.

Combining our derivations, we have the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}_d), \|\cdot\|_2^2) \geq \frac{\delta^2}{16}\left(1 - \frac{n\delta^2/2\sigma^2 + \log 2}{d \log 2}\right). \tag{8.4.7}$$

Then by taking $\delta^2 = d\sigma^2 \log 2/(2n)$, we see that

$$1 - \frac{n\delta^2/2\sigma^2 + \log 2}{d \log 2} = 1 - \frac{1}{d} - \frac{1}{4} \geq \frac{1}{4}$$

by assumption that $d \geq 2$, and inequality (8.4.7) implies the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}_d), \|\cdot\|_2^2) \geq \frac{d\sigma^2 \log 2}{32n} \cdot \frac{1}{4} \geq \frac{1}{185} \cdot \frac{d\sigma^2}{n}.$$

While the constant $1/185$ is not sharp, we do obtain the right scaling in d, n, and the variance σ^2; the sample mean attains the same risk. $\diamond$

Example 8.4.5 (Linear regression)**:** In this example, we show how local packings can give (up to some constant factors) sharp minimax rates for standard linear regression problems. In particular, for fixed matrix $X \in \mathbb{R}^{n \times d}$, we observe

$$Y = X\theta + \varepsilon,$$

where $\varepsilon \in \mathbb{R}^n$ consists of independent random variables ε_i with variance bounded by $\mathrm{Var}(\varepsilon_i) \leq \sigma^2$, and $\theta \in \mathbb{R}^d$ is allowed to vary over $\mathbb{R}^d$. For the purposes of our lower bound, we may assume that $\varepsilon \sim \mathsf{N}(0, \sigma^2 I_{n \times n})$. Let $\mathcal{P}$ denote the family of such normally distributed linear regression problems, and assume for simplicity that $d \geq 32$.
In this case, we use the Gilbert-Varshamov bound (Lemma 8.2.3) to construct a local packing and attain minimax rates. Indeed, let $\mathcal{V}$ be a packing of $\{-1, 1\}^d$ such that $\|v - v'\|_1 \geq d/2$ for distinct elements of $\mathcal{V}$, and let $|\mathcal{V}| \geq \exp(d/8)$ as guaranteed by the Gilbert-Varshamov bound. For fixed $\delta > 0$, if we set $\theta_v = \delta v$, then we have the packing guarantee for distinct elements v, v' that

$$\|\theta_v - \theta_{v'}\|_2^2 = \delta^2 \sum_{j=1}^d (v_j - v_j')^2 = 4\delta^2 \left\|v - v'\right\|_1 \geq 2d\delta^2.$$

Moreover, we have the upper bound

$$\begin{aligned} D_{\rm kl}\left(\mathsf{N}(X\theta_v, \sigma^2 I_{n\times n}) \| \mathsf{N}(X\theta_{v'}, \sigma^2 I_{n\times n})\right) &= \frac{1}{2\sigma^2} \|X(\theta_v - \theta_{v'})\|_2^2 \\ &\leq \frac{\delta^2}{2\sigma^2} \gamma_{\max}^2(X) \left\|v - v'\right\|_2^2 \leq \frac{2d}{\sigma^2} \gamma_{\max}^2(X) \delta^2, \end{aligned}$$

where $\gamma_{\max}(X)$ denotes the maximum singular value of X. Consequently, the bound (8.4.6) holds with $\kappa^2 \leq 2d\gamma_{\max}^2(X)/\sigma^2$, and we have the minimax lower bound

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|_2^2) \geq \frac{d\delta^2}{2} \left(1 - \frac{I(V;Y) + \log 2}{\log |\mathcal{V}|}\right) \geq \frac{d\delta^2}{2} \left(1 - \frac{\frac{2d\gamma_{\max}^2(X)}{\sigma^2}\delta^2 + \log 2}{d/8}\right).$$

Now, if we choose

$$\delta^2 = \frac{\sigma^2}{64\gamma_{\max}^2(X)}, \quad \text{then} \quad 1 - \frac{8\log 2}{d} - \frac{16d\gamma_{\max}^2(X)\delta^2}{d} \geq 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2},$$

by assumption that $d \geq 32$. In particular, we obtain the lower bound

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|_2^2) \geq \frac{1}{256} \frac{\sigma^2 d}{\gamma_{\max}^2(X)} = \frac{1}{256} \frac{\sigma^2 d}{n} \frac{1}{\gamma_{\max}^2(\frac{1}{\sqrt{n}} X)},$$

for a convergence rate (roughly) of $\sigma^2 d/n$ after rescaling the singular values of X by $1/\sqrt{n}$. This bound is sharp in terms of the dimension, dependence on n, and the variance σ^2, but it does not fully capture the dependence on X, as it depends only on the maximum singular value. Indeed, in this case, an exact calculation (cf. [130]) shows that the minimax value of the problem is exactly $\sigma^2 \operatorname{tr}((X^\top X)^{-1})$. Letting $\lambda_j(A)$ be the jth eigenvalue of a matrix A, we have

$$\sigma^2 \operatorname{tr}((X^\top X)^{-1}) = \frac{\sigma^2}{n} \operatorname{tr}((n^{-1} X^\top X)^{-1}) = \frac{\sigma^2}{n} \sum_{j=1}^{d} \frac{1}{\lambda_j(\frac{1}{n} X^\top X)} \\ \geq \frac{\sigma^2 d}{n} \min_j \frac{1}{\lambda_j(\frac{1}{n} X^\top X)} = \frac{\sigma^2 d}{n} \frac{1}{\gamma^2_{\max}(\frac{1}{\sqrt{n}} X)}.$$

Thus, the local Fano method captures most—but not all—of the difficulty of the problem. $\diamond$

8.4.2 A distance-based Fano method

While the testing lower bound (8.4.2) is sufficient for proving lower bounds for many estimation problems, for the sharpest results it sometimes requires a somewhat delicate construction of a well-separated packing (e.g. [43, 69]). To that end, we also provide extensions of inequalities (8.4.1) and (8.4.2) that more directly yield bounds on estimation error, allowing more direct and simpler proofs of a variety of minimax lower bounds (see also reference [67]).

More specifically, suppose that the distance function $\rho_{\mathcal{V}}$ is defined on $\mathcal{V}$, and we are interested in bounding the estimation error $\rho_{\mathcal{V}}(\widehat{V}, V)$. We begin by providing analogues of the lower bounds (8.4.1) and (8.4.2) that replace the testing error with the tail probability $\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t)$. By Markov's inequality, such control directly yields bounds on the expectation $\mathbb{E}[\rho_{\mathcal{V}}(\widehat{V}, V)]$. As we show in the sequel and in chapters to come, these distance-based Fano inequalities allow more direct proofs of a variety of minimax bounds without the need for careful construction of packing sets or metric entropy calculations as in other arguments.

We begin with the distance-based analogue of the usual discrete Fano inequality in Proposition 8.4.1. Let V be a random variable supported on a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| \geq 2$, and let $\rho : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a function defined on $\mathcal{V} \times \mathcal{V}$. In the usual setting, the function ρ is a metric on the space $\mathcal{V}$, but our theory applies to general functions. For a given scalar $t \geq 0$, the maximum and minimum *neighborhood sizes at radius* t are given by

$$N_t^{\max} := \max_{v \in \mathcal{V}} \left\{ \operatorname{card}\{v' \in \mathcal{V} \mid \rho(v, v') \leq t\} \right\} \quad \text{and} \quad N_t^{\min} := \min_{v \in \mathcal{V}} \left\{ \operatorname{card}\{v' \in \mathcal{V} \mid \rho(v, v') \leq t\} \right\}. \tag{8.4.8}$$

Defining the error probability $P_t = \mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t)$, we then have the following generalization of Fano's inequality:

Proposition 8.4.6. *For any Markov chain $V \to X \to \widehat{V}$, we have*

$$h_2(P_t) + P_t \log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}} + \log N_t^{\max} \geq H(V \mid \widehat{V}). \tag{8.4.9}$$

Before proving the proposition, which we do in Section 8.8.1, it is informative to note that it reduces to the standard form of Fano's inequality (8.4.1) in a special case. Suppose that we take $\rho_{\mathcal{V}}$ to be the 0-1 metric, meaning that $\rho_{\mathcal{V}}(v, v') = 0$ if $v = v'$ and 1 otherwise. Setting $t = 0$ in Proposition 8.4.6, we have $P_0 = \mathbb{P}[\widehat{V} \neq V]$ and $N_0^{\min} = N_0^{\max} = 1$, whence inequality (8.4.9) reduces

to inequality (8.4.1). Other weakenings allow somewhat clearer statements (see Section 8.8.2 for a proof):

Corollary 8.4.7. *If V is uniform on $\mathcal{V}$ and $(|\mathcal{V}| - N_t^{\min}) > N_t^{\max}$, then*

$$\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t) \geq 1 - \frac{I(V;X) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}. \tag{8.4.10}$$

Inequality (8.4.10) is the natural analogue of the classical mutual-information based form of Fano's inequality (8.4.2), and it provides a qualitatively similar bound. The main difference is that the usual cardinality $|\mathcal{V}|$ is replaced by the ratio $|\mathcal{V}|/N_t^{\max}$. This quantity serves as a rough measure of the number of possible "regions" in the space $\mathcal{V}$ that are distinguishable—that is, the number of subsets of $\mathcal{V}$ for which $\rho_{\mathcal{V}}(v, v') > t$ when v and v' belong to different regions. While this construction is similar in spirit to the usual construction of packing sets in the standard reduction from testing to estimation (cf. Section 8.2.1), our bound allows us to skip the packing set construction. We can directly compute $I(V;X)$ where V takes values over the full space, as opposed to computing the mutual information $I(V';X)$ for a random variable V' uniformly distributed over a packing set contained within $\mathcal{V}$. In some cases, the former calculation can be much simpler, as illustrated in examples and chapters to follow.

We now turn to providing a few consequences of Proposition 8.4.6 and Corollary 8.4.7, showing how they can be used to derive lower bounds on the minimax risk. Proposition 8.4.6 is a generalization of the classical Fano inequality (8.4.1), so it leads naturally to a generalization of the classical Fano lower bound on minimax risk, which we describe here. This reduction from estimation to testing is somewhat more general than the classical reductions, since we do not map the original estimation problem to a strict test, but rather a test that allows errors. Consider as in the standard reduction of estimation to testing in Section 8.2.1 a family of distributions $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ indexed by a finite set $\mathcal{V}$. This family induces an associated collection of parameters $\{\theta_v := \theta(P_v)\}_{v \in \mathcal{V}}$. Given a function $\rho_{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and a scalar t, we define the separation $\delta(t)$ of this set relative to the metric ρ on Θ via

$$\delta(t) := \sup \left\{ \delta \mid \rho(\theta_v, \theta_{v'}) \geq \delta \text{ for all } v, v' \in \mathcal{V} \text{ such that } \rho_{\mathcal{V}}(v, v') > t \right\}. \tag{8.4.11}$$

As a special case, when $t = 0$ and $\rho_{\mathcal{V}}$ is the discrete metric, this definition reduces to that of a packing set: we are guaranteed that $\rho(\theta_v, \theta_{v'}) \geq \delta(0)$ for all distinct pairs $v \neq v'$, as in the classical approach to minimax lower bounds. On the other hand, allowing for $t > 0$ lends greater flexibility to the construction, since only certain pairs θ_v and $\theta_{v'}$ are required to be well-separated.

Given a set $\mathcal{V}$ and associated separation function (8.4.11), we assume the canonical estimation setting: nature chooses $V \in \mathcal{V}$ uniformly at random, and conditioned on this choice $V = v$, a sample X is drawn from the distribution P_v. We then have the following corollary of Proposition 8.4.6, whose argument is completely identical to that for inequality (8.2.1):

Corollary 8.4.8. *Given V uniformly distributed over $\mathcal{V}$ with separation function $\delta(t)$, we have*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi\left(\frac{\delta(t)}{2}\right) \left[1 - \frac{I(X;V) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}\right] \quad \text{for all } t. \tag{8.4.12}$$

Notably, using the discrete metric $\rho_{\mathcal{V}}(v, v') = \mathbf{1}\{v \neq v'\}$ and taking $t = 0$ in the lower bound (8.4.12) gives the classical Fano lower bound on the minimax risk based on constructing a packing [110, 177, 175]. We now turn to an example illustrating the use of Corollary 8.4.8 in providing a minimax lower bound on the performance of regression estimators.

Example: Normal regression model Consider the d-dimensional linear regression model $Y = X\theta + \varepsilon$, where $\varepsilon \in \mathbb{R}^n$ is i.i.d. $\mathsf{N}(0, \sigma^2)$ and $X \in \mathbb{R}^{n \times d}$ is known, but θ is not. In this case, our family of distributions is

$$\mathcal{P}_X := \left\{Y \sim \mathsf{N}(X\theta, \sigma^2 I_{n\times n}) \mid \theta \in \mathbb{R}^d\right\} = \left\{Y = X\theta + \varepsilon \mid \varepsilon \sim \mathsf{N}(0, \sigma^2 I_{n\times n}), \theta \in \mathbb{R}^d\right\}.$$

We then obtain the following minimax lower bound on the minimax error in squared ℓ_2-norm: there is a universal (numerical) constant $c > 0$ such that

$$\mathfrak{M}_n(\theta(\mathcal{P}_X, \|\cdot\|_2^2) \ge c\frac{\sigma^2 d^2}{\|X\|_{\rm Fr}^2} \ge \frac{c}{\gamma_{\max}(X/\sqrt{n})^2} \cdot \frac{\sigma^2 d}{n}, \tag{8.4.13}$$

where $\gamma_{\max}$ denotes the maximum singular value. Notably, this inequality is nearly the sharpest known bound proved via Fano inequality-based methods [43], but our technique is essentially direct and straightforward.

To see inequality (8.4.13), let the set $\mathcal{V} = \{-1, 1\}^d$ be the d-dimensional hypercube, and define $\theta_v = \delta v$ for some fixed $\delta > 0$. Then letting $\rho_{\mathcal{V}}$ be the Hamming metric on $\mathcal{V}$ and ρ be the usual ℓ_2-norm, the associated separation function (8.4.11) satisfies $\delta(t) > \max\{\sqrt{t}, 1\}\delta$. Now, for any $t \le \lceil d/3 \rceil$, the neighborhood size satisfies

$$N_t^{\max} = \sum_{\tau=0}^{t} \binom{d}{\tau} \le 2\binom{d}{t} \le 2\left(\frac{de}{t}\right)^t.$$

Consequently, for $t \le d/6$, the ratio $|\mathcal{V}|/N_t^{\max}$ satisfies

$$\log \frac{|\mathcal{V}|}{N_t^{\max}} \ge d\log 2 - \log 2\binom{d}{t} \ge d\log 2 - \frac{d}{6}\log(6e) - \log 2 = d\log\frac{2}{2^{1/d}\sqrt[6]{6e}} > \max\left\{\frac{d}{6}, \log 4\right\}$$

for $d \ge 12$. (The case $2 \le d < 12$ can be checked directly). In particular, by taking $t = \lfloor d/6 \rfloor$ we obtain via Corollary 8.4.8 that

$$\mathfrak{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \ge \frac{\max\{\lfloor d/6 \rfloor, 2\}\delta^2}{4}\left(1 - \frac{I(Y;V) + \log 2}{\max\{d/6, 2\log 2\}}\right).$$

But of course, for V uniform on $\mathcal{V}$, we have $\mathbb{E}[VV^\top] = I_{d\times d}$, and thus for V, V' independent and uniform on $\mathcal{V}$,

$$\begin{aligned} I(Y;V) &\le n\frac{1}{|\mathcal{V}|^2}\sum_{v\in\mathcal{V}}\sum_{v'\in\mathcal{V}} D_{\rm kl}\left(\mathsf{N}(X\theta_v, \sigma^2 I_{n\times n}) \| \mathsf{N}(X\theta_{v'}, \sigma^2 I_{n\times n})\right) \\ &= \frac{\delta^2}{2\sigma^2}\mathbb{E}\left[\left\|XV - XV'\right\|_2^2\right] = \frac{\delta^2}{\sigma^2}\|X\|_{\rm Fr}^2. \end{aligned}$$

Substituting this into the preceding minimax bound, we obtain

$$\mathfrak{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \ge \frac{\max\{\lfloor d/6 \rfloor, 2\}\delta^2}{4}\left(1 - \frac{\delta^2\|X\|_{\rm Fr}^2/\sigma^2 + \log 2}{\max\{d/6, 2\log 2\}}\right).$$

Choosing $\delta^2 \asymp d\sigma^2/\|X\|_{\rm Fr}^2$ gives the result (8.4.13).

8.5 Assouad's method

Assouad's method provides a somewhat different technique for proving lower bounds. Instead of reducing the estimation problem to a multiple hypothesis test or simpler estimation problem, as with Le Cam's method and Fano's method from the preceding lectures, here we transform the original estimation problem into multiple binary hypothesis testing problems, using the structure of the problem in an essential way. Assouad's method applies only problems where the loss we care about is naturally related to identification of individual points on a hypercube.

8.5.1 Well-separated problems

To describe the method, we begin by encoding a notion of separation and loss, similar to what we did in the classical reduction of estimation to testing. For some $d \in \mathbb{N}$, let $\mathcal{V} = \{-1, 1\}^d$, and let us consider a family $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ indexed by the hypercube. We say that the the family P_v induces a 2δ-*Hamming separation* for the loss $\Phi \circ \rho$ if there exists a function $\widehat{\mathsf{v}} : \theta(\mathcal{P}) \to \{-1, 1\}^d$ satisfying

$$\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^{d} \mathbf{1}\left\{[\widehat{\mathsf{v}}(\theta)]_j \neq v_j\right\}. \tag{8.5.1}$$

That is, we can take the parameter θ and test the individual indices via $\widehat{\mathsf{v}}$.

Example 8.5.1 (Estimation in ℓ_1-error)**:** Suppose we have a family of multivariate Laplace distributions on $\mathbb{R}^d$—distributions with density proportional to $p(x) \propto \exp(-\left\|x - \mu\right\|_1)$—and we wish to estimate the mean in ℓ_1-distance. For $v \in \{-1, 1\}^d$ and some fixed $\delta > 0$ let p_v be the density

$$p_v(x) = \frac{1}{2} \exp\left(-\left\|x - \delta v\right\|_1\right),$$

which has mean $\theta(P_v) = \delta v$. Under the ℓ_1-loss, we have for any $\theta \in \mathbb{R}^d$ that

$$\left\|\theta - \theta(P_v)\right\|_1 = \sum_{j=1}^{d} \left|\theta_j - \delta v_j\right| \geq \delta \sum_{j=1}^{d} \mathbf{1}\left\{\operatorname{sign}(\theta_j) \neq v_j\right\},$$

so that this family induces a δ-Hamming separation for the ℓ_1-loss. $\diamond$

8.5.2 From estimation to multiple binary tests

As in the standard reduction from estimation to testing, we consider the following random process: nature chooses a vector $V \in \{-1, 1\}^d$ uniformly at random, after which the sample X is drawn from the distribution P_v conditional on $V = v$. Then, if we let $\mathbb{P}_{\pm j}$ denote the joint distribution over the random index V and X conditional on the jth coordinate $V_j = \pm 1$, we obtain the following sharper version of Assouad's lemma [10] (see also the paper [7]); we provide a proof in Section 8.8.3 to follow.

Lemma 8.5.2. *Under the conditions of the previous paragraph, we have*

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \inf_{\Psi} \left[\mathbb{P}_{+j}(\Psi(X) \neq +1) + \mathbb{P}_{-j}(\Psi(X) \neq -1)\right].$$

While Lemma 8.5.2 requires conditions on the loss Φ and metric ρ for the separation condition (8.5.1) to hold, it is sometimes easier to apply than Fano's method. Moreover, while we will not address this in class, several researchers [7, 68] have noted that it appears to allow easier application in so-called "interactive" settings—those for which the sampling of the X_i may not be precisely i.i.d. It is closely related to Le Cam's method, discussed previously, as we see that if we define $P_{+j} = 2^{1-d} \sum_{v:v_j=1} P_v$ (and similarly for $-j$), Lemma 8.5.2 is equivalent to

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \left[1 - \|P_{+j} - P_{-j}\|_{\mathrm{TV}}\right]. \tag{8.5.2}$$

There are standard weakenings of the lower bound (8.5.2) (and Lemma 8.5.2). We give one such weakening. First, we note that the total variation is convex, so that if we define $P_{v,+j}$ to be the distribution P_v where coordinate j takes the value $v_j = 1$ (and similarly for $P - v, -j$), we have

$$P_{+j} = \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} P_{v,+j} \quad \text{and} \quad P_{-j} = \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} P_{v,+j}.$$

Thus, by the triangle inequality, we have

$$\begin{aligned} \|P_{+j} - P_{-j}\|_{\mathrm{TV}} &= \left\| \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} P_{v,+j} - P_{v,-j} \right\|_{\mathrm{TV}} \\ &\leq \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}} \leq \max_{v,j} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}}. \end{aligned}$$

Then as long as the loss satisfies the per-coordinate separation (8.5.1), we obtain the following:

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq d\delta \left(1 - \max_{v,j} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}}\right). \tag{8.5.3}$$

This most common version of Assouad's lemma sometimes too brutally controls $\|P_{+j} - P_{-j}\|_{\mathrm{TV}}$.

We also note that by the Cauchy-Schwarz inequality and convexity of the variation-distance, we have

$$\sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{\mathrm{TV}} \leq \sqrt{d} \left(\sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2 \right)^{1/2} \leq \sqrt{d} \left(\sum_{j=1}^{d} \frac{1}{2^d} \sum_{v} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}}^2 \right)^{\frac{1}{2}},$$

and consequently we have a not quite so terribly weak version of inequality (8.5.2):

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta d \left[1 - \left(\frac{1}{d} \sum_{j=1}^{d} \sum_{v \in \{-1,1\}^d} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}}^2 \right)^{\frac{1}{2}}\right]. \tag{8.5.4}$$

Regardless of whether we use the sharper version (8.5.2) or weakened versions (8.5.3) or (8.5.4), the technique is essentially the same. We seek a setting of the distributions P_v so that the probability of making a mistake in the hypothesis test of Lemma 8.5.2 is high enough—say $1/2$—or the variation distance is small enough—such as $\|P_{+j} - P_{-j}\|_{\mathrm{TV}} \leq 1/2$ for all j. Once this is satisfied, we obtain a minimax lower bound of the form

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \left[1 - \frac{1}{2}\right] = \frac{d\delta}{2}.$$

8.5.3 Example applications of Assouad's method

We now provide two example applications of Assouad's method. The first is a standard finite-dimensional lower bound, where we provide a lower bound in a normal mean estimation problem. For the second, we consider estimation in a logistic regression problem, showing a similar lower bound. In Section 8.6 to follow, we show how to use Assouad's method to prove strong lower bounds in a standard nonparametric problem.

Example 8.5.3 (Normal mean estimation)**:** For some $\sigma^2 > 0$ and $d \in \mathbb{N}$, we consider estimation of mean parameter for the normal location family

$$\mathcal{N} := \left\{\mathsf{N}(\theta, \sigma^2 I_{d\times d}) : \theta \in \mathbb{R}^d\right\}$$

in squared Euclidean distance. We now show how for this family, the sharp Assouad's method implies the lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq \frac{d\sigma^2}{8n}. \tag{8.5.5}$$

Up to constant factors, this bound is sharp; the sample mean has mean squared error $d\sigma^2/n$. We proceed in (essentially) the usual way we have set up. Fix some $\delta > 0$ and define $\theta_v = \delta v$, taking $P_v = \mathsf{N}(\theta_v, \sigma^2 I_{d\times d})$ to be the normal distribution with mean θ_v. In this case, we see that the hypercube structure is natural, as our loss function decomposes on coordinates: we have $\|\theta - \theta_v\|_2^2 \geq \delta^2 \sum_{j=1}^d \mathbf{1}\left\{\mathrm{sign}(\theta_j) \neq v_j\right\}$. The family P_v thus induces a δ^2-Hamming separation for the loss $\|\cdot\|_2^2$, and by Assouad's method (8.5.2), we have

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq \frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\mathrm{TV}}\right],$$

where $P_{\pm j}^n = 2^{1-d} \sum_{v:v_j=\pm 1} P_v^n$. It remains to provide upper bounds on $\|P_{+j}^n - P_{-j}^n\|_{\mathrm{TV}}$. By the convexity of $\|\cdot\|_{\mathrm{TV}}^2$ and Pinsker's inequality, we have

$$\left\|P_{+j}^n - P_{-j}^n\right\|_{\mathrm{TV}}^2 \leq \max_{d_{\mathrm{ham}}(v,v')\leq 1} \|P_v^n - P_{v'}^n\|_{\mathrm{TV}}^2 \leq \frac{1}{2} \max_{d_{\mathrm{ham}}(v,v')\leq 1} D_{\mathrm{kl}}\left(P_v^n \| P_{v'}^n\right).$$

But of course, for any v and v' differing in only 1 coordinate,

$$D_{\mathrm{kl}}\left(P_v^n \| P_{v'}^n\right) = \frac{n}{2\sigma^2} \|\theta_v - \theta_{v'}\|_2^2 = \frac{2n}{\sigma^2}\delta^2,$$

giving the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq 2\delta^2 \sum_{j=1}^d \left[1 - \sqrt{2n\delta^2/\sigma^2}\right].$$

Choosing $\delta^2 = \sigma^2/8n$ gives the claimed lower bound (8.5.5). $\diamond$

Example 8.5.4 (Logistic regression)**:** In this example, consider the logistic regression model, where we have known (fixed) regressors $X_i \in \mathbb{R}^d$ and an unknown parameter $\theta \in \mathbb{R}^d$; the goal is to estimate θ after observing a sequence of $Y_i \in \{-1, 1\}$, where for $y \in \{-1, 1\}$ we have

$$P(Y_i = y \mid X_i, \theta) = \frac{1}{1 + \exp(-yX_i^\top \theta)}.$$

Denote this family by $\mathcal{P}_{\mathsf{log}}$, and for $P \in \mathcal{P}_{\mathsf{log}}$, let $\theta(P)$ be the predictor vector θ. We would like to estimate the vector θ in squared ℓ_2 error. As in Example 8.5.3, if we choose some $\delta > 0$ and for each $v \in \{-1,1\}^d$, we set $\theta_v = \delta v$, then we have the δ^2-separation in Hamming metric $\|\theta - \theta_v\|_2^2 \ge \delta^2 \sum_{j=1}^d \mathbf{1}\{\mathrm{sign}(\theta_j) \ne v_j\}$. Let P_v^n denote the distribution of the n independent observations Y_i when $\theta = \theta_v$. Then we have by Assouad's lemma (and the weakening (8.5.4)) that

$$\begin{aligned}\mathfrak{M}_n(\theta(\mathcal{P}_{\mathsf{log}}), \|\cdot\|_2^2) &\ge \frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\mathrm{TV}}\right] \\ &\ge \frac{d\delta^2}{2}\left[1 - \left(\frac{1}{d}\sum_{j=1}^d \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} \left\|P_{v,+j}^n - P_{v,-j}^n\right\|_{\mathrm{TV}}^2\right)^{\frac{1}{2}}\right]. \end{aligned} \tag{8.5.6}$$

It remains to bound $\|P_{v,+j}^n - P_{v,-j}^n\|_{\mathrm{TV}}^2$ to find our desired lower bound. To that end, use the shorthands $p_v(x) = 1/(1+\exp(\delta x^\top v))$ and let $D_{\mathrm{kl}}\left(p\|q\right)$ be the binary KL-divergence between $\mathsf{Bernoulli}(p)$ and $\mathsf{Bernoulli}(q)$ distributions. Then Pinsker's inequality (recall Proposition 2.2.8) implies that for any v, v',

$$\begin{aligned}\|P_v^n - P_{v'}^n\|_{\mathrm{TV}} &\le \frac{1}{4}[D_{\mathrm{kl}}\left(P_v^n \| P_{v'}^n\right) + D_{\mathrm{kl}}\left(P_{v'}^n \| P_v^n\right)] \\ &= \frac{1}{4}\sum_{i=1}^n \left[D_{\mathrm{kl}}\left(p_v(X_i) \| p_{v'}(X_i)\right) + D_{\mathrm{kl}}\left(p_{v'}(X_i) \| p_v(X_i)\right)\right].\end{aligned}$$

Let us upper bound the final KL-divergence. Let $p_a = 1/(1+e^a)$ and $p_b = 1/(1+e^b)$. We claim that

$$D_{\mathrm{kl}}\left(p_a \| p_b\right) + D_{\mathrm{kl}}\left(p_b \| p_a\right) \le (a-b)^2. \tag{8.5.7}$$

Deferring the proof of claim (8.5.7), we immediately see that

$$\|P_v^n - P_{v'}^n\|_{\mathrm{TV}} \le \frac{\delta^2}{4} \sum_{i=1}^n \left(X_i^\top (v - v')\right)^2.$$

Now we recall inequality (8.5.6) for motivation, and we see that the preceding display implies

$$\frac{1}{2^d d}\sum_{j=1}^d \sum_{v \in \{-1,1\}^d} \left\|P_{v,+j}^n - P_{v,-j}^n\right\|_{\mathrm{TV}}^2 \le \frac{\delta^2}{4d}\frac{1}{2^d} \sum_{v \in \{-1,1\}^d} \sum_{j=1}^d \sum_{i=1}^n (2X_{ij})^2 = \frac{\delta^2}{d}\sum_{i=1}^n \sum_{j=1}^d X_{ij}^2.$$

Replacing the final double sum with $\|X\|_{\mathrm{Fr}}^2$, where X is the matrix of the X_i, we have

$$\mathfrak{M}_n(\theta(\mathcal{P}_{\mathsf{log}}), \|\cdot\|_2^2) \ge \frac{d\delta^2}{2}\left[1 - \left(\frac{\delta^2}{d}\|X\|_{\mathrm{Fr}}^2\right)^{\frac{1}{2}}\right].$$

Setting $\delta^2 = d/4\,\|X\|_{\mathrm{Fr}}^2$, we obtain

$$\mathfrak{M}_n(\theta(\mathcal{P}_{\mathsf{log}}), \|\cdot\|_2^2) \ge \frac{d\delta^2}{4} = \frac{d^2}{16\,\|X\|_{\mathrm{Fr}}^2} = \frac{d}{n} \cdot \frac{1}{16\frac{1}{dn}\sum_{i=1}^n \|X_i\|_2^2}.$$

That is, we have a minimax lower bound scaling roughly as d/n for logistic regression, where "large" X_i (in ℓ_2-norm) suggest that we may obtain better performance in estimation. This is intuitive, as a larger X_i gives a better signal to noise ratio.

We return to prove the claim (8.5.7). Indeed, by a straightforward expansion, we have

$$\begin{aligned} D_{\rm kl}\left(p_a\|p_b\right) + D_{\rm kl}\left(p_b\|p_a\right) &= p_a \log \frac{p_a}{p_b} + (1-p_a)\log\frac{1-p_a}{1-p_b} + p_b \log\frac{p_b}{p_a} + (1-p_b)\log\frac{1-p_b}{1-p_a} \\ &= (p_a - p_b)\log\frac{p_a}{p_b} + (p_b - p_a)\log\frac{1-p_a}{1-p_b} = (p_a - p_b)\log\left(\frac{p_a}{1-p_a}\frac{1-p_b}{p_b}\right). \end{aligned}$$

Now note that $p_a/(1-p_a) = e^{-a}$ and $(1-p_b)/p_b = e^b$. Thus we obtain

$$D_{\rm kl}\left(p_a\|p_b\right) + D_{\rm kl}\left(p_b\|p_a\right) = \left(\frac{1}{1+e^a} - \frac{1}{1+e^b}\right)\log\left(e^{b-a}\right) = (b-a)\left(\frac{1}{1+e^a} - \frac{1}{1+e^b}\right)$$

Assume without loss of generality that $b \geq a$. Noting that $e^x \geq 1 + x$ by convexity, we have

$$\frac{1}{1+e^a} - \frac{1}{1+e^b} = \frac{e^b - e^a}{(1+e^a)(1+e^b)} \leq \frac{e^b - e^a}{e^b} = 1 - e^{a-b} \leq 1 - (1 + (a-b)) = b - a,$$

yielding claim (8.5.7). $\diamond$

8.6 Nonparametric regression: minimax upper and lower bounds

To show further applications of the minimax optimality ideas we have developed, we consider one of the two the most classical non-parametric (meaning that the number of parameters can grow with the sample size n) problems: estimating a regression function on a subset of the real line (the most classical problem being estimation of a density). In non-parametric regression, we assume there is an unknown function $f : \mathbb{R} \to \mathbb{R}$, where f belongs to a pre-determined class of functions $\mathcal{F}$; usually this class is parameterized by some type of smoothness guarantee. To make our problems concrete, we will assume that the unknown function f is L-Lipschitz and defined on $[0, 1]$. Let $\mathcal{F}$ denote this class.

In the standard non-parametric regression problem, we obtain observations of the form

$$Y_i = f(X_i) + \varepsilon_i \tag{8.6.1}$$

where ε_i are independent, mean zero conditional on X_i, and $\mathbb{E}[\varepsilon_i^2] \leq \sigma^2$. See Figure 8.2 for an example. We also assume that we fix the locations of the X_i as $X_i = i/n \in [0, 1]$, that is, the X_i are evenly spaced in $[0, 1]$. Given n observations Y_i, we ask two questions: (1) how can we estimate f? and (2) what are the optimal rates at which it is possible to estimate f?

8.6.1 Kernel estimates of the function

A natural strategy is to place small "bumps" around the observed points, and estimate f in a neighborhood of a point x by weighted averages of the Y values for other points near x. We now formalize a strategy for doing this. Suppose we have a kernel function $K : \mathbb{R} \to \mathbb{R}_+$, which is continuous, not identically zero, has support $\operatorname{supp} K = [-1, 1]$, and satisfies the technical condition

$$\lambda_0 \sup_x K(x) \leq \inf_{|x| \leq 1/2} K(x), \tag{8.6.2}$$

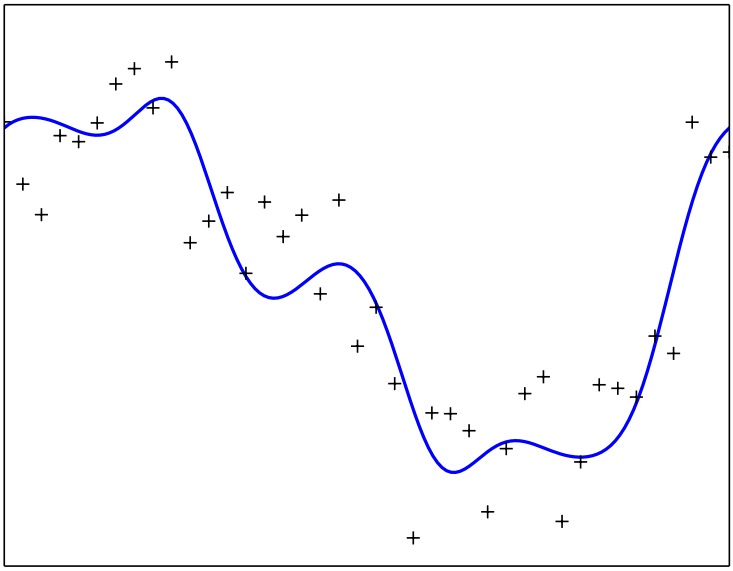

Figure 8.2. Observations in a non-parametric regression problem, with function f plotted. (Here $f(x) = \sin(2x + \cos^2(3x))$.)

where $\lambda_0 > 0$ (this says the kernel has some width to it). A natural example is the "tent" function given by $K_{\text{tent}}(x) = [1 - |x|]_+$, which satisfies inequality (8.6.2) with $\lambda_0 = 1/2$. See Fig. 8.3 for two examples, one the tent function and the other the function

$$K(x) = \mathbf{1}\left\{|x| < 1\right\} \exp\left(-\frac{1}{(x-1)^2}\right) \exp\left(-\frac{1}{(x+1)^2}\right),$$

which is infinitely differentiable and supported on $[-1, 1]$.

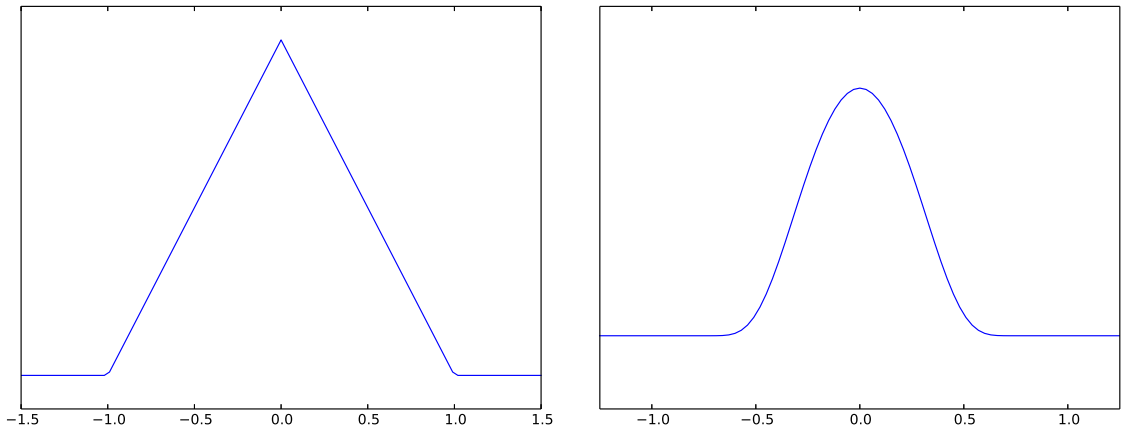

Figure 8.3: Left: "tent" kernel. Right: infinitely differentiable compactly supported kernel.

Now we consider a natural estimator of the function f based on observations (8.6.2) known as the Nadaraya-Watson estimator. Fix a bandwidth h, which we will see later smooths the estimated

functions f. For all x, define weights

$$W_{ni}(x) := \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)}$$

and define the estimated function

$$\widehat{f}_n(x) := \sum_{i=1}^n Y_i W_{ni}(x).$$

The intuition here is that we have a locally weighted regression function, where points X_i in the neighborhood of x are given higher weight than further points. Using this function $\widehat{f}_n$ as our estimator, it is possible to provide a guarantee on the bias and variance of the estimated function at each point $x \in [0,1]$.

Proposition 8.6.1. *Let the observation model (8.6.1) hold and assume condition (8.6.2). In addition assume the bandwidth is suitably large that $h \geq 2/n$ and that the X_i are evenly spaced on $[0,1]$. Then for any $x \in [0,1]$, we have*

$$|\mathbb{E}[\widehat{f}_n(x)] - f(x)| \leq Lh \quad \text{and} \quad \mathrm{Var}(\widehat{f}_n(x)) \leq \frac{2\sigma^2}{\lambda_0 nh}.$$

Proof To bound the bias, we note that (conditioning implicitly on X_i)

$$\mathbb{E}[\widehat{f}_n(x)] = \sum_{i=1}^n \mathbb{E}[Y_i W_{ni}(x)] = \sum_{i=1}^n \mathbb{E}[f(X_i)W_{ni}(x) + \varepsilon_i W_{ni}(x)] = \sum_{i=1}^n f(X_i) W_{ni}(x).$$

Thus we have that the bias is bounded as

$$\begin{aligned}\left|\mathbb{E}[\widehat{f}_n(x)] - f(x)\right| &\leq \sum_{i=1}^n |f(X_i) - f(x)| W_{ni}(x) \\ &\leq \sum_{i:|X_i - x| \leq h} |f(X_i) - f(x)| W_{ni}(x) \leq Lh \sum_{i=1}^n W_{ni}(x) = Lh.\end{aligned}$$

To bound the variance, we claim that

$$W_{ni}(x) \leq \min\left\{\frac{2}{\lambda_0 nh}, 1\right\}. \tag{8.6.3}$$

Indeed, we have that

$$W_{ni}(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)} = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j:|X_j - x| \leq h/2} K\left(\frac{X_j - x}{h}\right)} \leq \frac{K\left(\frac{X_i - x}{h}\right)}{\lambda_0 \sup_x K(x) |\{j : |X_j - x| \leq h/2\}|},$$

and because there are at least $nh/2$ indices satisfying $|X_j - x| \leq h$, we obtain the claim (8.6.3). Using the claim, we have

$$\begin{aligned}\mathrm{Var}(\widehat{f}_n(x)) = \mathbb{E}\left[\left(\sum_{i=1}^n (Y_i - f(X_i)) W_{ni}(x)\right)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n \varepsilon_i W_{ni}(x)\right)^2\right] \\ &= \sum_{i=1}^n W_{ni}(x)^2 \mathbb{E}[\varepsilon_i^2] \leq \sum_{i=1}^n \sigma^2 W_{ni}(x)^2.\end{aligned}$$

Noting that $W_{ni}(x) \le 2/\lambda_0 nh$ and $\sum_{i=1}^n W_{ni}(x) = 1$, we have

$$\sum_{i=1}^n \sigma^2 W_{ni}(x)^2 \le \sigma^2 \max_i W_{ni}(x) \underbrace{\sum_{i=1}^n W_{ni}(x)}_{=1} \le \sigma^2 \frac{2}{\lambda_0 nh},$$

completing the proof. □

With the proposition in place, we can then provide a theorem bounding the worst case pointwise mean squared error for estimation of a function $f \in \mathcal{F}$.

Theorem 8.6.2. *Under the conditions of Proposition 8.6.1, choose $h = (\sigma^2/L^2\lambda_0)^{1/3} n^{-1/3}$. Then there exists a universal (numerical) constant $C < \infty$ such that for any $f \in \mathcal{F}$,*

$$\sup_{x \in [0,1]} \mathbb{E}[(\widehat{f}_n(x) - f(x))^2] \le C \left(\frac{L\sigma^2}{\lambda_0}\right)^{2/3} n^{-\frac{2}{3}}.$$

Proof Using Proposition 8.6.1, we have for any $x \in [0,1]$ that

$$\mathbb{E}[(\widehat{f}_n(x) - f(x))^2] = \left(\mathbb{E}[\widehat{f}_n(x)] - f(x)\right)^2 + \mathbb{E}[(\widehat{f}_n(x) - \mathbb{E}[\widehat{f}_n(x)])^2] \le \frac{2\sigma^2}{\lambda_0 nh} + L^2 h^2.$$

Choosing h to balance the above bias/variance tradeoff, we obtain the thoerem. □

By integrating the result in Theorem 8.6.2 over the interval $[0,1]$, we immediately obtain the following corollary.

Corollary 8.6.3. *Under the conditions of Theorem 8.6.2, if we use the tent kernel K_{tent}, we have*

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f[\|\widehat{f}_n - f\|_2^2] \le C \left(\frac{L\sigma^2}{n}\right)^{2/3},$$

where C is a universal constant.

In Proposition 8.6.1, it is possible to show that a more clever choice of kernels—ones that are not always positive—can attain bias $\mathbb{E}[\widehat{f}_n(x)] - f(x) = \mathcal{O}(h^\beta)$ if f has Lipschitz $(\beta-1)$th derivative. In this case, we immediately obtain that the rate can be improved to

$$\sup_x \mathbb{E}[(\widehat{f}_n(x) - f(x))^2] \le C n^{-\frac{2\beta}{2\beta+1}},$$

and every additional degree of smoothness gives a corresponding improvement in convergence rate. We also remark that rates of this form, which are much larger than n^{-1}, are characteristic of non-parametric problems; essentially, we must adaptively choose a dimension that balances the sample size, so that rates of $1/n$ are difficult or impossible to achieve.

8.6.2 Minimax lower bounds on estimation with Assouad's method

Now we can ask whether the results we have given are in fact sharp; do there exist estimators attaining a faster rate of convergence than our kernel-based (locally weighted) estimator? Using Assouad's method, we show that, in fact, these results are all tight. In particular, we prove the following result on minimax estimation of a regression function $f \in \mathcal{F}$, where $\mathcal{F}$ consists of 1-Lipschitz functions defined on $[0,1]$, in the $\|\cdot\|_2^2$ error, that is, $\|f-g\|_2^2 = \int_0^1 (f(t)-g(t))^2 dt$.

Theorem 8.6.4. *Let the observation points X_i be spaced evenly on $[0,1]$, and assume the observation model (8.6.1). Then there exists a universal constant $c > 0$ such that*

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) := \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[\|\widehat{f}_n - f\|_2^2 \right] \geq c \left(\frac{\sigma^2}{n} \right)^{\frac{2}{3}}.$$

Deferring the proof of the theorem temporarily, we make a few remarks. It is in fact possible to show—using a completely identical technique—that if $\mathcal{F}_\beta$ denotes the class of functions with $\beta - 1$ derivatives, where the $(\beta-1)$th derivative is Lipschitz, then

$$\mathfrak{M}_n(\mathcal{F}_\beta, \|\cdot\|_2^2) \geq c \left(\frac{\sigma^2}{n} \right)^{\frac{2\beta}{2\beta+1}}.$$

So for any smoothness class, we can never achieve the parametric σ^2/n rate, but we can come arbitrarily close. As another remark, which we do not prove, in dimensions $d \geq 1$, the minimax rate for estimation of functions f with Lipschitz $(\beta-1)$th derivative scales as

$$\mathfrak{M}_n(\mathcal{F}_\beta, \|\cdot\|_2^2) \geq c \left(\frac{\sigma^2}{n} \right)^{\frac{2\beta}{2\beta+d}}. \tag{8.6.4}$$

This result can, similarly, be proved using a variant of Assouad's method or a local Fano method; see, for example, Györfi et al. [99, Chapter 3]. Exercise 8.9 works through a particular case of this lower bound. This is a striking example of the curse of dimensionality: the penalty for increasing dimension results in worse rates of convergence. For example, suppose that $\beta = 1$. In 1 dimension, we require $n \geq 90 \approx (.05)^{-3/2}$ observations to achieve accuracy .05 in estimation of f, while we require $n \geq 8000 = (.05)^{-(2+d)/2}$ even when the dimension $d = 4$, and $n \geq 64 \cdot 10^6$ observations even in 10 dimensions, which is a relatively small problem. That is, the problem is made exponentially more difficult by dimension increases.

We now prove Theorem 8.6.4. To establish the result, we show how to construct a family of problems—indexed by binary vectors $v \in \{-1,1\}^k$—so that our estimation problem satisfies the separation (8.5.1), then we show that the information based on observing noisy versions of the functions we have defined is small. Choosing k to make our resulting lower bound as high as possible completes the argument.

Construction of a separated family of functions To construct our separation in Hamming metric, as required by Eq. (8.5.1), fix some $k \in \mathbb{N}$; we will choose k later. This approach is somewhat different from our standard approach of using a fixed dimensionality and scaling the separation directly; in non-parametric problems, we scale the "dimension" itself to adjust the difficulty of the estimation problem. Define the function $g(x) = [1/2 - |x - 1/2|]_+$, so that g is 1-Lipschitz and is

0 outside of the interval $[0, 1]$. Then for any $v \in \{-1, 1\}^k$, define the "bump" functions

$$g_j(x) := \frac{1}{k} g\left(k\left(x - \frac{j-1}{k}\right)\right) \quad \text{and} \quad f_v(x) := \sum_{j=1}^k v_j g_j(x),$$

which we see is 1-Lipschitz. Now, consider any function $f : [0, 1] \to \mathbb{R}$, and let E_j be shorthand for the intervals $E_j = [(j-1)/k, j/k]$ for $j = 1, \ldots, k$. We must find a mapping identifying a function f with points in the hypercube $\{-1, 1\}^k$. To that end, we may define a vector $\widehat{\mathsf{v}}(f) \in \{-1, 1\}^k$ by

$$\widehat{\mathsf{v}}_j(f) = \operatorname*{argmin}_{s \in \{-1,1\}} \int_{E_j} (f(t) - s g_j(t))^2 \, dt.$$

We claim that for any function f,

$$\left(\int_{E_j} (f(t) - f_v(t))^2 dt\right)^{\frac{1}{2}} \geq \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\} \left(\int_{E_j} f_v(t)^2 dt\right)^{\frac{1}{2}}. \tag{8.6.5}$$

Indeed, on the set E_j, we have $v_j g_j(t) = f_v(t)$, and thus $\int_{E_j} g_j(t)^2 dt = \int_{E_j} f_v(t)^2 dt$. Then by the triangle inequality, we have

$$\begin{aligned}
2 \cdot \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\} \left(\int_{E_j} g_j(t)^2 dt\right)^{\frac{1}{2}} &= \left(\int_{E_j} \left((\widehat{\mathsf{v}}_j(f) - v_j) g_j(t)\right)^2 dt\right)^{\frac{1}{2}} \\
&\leq \left(\int_{E_j} \left(f(t) - v_j g_j(t)\right)^2 dt\right)^{\frac{1}{2}} + \left(\int_{E_j} \left(f(t) - \widehat{\mathsf{v}}_j(f) g_j(t)\right)^2 dt\right)^{\frac{1}{2}} \\
&\leq 2\left(\int_{E_j} \left(f(t) - f_v(t)\right)^2 dt\right)^{\frac{1}{2}},
\end{aligned}$$

by definition of the sign $\widehat{\mathsf{v}}_j(f)$.

With the definition of $\widehat{\mathsf{v}}$ and inequality (8.6.5), we see that for any vector $v \in \{-1, 1\}^k$, we have

$$\|f - f_v\|_2^2 = \sum_{j=1}^k \int_{E_j} \left(f(t) - f_v(t)\right)^2 dt \geq \sum_{j=1}^k \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\} \int_{E_j} f_v(t)^2 dt.$$

In particular, we know that

$$\int_{E_j} f_v(t)^2 dt = \frac{1}{k^2} \int_0^{1/k} g(kt)^2 dt = \frac{1}{k^3} \int_0^1 g(u)^2 du \geq \frac{c}{k^3},$$

where c is a numerical constant. In particular, we have the desired separation

$$\|f - f_v\|_2^2 \geq \frac{c}{k^3} \sum_{j=1}^k \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\}. \tag{8.6.6}$$

Bounding the binary testing error Let P_v^n denote the distribution of the n observations $Y_i = f_v(X_i) + \varepsilon_i$ when f_v is the true regression function. Then inequality (8.6.6) implies via Assouad's lemma that

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) \geq \frac{c}{k^3} \sum_{j=1}^{k} \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\rm TV}\right]. \tag{8.6.7}$$

Now, we use convexity and Pinsker's inequality to note that

$$\left\|P_{+j}^n - P_{-j}^n\right\|_{\rm TV}^2 \leq \max_v \left\|P_{v,+j}^n - P_{v,-j}^n\right\|_{\rm TV}^2 \leq \max_v \frac{1}{2} D_{\rm kl}\left(P_{v,+j}^n \| P_{v,-j}^n\right).$$

For any two functions f_v and $f_{v'}$, we have that the observations Y_i are independent and normal with means $f_v(X_i)$ or $f_{v'}(X_i)$, respectively. Thus

$$\begin{aligned} D_{\rm kl}\left(P_v^n \| P_{v'}^n\right) &= \sum_{i=1}^{n} D_{\rm kl}\left(\mathsf{N}(f_v(X_i), \sigma^2) \| \mathsf{N}(f_{v'}(X_i), \sigma^2)\right) \\ &= \sum_{i=1}^{n} \frac{1}{2\sigma^2}(f_v(X_i) - f_{v'}(X_i))^2. \end{aligned} \tag{8.6.8}$$

Now we must show that the expression (8.6.8) scales more slowly than n, which we will see must be the case as whenever $d_{\rm ham}(v, v') \leq 1$. Intuitively, most of the observations have the same distribution by our construction of the f_v as bump functions; let us make this rigorous.

We may assume without loss of generality that $v_j = v'_j$ for $j > 1$. As the $X_i = i/n$, we thus have that only X_i for i near 1 can have non-zero values in the tensorization (8.6.8). In particular,

$$f_v(i/n) = f_{v'}(i/n) \;\; \text{for all } i \text{ s.t. } \frac{i}{n} \geq \frac{2}{k}, \;\; \text{i.e. } \; i \geq \frac{2n}{k}.$$

Rewriting expression (8.6.8), then, and noting that $f_v(x) \in [-1/k, 1/k]$ for all x by construction, we have

$$\sum_{i=1}^{n} \frac{1}{2\sigma^2}(f_v(X_i) - f_{v'}(X_i))^2 \leq \sum_{i=1}^{2n/k} \frac{1}{2\sigma^2}(f_v(X_i) - f_{v'}(X_i))^2 \leq \frac{1}{2\sigma^2}\frac{2n}{k}\frac{1}{k^2} = \frac{n}{k^3\sigma^2}.$$

Combining this with inequality (8.6.8) and the minimax bound (8.6.7), we obtain

$$\left\|P_{+j}^n - P_{-j}^n\right\|_{\rm TV} \leq \sqrt{\frac{n}{2k^3\sigma^2}},$$

so

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) \geq \frac{c}{k^3} \sum_{j=1}^{k} \left[1 - \sqrt{\frac{n}{2k^3\sigma^2}}\right].$$

Choosing k for optimal tradeoffs Now we simply choose k; in particular, setting

$$k = \left\lceil \left(\frac{n}{2\sigma^2}\right)^{1/3} \right\rceil \quad \text{then} \quad 1 - \sqrt{\frac{n}{2k^3\sigma^2}} \geq 1 - \sqrt{1/4} = \frac{1}{2},$$

and we arrive at

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) \geq \frac{c}{k^3} \sum_{j=1}^{k} \frac{1}{2} = \frac{c}{2k^2} \geq c' \left(\frac{\sigma^2}{n}\right)^{2/3},$$

where $c' > 0$ is a universal constant. Theorem 8.6.4 follows.

8.7 Global Fano Method

In this section, we extend the techniques of Section 8.4 on Fano's method (the local Fano method) to a more global construction. In particular, we show that, rather than constructing a local packing, choosing a scaling $\delta > 0$, and then optimizing over this δ, it is actually, in many cases, possible to prove lower bounds on minimax error directly using packing and covering numbers (metric entropy and packing entropy).

8.7.1 A mutual information bound based on metric entropy

To begin, we recall the classical Fano inequality in Corollary 8.4.2, which says that for any Markov chain $V \to X \to \widehat{V}$, where V is uniform on the finite set $\mathcal{V}$, we have

$$\mathbb{P}(\widehat{V} \neq V) \geq 1 - \frac{I(V;X) + \log 2}{\log(|\mathcal{V}|)}.$$

Thus, there are two ingredients in proving lower bounds on the error in a hypothesis test: upper bounding the mutual information and lower bounding the size $|\mathcal{V}|$. The key in the global Fano method is an upper bound on the former (the information $I(V;X)$) using covering numbers.

Before stating our result, we require a bit of notation. First, we assume that V is drawn from a distribution μ, and conditional on $V = v$, assume the sample $X \sim P_v$. Then a standard calculation (or simply the definition of mutual information; recall equation (8.4.4)) gives that

$$I(V;X) = \int D_{\rm kl}\left(P_v \| \overline{P}\right) d\mu(v), \quad \text{where} \quad \overline{P} = \int P_v d\mu(v).$$

Now, we show how to connect this mutual information quantity to a covering number of a set of distributions.

Assume that for all v, we have $P_v \in \mathcal{P}$, where $\mathcal{P}$ is a collection of distributions. In analogy with Definition 4.7, we say that the collection of distributions $\{Q_i\}_{i=1}^N$ form an ϵ-cover of $\mathcal{P}$ in KL-divergence if for all $P \in \mathcal{P}$, there exists some i such that $D_{\rm kl}\left(P \| Q_i\right) \leq \epsilon^2$. With this, we may define the KL-covering number of the set $\mathcal{P}$ as

$$N_{\rm kl}\left(\epsilon, \mathcal{P}\right) := \inf\left\{N \in \mathbb{N} \mid \exists\, Q_i, i = 1, \ldots, N, \; \sup_{P \in \mathcal{P}} \min_i D_{\rm kl}\left(P \| Q_i\right) \leq \epsilon^2\right\}, \tag{8.7.1}$$

where $N_{\rm kl}\left(\epsilon, \mathcal{P}\right) = +\infty$ if no such cover exists. With definition (8.7.1) in place, we have the following proposition.

Proposition 8.7.1. *Under conditions of the preceding paragraphs, we have*

$$I(V;X) \leq \inf_{\epsilon > 0}\left\{\epsilon^2 + \log N_{\rm kl}\left(\epsilon, \mathcal{P}\right)\right\}. \tag{8.7.2}$$

Proof First, we claim that

$$\int D_{\rm kl}\left(P_v \| \overline{P}\right) d\mu(v) \leq \int D_{\rm kl}\left(P_v \| Q\right) d\mu(v) \tag{8.7.3}$$

for any distribution Q. Indeed, we have

$$\begin{aligned}\int D_{\rm kl}\left(P_v \| \overline{P}\right) d\mu(v) &= \int_{\mathcal{V}} \int_{\mathcal{X}} dP_v \log \frac{dP_v}{d\overline{P}} d\mu(v) = \int_{\mathcal{V}} \int_{\mathcal{X}} dP_v \left[\log \frac{dP_v}{Q} + \log \frac{dQ}{d\overline{P}}\right] d\mu(v) \\ &= \int_{\mathcal{V}} D_{\rm kl}\left(P_v \| Q\right) d\mu(v) + \int_{\mathcal{X}} \underbrace{\int_{\mathcal{V}} d\mu(v) dP_v}_{= d\overline{P}} \log \frac{dQ}{d\overline{P}} \\ &= \int D_{\rm kl}\left(P_v \| Q\right) d\mu(v) - D_{\rm kl}\left(\overline{P} \| Q\right) \le \int D_{\rm kl}\left(P_v \| Q\right) d\mu(v),\end{aligned}$$

so that inequality (8.7.3) holds. By carefully choosing the distribution Q in the upper bound (8.7.3), we obtain the proposition.

Now, assume that the distributions Q_i, $i = 1, \ldots, N$ form an ϵ^2-cover of the family $\mathcal{P}$, meaning that

$$\min_{i \in [N]} D_{\rm kl}\left(P \| Q_i\right) \le \epsilon^2 \quad \text{for all } P \in \mathcal{P}.$$

Let p_v and q_i denote the densities of P_v and Q_i with respect to some fixed base measure on $\mathcal{X}$ (the choice of based measure does not matter). Then defininining the distribution $Q = (1/N) \sum_{i=1}^N Q_i$, we obtain for any v that in expectation over $X \sim P_v$,

$$\begin{aligned}D_{\rm kl}\left(P_v \| Q\right) &= \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{q(X)}\right] = \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{N^{-1} \sum_{i=1}^n q_i(X)}\right] \\ &= \log N + \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{\sum_{i=1}^N q_i(X)}\right] \le \log N + \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{\max_i q_i(X)}\right] \\ &\le \log N + \min_i \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{q_i(X)}\right] = \log N + \min_i D_{\rm kl}\left(P_v \| Q_i\right).\end{aligned}$$

By our assumption that the Q_i form a cover, this gives the desired result, as $\epsilon \ge 0$ was arbitrary, as was our choice of the cover. □

By a completely parallel proof, we also immediately obtain the following corollary.

Corollary 8.7.2. *Assume that $X_1, \ldots, X_n$ are drawn i.i.d. from P_v conditional on $V = v$. Let $N_{\rm kl}\left(\epsilon, \mathcal{P}\right)$ denote the KL-covering number of a collection $\mathcal{P}$ containing the distributions (over a single observation) P_v for all $v \in \mathcal{V}$. Then*

$$I(V; X_1, \ldots, X_n) \le \inf_{\epsilon \ge 0} \left\{n\epsilon^2 + \log N_{\rm kl}\left(\epsilon, \mathcal{P}\right)\right\}.$$

With Corollary 8.7.2 and Proposition 8.7.1 in place, we thus see that the global covering numbers in KL-divergence govern the behavior of information.

We remark in passing that the quantity (8.7.2), and its i.i.d. analogue in Corollary 8.7.2, is known as the *index of resolvability*, and it controls estimation rates and redundancy of coding schemes for unknown distributions in a variety of scenarios; see, for example, Barron [17] and Barron and Cover [18]. It is also similar to notions of complexity in Dudley's entropy integral (cf. Dudley [71]) in empirical process theory, where the fluctuations of an empirical process are governed by a tradeoff between covering number and approximation of individual terms in the process.

8.7.2 Minimax bounds using global packings

There is now a four step process to proving minimax lower bounds using the global Fano method. Our starting point is to recall the Fano minimax lower bound in Proposition 8.4.3, which begins with the construction of a set of points $\{\theta(P_v)\}_{v\in\mathcal{V}}$ that form a 2δ-packing of a set Θ in some ρ-semimetric. With this inequality in mind, we perform the following four steps:

(i) *Bound the packing entropy.* Give a lower bound on the packing number of the set Θ with 2δ-separation (call this lower bound $M(\delta)$).

(ii) *Bound the metric entropy.* Give an upper bound on the KL-metric entropy of the class $\mathcal{P}$ of distributions containing all the distributions P_v, that is, an upper bound on $\log N_{\rm kl}(\epsilon, \mathcal{P})$.

(iii) *Find the critical radius.* Noting as in Corollary 8.7.2 that with n i.i.d. observations, we have

$$I(V; X_1, \ldots, X_n) \le \inf_{\epsilon \ge 0} \left\{ n\epsilon^2 + \log N_{\rm kl}(\epsilon, \mathcal{P}) \right\},$$

we now balance the information $I(V; X_1^n)$ and the packing entropy $\log M(\delta)$. To that end, we choose ϵ_n and $\delta > 0$ at the *critical radius*, defined as follows: choose the any ϵ_n such that

$$n\epsilon_n^2 \ge \log N_{\rm kl}(\epsilon_n, \mathcal{P}),$$

and choose the largest $\delta_n > 0$ such that

$$\log M(\delta_n) \ge 4n\epsilon_n^2 + 2\log 2 \ge 2N_{\rm kl}(\epsilon_n, \mathcal{P}) + 2n\epsilon_n^2 + 2\log 2 \ge 2\left(I(V; X_1^n) + \log 2\right).$$

(We could have chosen the ϵ_n attaining the infimum in the mutual information, but this way we need only an upper bound on $\log N_{\rm kl}(\epsilon, \mathcal{P})$.)

(iv) *Apply the Fano minimax bound.* Having chosen δ_n and ϵ_n as above, we immediately obtain that for the Markov chain $V \to X_1^n \to \widehat{V}$,

$$\mathbb{P}(V \neq \widehat{V}) \ge 1 - \frac{I(V; X_1, \ldots, X_n) + \log 2}{\log M(\delta_n)} \ge 1 - \frac{1}{2} = \frac{1}{2},$$

and thus, applying the Fano minimax bound in Proposition 8.4.3, we obtain

$$\mathfrak{M}_n(\theta(\mathcal{P}); \Phi \circ \rho) \ge \frac{1}{2}\Phi(\delta_n).$$

8.7.3 Example: non-parametric regression

In this section, we flesh out the outline in the prequel to show how to obtain a minimax lower bound for a non-parametric regression problem directly with packing and metric entropies. In this example, we sketch the result, leaving explicit constant calculations to the dedicated reader. Nonetheless, we recover an analogue of Theorem 8.6.4 on minimax risks for estimation of 1-Lipschitz functions on $[0, 1]$.

We use the standard non-parametric regression setting, where our observations Y_i follow the independent noise model (8.6.1), that is, $Y_i = f(X_i) + \varepsilon_i$. Letting

$$\mathcal{F} := \{f : [0,1] \to \mathbb{R}, \ f(0) = 0, \ f \text{ is Lipschitz}\}$$

be the family of 1-Lipschitz functions with $f(0) = 0$, we have

Proposition 8.7.3. *There exists a universal constant $c > 0$ such that*

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_\infty) := \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[\|\widehat{f}_n - f\|_\infty \right] \geq c \left(\frac{\sigma^2}{n} \right)^{1/3},$$

where $\widehat{f}_n$ is constructed based on the n independent observations $f(X_i) + \varepsilon_i$.

The rate in Proposition 8.7.3 is sharp to within factors logarithmic in n; a more precise analysis of the upper and lower bounds on the minimax rate yields

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_\infty) := \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[\|\widehat{f}_n - f\|_\infty \right] \asymp \left(\frac{\sigma^2 \log n}{n} \right)^{1/3}.$$

See, for example, Tsybakov [167] for a proof of this fact.

Proof Our first step is to note that the covering and packing numbers of the set $\mathcal{F}$ in the ℓ_∞ metric satisfy

$$\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \log M(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \frac{1}{\delta}. \tag{8.7.4}$$

To see this, fix some $\delta \in (0, 1)$ and assume for simplicity that $1/\delta$ is an integer. Define the sets $E_j = [\delta(j-1), \delta j)$, and for each $v \in \{-1, 1\}^{1/\delta}$ define $h_v(x) = \sum_{j=1}^{1/\delta} v_j \mathbf{1}\{x \in E_j\}$. Then define the function $f_v(t) = \int_0^t h_v(t)dt$, which increases or decreases linearly on each interval of width δ in $[0, 1]$. Then these f_v form a 2δ-packing and a 2δ-cover of $\mathcal{F}$, and there are $2^{1/\delta}$ such f_v. Thus the asymptotic approximation (8.7.4) holds.

JCD Comment: TODO: Draw a picture

Now, if for some fixed $x \in [0, 1]$ and $f, g \in \mathcal{F}$ we define P_f and P_g to be the distributions of the observations $f(x) + \varepsilon$ or $g(x) + \varepsilon$, we have that

$$D_{\rm kl}\left(P_f \| P_g\right) = \frac{1}{2\sigma^2}(f(X_i) - g(X_i))^2 \leq \frac{\|f - g\|_\infty^2}{2\sigma^2},$$

and if P_f^n is the distribution of the n observations $f(X_i) + \varepsilon_i$, $i = 1, \ldots, n$, we also have

$$D_{\rm kl}\left(P_f^n \| P_g^n\right) = \sum_{i=1}^n \frac{1}{2\sigma^2}(f(X_i) - g(X_i))^2 \leq \frac{n}{2\sigma^2} \|f - g\|_\infty^2.$$

In particular, this implies the upper bound

$$\log N_{\rm kl}\left(\epsilon, \mathcal{P}\right) \lesssim \frac{1}{\sigma\epsilon}$$

on the KL-metric entropy of the class $\mathcal{P} = \{P_f : f \in \mathcal{F}\}$, as $\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \delta^{-1}$. Thus we have completed steps (i) and (ii) in our program above.

It remains to choose the critical radius in step (iii), but this is now relatively straightforward: by choosing $\epsilon_n \asymp (1/\sigma n)^{1/3}$, and whence $n\epsilon_n^2 \asymp (n/\sigma^2)^{1/3}$, we find that taking $\delta \asymp (\sigma^2/n)^{1/3}$ is sufficient to ensure that $\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \gtrsim \delta^{-1} \geq 4n\epsilon_n^2 + 2\log 2$. Thus we have

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_\infty) \gtrsim \delta_n \cdot \frac{1}{2} \gtrsim \left(\frac{\sigma^2}{n} \right)^{1/3}$$

as desired. □

8.8 Deferred proofs

8.8.1 Proof of Proposition 8.4.6

Our argument for proving the proposition parallels that of the classical Fano inequality by Cover and Thomas [53]. Letting E be a $\{0, 1\}$-valued indicator variable for the event $\rho(\widehat{V}, V) \le t$, we compute the entropy $H(E, V \mid \widehat{V})$ in two different ways. On one hand, by the chain rule for entropy, we have

$$H(E, V \mid \widehat{V}) = H(V \mid \widehat{V}) + \underbrace{H(E \mid V, \widehat{V})}_{=0}, \tag{8.8.1}$$

where the final term vanishes since E is $(V, \widehat{V})$-measurable. On the other hand, we also have

$$H(E, V \mid \widehat{V}) = H(E \mid \widehat{V}) + H(V \mid E, \widehat{V}) \le H(E) + H(V \mid E, \widehat{V}),$$

using the fact that conditioning reduces entropy. Applying the definition of conditional entropy yields

$$H(V \mid E, \widehat{V}) = \mathbb{P}(E = 0)H(V \mid E = 0, \widehat{V}) + \mathbb{P}(E = 1)H(V \mid E = 1, \widehat{V}),$$

and we upper bound each of these terms separately. For the first term, we have

$$H(V \mid E = 0, \widehat{V}) \le \log(|\mathcal{V}| - N_t^{\min}),$$

since conditioned on the event $E = 0$, the random variable V may take values in a set of size at most $|\mathcal{V}| - N_t^{\min}$. For the second, we have

$$H(V \mid E = 1, \widehat{V}) \le \log N_t^{\max},$$

since conditioned on $E = 1$, or equivalently on the event that $\rho(\widehat{V}, V) \le t$, we are guaranteed that V belongs to a set of cardinality at most $N_t^{\max}$.

Combining the pieces and and noting $\mathbb{P}(E = 0) = P_t$, we have proved that

$$H(E, V \mid \widehat{V}) \le H(E) + P_t \log\left(|\mathcal{V}| - N^{\min}\right) + (1 - P_t) \log N_t^{\max}.$$

Combining this inequality with our earlier equality (8.8.1), we see that

$$H(V \mid \widehat{V}) \le H(E) + P_t \log(|\mathcal{V}| - N_t^{\min}) + (1 - P_t) \log N_t^{\max}.$$

Since $H(E) = h_2(P_t)$, the claim (8.4.9) follows.

8.8.2 Proof of Corollary 8.4.7

First, by the information-processing inequality [e.g. 53, Chapter 2], we have $I(V; \widehat{V}) \le I(V; X)$, and hence $H(V \mid X) \le H(V \mid \widehat{V})$. Since $h_2(P_t) \le \log 2$, inequality (8.4.9) implies that

$$H(V \mid X) - \log N_t^{\max} \le H(V \mid \widehat{V}) - \log N_t^{\max} \le \mathbb{P}(\rho(\widehat{V}, V) > t) \log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}} + \log 2.$$

Rearranging the preceding equations yields

$$\mathbb{P}(\rho(\widehat{V}, V) > t) \ge \frac{H(V \mid X) - \log N_t^{\max} - \log 2}{\log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}}}. \tag{8.8.2}$$

Note that his bound holds without any assumptions on the distribution of V.

By definition, we have $I(V;X) = H(V) - H(V \mid X)$. When V is uniform on $\mathcal{V}$, we have $H(V) = \log|\mathcal{V}|$, and hence $H(V \mid X) = \log|\mathcal{V}| - I(V;X)$. Substituting this relation into the bound (8.8.2) yields the inequality

$$\mathbb{P}(\rho(\widehat{V}, V) > t) \geq \frac{\log \frac{|\mathcal{V}|}{N_t^{\max}}}{\log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}}} - \frac{I(V;X) + \log 2}{\log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}}} \geq 1 - \frac{I(V;X) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}.$$

8.8.3 Proof of Lemma 8.5.2

Fix an (arbitrary) estimator $\widehat{\theta}$. By assumption (8.5.1), we have

$$\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^{d} \mathbf{1}\left\{[\widehat{\mathrm{v}}(\theta)]_j \neq v_j\right\}.$$

Taking expectations, we see that

$$\begin{aligned}\sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\Phi(\rho(\widehat{\theta}(X), \theta(P)))\right] &\geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{P_v}\left[\Phi(\rho(\widehat{\theta}(X), \theta_v))\right] \\ &\geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} 2\delta \sum_{j=1}^{d} \mathbb{E}_{P_v}\left[\mathbf{1}\left\{[\psi(\widehat{\theta})]_j \neq v_j\right\}\right]\end{aligned}$$

as the average is smaller than the maximum of a set and using the separation assumption (8.5.1). Recalling the definition of the mixtures $\mathbb{P}_{\pm j}$ as the joint distribution of V and X conditional on $V_j = \pm 1$, we swap the summation orders to see that

$$\begin{aligned}\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v\left([\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right) &= \frac{1}{|\mathcal{V}|} \sum_{v : v_j = 1} P_v\left([\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right) + \frac{1}{|\mathcal{V}|} \sum_{v : v_j = -1} P_v\left([\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right) \\ &= \frac{1}{2}\mathbb{P}_{+j}\left([\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right) + \frac{1}{2}\mathbb{P}_{-j}\left([\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right).\end{aligned}$$

This gives the statement claimed in the lemma, while taking an infimum over all testing procedures $\Psi : \mathcal{X} \to \{-1, +1\}$ gives the claim (8.5.2).

8.9 Bibliography

For a fuller technical introduction into nonparametric estimation, see the book by Tsybakov [167]. Has'minskii [100].

The material in Section 8.7 is based on a paper of Yang and Barron [175].

8.10 Exercises

Exercise 8.1 (A generalized version of Fano's inequality; cf. Proposition 8.4.6)**:** Let $\mathcal{V}$ and $\widehat{\mathcal{V}}$ be arbitrary sets, and suppose that π is a (prior) probability measure on $\mathcal{V}$, where V is distributed according to π. Let $V \to X \to \widehat{V}$ be Markov chain, where V takes values in $\mathcal{V}$ and $\widehat{V}$ takes values

in $\widehat{\mathcal{V}}$. Let $\mathcal{N} \subset \mathcal{V} \times \widehat{\mathcal{V}}$ denote a measurable subset of $\mathcal{V} \times \widehat{\mathcal{V}}$ (a collection of neighborhoods), and for any $\widehat{v} \in \widehat{\mathcal{V}}$, denote the slice

$$\mathcal{N}_{\widehat{v}} := \{v \in \mathcal{V} : (v, \widehat{v}) \in \mathcal{N}\}. \tag{8.10.1}$$

That is, $\mathcal{N}$ denotes the neighborhoods of points v for which we do not consider a prediction $\widehat{v}$ for v to be an error, and the slices (8.10.1) index the neighborhoods. Define the "volume" constants

$$p^{\max} := \sup_{\widehat{v}} \pi(V \in \mathcal{N}_{\widehat{v}}) \quad \text{and} \quad p^{\min} := \inf_{\widehat{v}} \pi(V \in \mathcal{N}_{\widehat{v}}).$$

Define the error probability $P_{\mathsf{error}} = \mathbb{P}[(V, \widehat{V}) \notin \mathcal{N}]$ and entropy $h_2(p) = -p \log p - (1-p) \log(1-p)$.

(a) Prove that for any Markov chain $V \to X \to \widehat{V}$, we have

$$h_2(P_{\mathsf{error}}) + P_{\mathsf{error}} \log \frac{1 - p^{\min}}{p^{\max}} \geq \log \frac{1}{p^{\max}} - I(V; \widehat{V}). \tag{8.10.2}$$

(b) Conclude from inequality (8.10.2) that

$$\mathbb{P}[(V, \widehat{V}) \notin \mathcal{N}] \geq 1 - \frac{I(V; X) + \log 2}{\inf_{\widehat{v}} \log \frac{1}{\pi(\mathcal{N}_{\widehat{v}})}}.$$

(c) Now we give a version explicitly using distances. Let $\mathcal{V} \subset \mathbb{R}^d$ and define $\mathcal{N} = \{(v, v') : \|v - v'\| \leq \delta\}$ to be the points within δ of one another. Let $\mathbb{B}_v$ denote the $\|\cdot\|$-ball of radius 1 centered at v. Conclude that for any prior π on $\mathbb{R}^d$ that

$$\mathbb{P}\left(\|V - \widehat{V}\|_2 \geq \delta\right) \geq 1 - \frac{I(V; X) + \log 2}{\log \frac{1}{\sup_v \pi(\delta \mathbb{B}_v)}}.$$

Exercise 8.2: In this question, we will show that the minimax rate of estimation for the parameter of a uniform distribution (in squared error) scales as $1/n^2$. In particular, assume that $X_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}(\theta, \theta + 1)$, meaning that X_i have densities $p(x) = \mathbf{1}\{x \in [\theta, \theta + 1]\}$. Let $X_{(1)} = \min_i\{X_i\}$ denote the first order statistic.

(a) Prove that

$$\mathbb{E}[(X_{(1)} - \theta)^2] = \frac{2}{(n+1)(n+2)}.$$

(Hint: the fact that $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \geq t) dt$ for any positive Z may be useful.)

(b) Using Le Cam's two-point method, show that the minimax rate for estimation of $\theta \in \mathbb{R}$ for the uniform family $\mathcal{U} = \{\mathsf{Uniform}(\theta, \theta + 1) : \theta \in \mathbb{R}\}$ in squared error has lower bound c/n^2, where c is a numerical constant.

Exercise 8.3 (Sign identification in sparse linear regression)**:** In sparse linear regression, we have n observations $Y_i = \langle X_i, \theta^* \rangle + \varepsilon_i$, where $X_i \in \mathbb{R}^d$ are known (fixed) matrices and the vector θ^* has a small number $k \ll d$ of non-zero indices, and $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$. In this problem, we investigate the problem of *sign recovery*, that is, identifying the vector of signs $\mathrm{sign}(\theta_j^*)$ for $j = 1, \ldots, d$, where $\mathrm{sign}(0) = 0$.

Assume we have the following process: fix a signal threshold $\theta_{\min} > 0$. First, a vector $S \in \{-1, 0, 1\}^d$ is chosen uniformly at random from the set of vectors $\mathcal{S}_k := \{s \in \{-1, 0, 1\}^d : \|s\|_1 = k\}$. Then we define vectors θ^s so that $\theta^s_j = \theta_{\min} s_j$, and conditional on $S = s$, we observe

$$Y = X\theta^s + \varepsilon, \quad \varepsilon \sim \mathsf{N}(0, \sigma^2 I_{n\times n}).$$

(Here $X \in \mathbb{R}^{n\times d}$ is a known fixed matrix.)

(a) Use Fano's inequality to show that for any estimator $\widehat{S}$ of S, we have

$$\mathbb{P}(\widehat{S} \neq S) \geq \frac{1}{2} \quad \text{unless} \quad n \geq c \frac{\frac{d}{k} \log \binom{d}{k}}{\left\|n^{-1/2} X\right\|_{\mathrm{Fr}}^2} \frac{\sigma^2}{\theta_{\min}^2},$$

where c is a numerical constant. You may assume that $k \geq 4$ or $\log \binom{d}{k} \geq 4 \log 2$.

(b) Assume that $X \in \{-1, 1\}^{n\times d}$. Give a lower bound on how large n must be for sign recovery. Give a one sentence interpretation of $\sigma^2/\theta_{\min}^2$.

Exercise 8.4 (General minimax lower bounds)**:** In this exercise, we outline a more general approach to minimax risk than that afforded by studying losses applied to parameter error. In particular, we may instead consider losses of the form

$$L : \Theta \times \mathcal{P} \to \mathbb{R}_+$$

where $\mathcal{P}$ is a collection of distributions and Θ is a parameter space, where additionally the losses satisfy the condition

$$\inf_{\theta\in\Theta} L(\theta, P) = 0 \text{ for all } P \in \mathcal{P}.$$

(a) Consider a statistical risk minimization problem, where we have a distribution P on random variable $X \in \mathcal{X}$, loss function $f : \Theta \times \mathcal{X} \to \mathbb{R}$, and for $P \in \mathcal{P}$ define the population risk $F_P(\theta) := \mathbb{E}_P[f(\theta, X)]$. Show that

$$L(\theta, P) := F_P(\theta) - \inf_{\theta\in\Theta} F_P(\theta)$$

satisfies the conditions above.

(b) For distributions P_0, P_1, define the *separation* between them (for the loss L) by

$$\mathsf{sep}_L(P_0, P_1; \Theta) := \sup \left\{ \delta \geq 0 : \begin{array}{l} L(\theta, P_0) \leq \delta \text{ implies } L(\theta, P_1) \geq \delta \\ L(\theta, P_1) \leq \delta \text{ implies } L(\theta, P_0) \geq \delta \end{array} \text{ for any } \theta \in \Theta \right\}. \quad (8.10.3)$$

That is, having small loss on P_0 implies large loss on P_1 and vice versa.

We say a collection of distributions $\{P_v\}_{v\in\mathcal{V}}$ indexed by $\mathcal{V}$ is δ-*separated* if $\mathsf{sep}_L(P_v, P_{v'}; \Theta) \geq \delta$. Show that if $\{P_v\}_{v\in\mathcal{V}}$ is δ-separated, then for any estimator $\widehat{\theta}$

$$\frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} \mathbb{E}_{P_v}[L(\widehat{\theta}, P_v)] \geq \delta \inf_{\widehat{v}} \mathbb{P}(\widehat{v} \neq V),$$

where $\mathbb{P}$ is the joint distribution over the random index V chosen uniformly and then X sampled $X \sim P_v$ conditional on $V = v$.

(c) Show that if $\mathcal{P}$ has a δ-separated subset $\{P_v\}_{v\in\mathcal{V}}$, then

$$\mathfrak{M}(\mathcal{P}, L) := \inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_P[L(\widehat{\theta}, P)] \geq \delta \inf_{\widehat{v}} \mathbb{P}(\widehat{v} \neq V).$$

Exercise 8.5 (Optimality in stochastic optimization)**:** In this question, we prove minimax lower bounds on the convergence rates in stochastic optimization problems based on the size of the domain over which we optimize and certain Lipschitz conditions of the functions themselves. You may assume the dimension d in the problems we consider is as large as you wish.

The setting is as follows: we have a domain $\Theta \subset \mathbb{R}^d$, function $f : \Theta \times \mathcal{X} \to \mathbb{R}$, which is convex in its first argument, and population risks $F_P(\theta) := \mathbb{E}_P[f(\theta, X)]$, where the expectation is taken over $X \sim P$. For any two functions F_0, F_1, let $\theta^v \in \operatorname{argmin}_{\theta\in\Theta} F_v(\theta)$, and define the *optimization distance* between F_0 and F_1 by

$$d_{\rm opt}(F_0, F_1; \Theta) := \inf_{\theta\in\Theta} \left\{F_0(\theta) + F_1(\theta) - F_0(\theta^0) - F_1(\theta^1)\right\}.$$

Define also the loss $L(\theta, P) := F_P(\theta) - \inf_{\theta\in\Theta} F_P(\theta)$.

(a) Show for any $\delta \geq 0$ that if $d_{\rm opt}(F_0, F_1; \Theta) \geq \delta$, then $\mathsf{sep}_L(P_0, P_1; \Theta) \geq \frac{\delta}{2}$, where sep is defined in Eq. (8.10.3).

We consider lower bounds for stochastic optimization problems with appropriately Lipschitz f.

(b) Let the sample space $\mathcal{X} = \{\pm e_j\}_{j=1}^d$ be the signed standard basis vectors, and for $\theta \in \mathbb{R}^d$, define

$$f(\theta; x) := \begin{cases} |\theta_j - 1| & \text{if } x = e_j \\ |\theta_j + 1| & \text{if } x = -e_j. \end{cases}$$

Let $v \in \{-1, 1\}^d$. For some $\delta > 0$ to be chosen, define the distribution P_v on X by

$$X = \begin{cases} v_j e_j & \text{w.p. } \frac{1+\delta}{2d} \\ -v_j e_j & \text{w.p. } \frac{1-\delta}{2d}. \end{cases}$$

(Note that $\|X\|_0 = 1$.) Give an explicit formula for

$$F_v(\theta) := \mathbb{E}_{P_v}[f(\theta, X)].$$

(c) Show that $\theta^v = \operatorname{argmin}_\theta F_v(\theta) = v$ and that $F_v(\theta^v) = 1 - \delta$.

(d) Let $\mathcal{V} \subset \{\pm 1\}^d$ be a $d/2$-packing in ℓ_1-distance of cardinality at least $\exp(d/8)$ (by Gilbert-Varshamov, Lemma 8.2.3). Assume that $\Theta \supset [-1, 1]^d$. Show that $d_{\rm opt}(F_v, F_{v'}) \geq \delta \|v - v'\|_1 / d$ for all distinct $v, v' \in \mathcal{V}$.

(e) For our loss $L(\theta, P) = F_P(\theta) - \inf_{\theta\in\Theta} F_P(\theta)$, show that the minimax loss gap

$$\mathfrak{M}_n(\mathcal{P}, \Theta, L) := \inf_{\widehat{\theta}_n} \sup_{P\in\mathcal{P}} \mathbb{E}_P[L(\widehat{\theta}_n(X_1^n), P)] = \inf_{\widehat{\theta}_n} \sup_{P\in\mathcal{P}} \left\{\mathbb{E}_P[F_P(\widehat{\theta}_n(X_1^n)) - F_P^\star]\right\}$$

(where $F_P^\star = \inf_{\theta\in\Theta} F_P(\theta)$ and $X_1^n \stackrel{\rm iid}{\sim} P$) satisfies

$$\mathfrak{M}_n(\mathcal{P}, L) \geq c \min\left\{\frac{\sqrt{d}}{\sqrt{n}}, 1\right\}.$$

where $c > 0$ is a constant. You may assume $d \geq 8$ (or any other large constant) for simplicity.

(f) Show how to modify this construction so that for constants $L, R > 0$, if $\Theta \supset [-R, R]^d$, there are functions f that are L-Lipschitz with respect to the ℓ_∞ norm, meaning

$$|f(\theta; x) - f(\theta'; x)| \le L \left\|\theta - \theta'\right\|_\infty,$$

such that for this domain Θ, loss f (and induced L), and the same family of distributions $\mathcal{P}$ as above,

$$\mathfrak{M}_n(\mathcal{P}, \Theta, L) \ge cLR \min\left\{\frac{\sqrt{d}}{\sqrt{n}}, 1\right\}.$$

(g) Suppose that instead, we have $\Theta \supset \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \le R_2\}$, the ℓ_2-ball of radius R_2, and allow f to be L_2-Lipschitz with respect to the ℓ_2-norm (instead of ℓ_∞). Show that

$$\mathfrak{M}_n(\mathcal{P}, \Theta, L) \ge c\frac{L_2 R_2}{\sqrt{n}}.$$

(h) What do these results say about stochastic gradient methods?

Exercise 8.6 (Optimality in high-dimensional stochastic optimization)**:** We revisit the setting in Question 8.5, except that we consider a high-dimensional regime. In particular, we will prove lower bounds on optimization when the domain $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \le r\}$, the ℓ_1-ball, and the loss functions f are M-Lipschitz with respect to the ℓ_1-norm, equivalently, that $\|\nabla_\theta f(\theta, x)\|_\infty \le M$ for all $\theta \in \Theta$. For distributions P on X, define $F_P(\theta) = \mathbb{E}_P[f(\theta, X)]$ and $F_P^\star = \inf_{\theta\in\Theta} F_P(\theta)$.

We now give an explicit construction. Let the sample space $\mathcal{X} = \{-1, 1\}^d$ be the hypercube, and consider linear losses

$$f(\theta; x) = M\langle\theta, x\rangle,$$

which are evidently M-Lipschitz w.r.t. the ℓ_1-norm. Now, for the packing set $\mathcal{V} = \{\pm e_j\}_{j=1}^d$ of the standard basis vectors, define the distribution P_v on $X \in \{\pm 1\}^d$ to have independent coordinates with

$$X_j = \begin{cases} 1 & \text{w.p. } \frac{1+\delta v_j}{2} \\ -1 & \text{w.p. } \frac{1-\delta v_j}{2}. \end{cases}$$

That is, $X \sim P_v$ has independent random sign coordinates except in coordinate j when $v = e_j$, where $P_{\pm e_j}(X_j = \pm 1) = \frac{1\pm\delta}{2}$. Let

$$F_v(\theta) = \mathbb{E}_{P_v}[f(\theta, X)] = M\delta\langle v, \theta\rangle.$$

(a) Give $\theta^v := \operatorname{argmin}_{\theta\in\Theta} F_v(\theta)$.

(b) Using the optimization distance $d_{\rm opt}(F_0, F_1; \Theta) = \inf_{\theta\in\Theta}\{F_0(\theta) + F_1(\theta) - F_0^\star - F_1^\star\}$, where $F_v^\star = \inf_{\theta\in\Theta} F_v(\theta)$, defined in Question 8.5, show the separation

$$\min_{v\neq v'} d_{\rm opt}(F_v, F_{v'}; \Theta) = M\delta r.$$

(c) Let the loss $L(\theta, P) = F_P(\theta) - \inf_{\theta\in\Theta} F_P(\theta)$ as in Question 8.5, let $\mathcal{P}$ be the collection of distributions supported on $[-1, 1]^d$, and define the minimax loss gap

$$\mathfrak{M}_n(\mathcal{P}, \Theta, L) := \inf_{\widehat{\theta}_n} \sup_{P\in\mathcal{P}} \left\{\mathbb{E}_P\left[F_P(\widehat{\theta}_n(X_1^n)) - F_P^\star\right]\right\}$$

where $X_1^n \stackrel{\text{iid}}{\sim} P$. Show that there exists a numerical constant $c > 0$ such that

$$\mathfrak{M}_n(\mathcal{P}, \Theta, L) \geq c \frac{\sqrt{\log(2d)}}{\sqrt{n}}.$$

(You may assume $d \geq 2$ to avoid trivial cases.) *Hint.* Use the result of Question 8.4 part (c).

Exercise 8.7: In this question, we study the question of whether adaptivity can give better estimation performance for linear regression problems. That is, for $i = 1, \ldots, n$, assume that we observe variables Y_i in the usual linear regression setup,

$$Y_i = \langle X_i, \theta \rangle + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2), \tag{8.10.4}$$

where $\theta \in \mathbb{R}^d$ is unknown. But now, based on observing $Y_1^{i-1} = \{Y_1, \ldots, Y_{i-1}\}$, we allow an adaptive choice of the next predictor variables $X_i \in \mathbb{R}^d$. Let $\mathcal{L}_{\text{ada}}^n(\mathsf{F}^2)$ denote the family of linear regression problems under this adaptive setting (with n observations) where we constrain the Frobenius norm of the data matrix $X^\top = [X_1 \; \cdots \; X_n]$, $X \in \mathbb{R}^{n \times d}$, to have bound $\|X\|_{\text{Fr}}^2 = \sum_{i=1}^n \|X_i\|_2^2 \leq \mathsf{F}^2$. We use Assouad's method to show that the minimax mean-squared error satisfies the following bound:

$$\mathfrak{M}(\mathcal{L}_{\text{ada}}^n(\mathsf{F}^2), \|\cdot\|_2^2) := \inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \geq \frac{d\sigma^2}{n} \cdot \frac{1}{16 \frac{1}{dn} \mathsf{F}^2}. \tag{8.10.5}$$

Here the infimum is taken over all adaptive procedures satisfying $\|X\|_{\text{Fr}}^2 \leq \mathsf{F}^2$.

In general, when we choose X_i based on the observations Y_1^{i-1}, we are taking $X_i = F_i(Y_1^{i-1}, U_1^i)$, where U_i is a random variable independent of ε_i and Y_1^{i-1} and F_i is some function. Justify the following steps in the proof of inequality (8.10.5):

(i) Assume that nature chooses $v \in \mathcal{V} = \{-1, 1\}^d$ uniformly at random and, conditionally on v, let $\theta = \theta_v$. Justify

$$\mathfrak{M}(\mathcal{L}_{\text{ada}}^n(\mathsf{F}^2), \|\cdot\|_2^2) \geq \inf_{\widehat{\theta}} \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{\theta_v}[\|\widehat{\theta} - \theta_v\|_2^2].$$

Argue it is no loss of generality to assume that the choices for X_i are deterministic based on the Y_1^{i-1}. Thus, throughout we assume that $X_i = F_i(Y_1^{i-1}, u_1^i)$, where u_1^n is a fixed sequence, or, for simplicity, that X_i is a function of Y_1^{i-1}.

(ii) Fix $\delta > 0$. Let $v \in \{-1, 1\}^d$, and for each such v, define $\theta_v = \delta v$. Also let P_v^n denote the joint distribution (over all adaptively chosen X_i) of the observed variables $Y_1, \ldots, Y_n$, and define $P_{+j}^n = \frac{1}{2^{d-1}} \sum_{v : v_j = 1} P_v^n$ and $P_{-j}^n = \frac{1}{2^{d-1}} \sum_{v : v_j = -1} P_v^n$, so that $P_{\pm j}^n$ denotes the distribution of the Y_i when $v \in \{-1, 1\}^d$ is chosen uniformly at random but conditioned on $v_j = \pm 1$. Then

$$\inf_{\widehat{\theta}} \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{\theta_v}[\|\widehat{\theta} - \theta_v\|_2^2] \geq \frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\text{TV}}\right].$$

(iii) We have

$$\frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\text{TV}}\right] \geq \frac{\delta^2 d}{2} \left[1 - \left(\frac{1}{d} \sum_{j=1}^d \left\|P_{+j}^n - P_{-j}^n\right\|_{\text{TV}}^2\right)^{\frac{1}{2}}\right].$$

(iv) Let $P_{+j}^{(i)}$ be the distribution of the random variable Y_i conditioned on $v_j = +1$ (with the other coordinates of v chosen uniformly at random), and let $P_{+j}^{(i)}(\cdot \mid y_1^{i-1}, x_i)$ denote the distribution of Y_i conditioned on $v_j = +1$, $Y_1^{i-1} = y_1^{i-1}$, and x_i. Justify

$$\left\|P_{+j}^n - P_{-j}^n\right\|_{\rm TV}^2 \le \frac{1}{2} D_{\rm kl}\left(P_{+j}^n \| P_{-j}^n\right) \le \frac{1}{2}\sum_{i=1}^n \int D_{\rm kl}\left(P_{+j}^{(i)}(\cdot \mid y_1^{i-1}, x_i) \| P_{-j}^{(i)}(\cdot \mid y_1^{i-1}, x_i)\right) dP_{+j}^{i-1}(y_1^{i-1}, x_i).$$

(v) Then we have

$$\sum_{j=1}^d D_{\rm kl}\left(P_{+j}^{(i)}(\cdot \mid y_1^{i-1}, x_i) \| P_{-j}^{(i)}(\cdot \mid y_1^{i-1}, x_i)\right) \le \frac{2\delta^2}{\sigma^2}\left\|x_i\right\|_2^2.$$

(vi) We have

$$\sum_{j=1}^d \left\|P_{+j}^n - P_{-j}^n\right\|_{\rm TV}^2 \le \frac{\delta^2}{\sigma^2}\mathbb{E}[\|X\|_{\rm Fr}^2],$$

where the final expectation is over V drawn uniformly in $\{-1, 1\}^d$ and all Y_i, X_i.

(vii) Show how to choose δ appropriately to conclude the minimax bound (8.10.5).

Exercise 8.8: Suppose under the setting of Question 8.7 that we may no longer be adaptive, meaning that the matrix $X \in \mathbb{R}^{n\times d}$ must be chosen ahead of time (without seeing any data). Assuming $n \ge d$, is it possible to attain (within a constant factor) the risk (8.10.5)? If so, give an example construction, if not, explain why not.

Exercise 8.9 (The curse of dimensionality in nonparametric regression)**:** Consider the nonparametric regression problem in Section 8.6. Let $\mathbb{B}^d$ be the unit ℓ_2-ball in $\mathbb{R}^d$ and consider the function class $\mathcal{F}$ of 1-Lipschitz functions taking values in $[-1, 1]$ on $\mathbb{B}^d$, and consider the error $\|f - g\|_2^2 = \int_{\mathbb{B}^d}(f(x) - g(x))^2 dx$. (Here, 1-Lipschitz means $|f(x) - f(x')| \le \|x - x'\|_2$ for any x, x'.) We show the minimax lower bound (8.6.4) for this function class using Fano's method. Fix $\delta \in [0, 1]$ to be chosen and let $\{x_j\}_{j=1}^M$ be the centers of a maximal 2δ-packing of $\mathbb{B}^d$, so that $M \ge (\frac{1}{2\delta})^d$ (by Lemma 4.3.10), and define the "bump" functions

$$g_j(x) = \delta\left[1 - \left\|x - x_j\right\|_2 / \delta\right]_+,$$

which all have disjoint support. Then for a vector $v \in \{\pm 1\}^M$, define

$$f_v(x) := \sum_{j=1}^M v_j g_j(x).$$

(a) Show that $f_v \in \mathcal{F}$.

(b) Show that $\int g_j(x)^2 dx = \frac{2\cdot \mathrm{SA}(d)}{d(d+1)(d+2)}\delta^{2+d}$, where $\mathrm{SA}(d)$ denotes the surface area of $\mathbb{B}^d$.

(c) Use the Gilbert-Varshamov bound (Lemma 8.2.3) to show there is a collection $\mathcal{V} \subset \{\pm 1\}^M$ of cardinality $\exp(M/8)$ with $\|f_v - f_{v'}\|_2^2 \ge c_d\delta^2$ for all $v \ne v' \in \mathcal{V}$, where c_d depends only on the dimension d.

(d) Prove the minimax lower bound (8.6.4) for $\beta = 1$.

Exercise 8.10 (Optimal algorithms for memory access)**:** In a modern CPU, memory is organized in a hierarchy, so that data upon which computations are being actively performed lies in a very small memory close to the logic units of the processor for which access is extraordinarily fast, while data not being actively used lies in slower memory slightly farther from the processor. (Modern processor memory is generally organized into the registers—a small number of 4- or 8-byte memory locations on the processor—and level 1, 2, (and sometimes 3 or more) cache, which contain small amounts of data and increasing access times, and RAM (random access memory).) Moving data—communicating—between levels of the memory hierarchy is both power intensive and very slow relative to computation on the data itself, so that in many algorithms the bulk of the time of the algorithm is in moving data from one place to another to be computed upon. Thus, developing very fast algorithms for numerical (and other) tasks on modern computers requires careful tracking of memory access and communication, and careful control of these quantities can often yield orders of magnitude speed improvements in execution. In this problem, you will prove a lower bound on the number of communication steps that a variety of numerical-type methods must perform, giving a concrete (attainable) inequality that allows one to certify optimality of *specific* algorithms.

In particular, we consider matrix multiplication, as it is a proxy for a class of cubic algorithms that are well behaved. Let $A, B \in \mathbb{R}^{n \times n}$ be matrices, and assume we wish to compute $C = AB$, via the simple algorithm that for all i, j sets

$$C_{ij} = \sum_{l=1}^{n} A_{il} B_{lj}.$$

Computationally, this forces us to repeatedly execute operations of the form

$$\mathsf{Mem}(C_{ij}) = F(\mathsf{Mem}(A_{il}), \mathsf{Mem}(B_{lj}), \mathsf{Mem}(C_{ij})),$$

where F is some function—that may depend on i, j, l—and $\mathsf{Mem}(\cdot)$ indicates that we access the memory associated with the argument. (In our case, we have $C_{ij} = C_{ij} + A_{il} \cdot B_{lj}$.) We assume that executing F requires that $\mathsf{Mem}(A_{il})$, $\mathsf{Mem}(B_{lj})$, and $\mathsf{Mem}(C_{ij})$ belong to fast memory, and that each are distinct (stored in a separate place in flow and fast memory). We assume that the order of the computations does *not* matter, so we may re-order them in any way. We call $\mathsf{Mem}(A_{il})$ (respectively B or C) and *operand* in our computation. We let M denote the size of fast/local memory, and we would like to lower bound the number of times we must communicate an operand into or out of the fast local memory as a function of n, the matrix size, and M, the fast memory size, when all we may do is re-order the computation being executed. We let N_{Store} denote the number of times we write something from fast memory out to slow memory and let N_{Load} the number of times we load something from slow memory to fast memory. Let N be the total number of operations we execute (for simple matrix multiplication, we have $N = n^3$, though with sparse matrices, this can be smaller).

We analyze the procedure by breaking the computation into a number of segments, where each segment contains precisely M load or store (communication-causing) instructions.

(a) Let N_{seg} be an upper bound on the number of evaluations with the function $F(\cdot)$ in any given segment (you will upper bound this in a later part of the problem). Justify that

$$N_{\mathsf{Store}} + N_{\mathsf{Load}} \geq M \left\lfloor N / N_{\mathsf{seg}} \right\rfloor.$$

(b) Within a segment, all operands involved must be in fast memory at least once to be computed with. Assume that memory locations $\mathsf{Mem}(A_{il})$, $\mathsf{Mem}(B_{lj})$, and $\mathsf{Mem}(C_{ij})$ do not overlap. For any operand involved in a memory operation in one of the segments, the operand (1) was already in fast memory at the beginning of the segment, (2) was read from slow memory, (3) is still in fast memory at the end of the segment, or (4) is written to slow memory at the end of the segment. (There are also operands potentially created during execution that are simply discarded; we do not bound those.) Justify the following: within a segment, for each type of operand $\mathsf{Mem}(A_{ij})$, $\mathsf{Mem}(B_{ij})$, or $\mathsf{Mem}(C_{ij})$, there are at most $c \cdot M$ such operands (i.e. there are at most cM operands of type $\mathsf{Mem}(A_{ij})$, independent of the others, and so on), where c is a numerical constant. What value of c can you attain?

(c) Using the result of question 6.1, argue that $N_{\mathsf{seg}} \le c' \sqrt{M^3}$ for a numerical constant c'. What value of c' do you get?

(d) Using the result of part (c), argue that the number of loads and stores satisfies

$$N_{\mathsf{Store}} + N_{\mathsf{Load}} \ge c'' \frac{N}{\sqrt{M}} - M$$

for a numerical constant c''. What is your constant?

JCD Comment: A few additional question ideas:

1. Use the global Fano method technique to give lower bounds for density estimation
2. Curse of dimensionality in high-dimensional regression? The idea would be to take disjoint δ-balls $B_j \subset \mathbb{B}^d$, where $\mathbb{B}^d = \{x \mid \|x\| \le 1\}$ is the unit ball, with centers x_j, where j runs from 1 to $(1/\delta)^d$, then define the bump function $g_j(x) = \delta \left[1 - \|x - x_j\| / \delta\right]_+$. Then set $f_v(x) = \sum_j v_j g_j(x)$, which is 1-Lipschitz for the norm $\|\cdot\|$. Then the separation is δ, while the log cardinality is $2^{\delta^{-d}}$, giving $\delta^2(1 - n\delta^{2+d})$ as the lower bound. Take $\delta = n^{-1/(2+d)}$.

Chapter 9

Constrained risk inequalities

In this chapter, we revisit our minimax bounds in the context of what we term *constrained* risk inequalities. While the minimax risk of provides a first approach for providing fundamental limits on procedures, its reliance on the collection of *all* measurable functions as its class of potential estimators is somewhat limiting. Indeed, in most statistical and statistical learning problems, we have some type of constraint on our procedures: they must be efficiently computable, they must work with data arriving in a sequential stream, they must be robust, or they must protect the privacy of the providers of the data. In modern computational hardware, where physical limits prevent increasing clock speeds, we may like to use as much parallel computation as possible, though there are potential tradeoffs between "sequentialness" of procedures and their parallelism.

With this as context, we replace the minimax risk of Chapter 8.1 with the *constrained minimax risk*, which, given a collection $\mathcal{C}$ of possible procedures—private, communication limited, or otherwise—defines

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho, \mathcal{C}) := \inf_{\widehat{\theta} \in \mathcal{C}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X), \theta(P))\big) \right], \tag{9.0.1}$$

where as in the original defining equation (8.1.1) of the minimax risk, $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a nondecreasing loss, ρ is a semimetric on the space Θ, and the expectation is taken over the sample $X \sim P$. In this chapter, we study the quantity (9.0.1) via a few examples, highlighting possibilities and challenges with its analysis. We will focus on a restricted class of examples—many procedures do not fall in the framework we consider—that assumes, given a sample $X_1, \ldots, X_n$, we can represent the class $\mathcal{C}$ of estimators under consideration as acting on some view or processed version Z_i of X_i. This allows us to study communication complexity, memory complexity, and certain private estimators.

9.1 Strong data processing inequalities

The starting point for our results is to consider *strong data processing inequalities*, which improve upon the standard data processing inequality for divergences, as in Chapter 2.1.3, to provide more quantitative versions. The initial setting is straightforward: we have distributions P_0 and P_1 on a space $\mathcal{X}$, and a channel (Markov kernel) Q from $\mathcal{X}$ to $\mathcal{Z}$. When Q is contractive on the space of distributions, we have a strong data processing inequality.

Definition 9.1 (Strong data processing inequalities)**.** *Let $f : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ be convex and satisfy $f(1) = 0$. For distributions P_0, P_1 on $\mathcal{X}$ and a channel Q from $\mathcal{X}$ to a space $\mathcal{Z}$, define*

the marginal distribution $M_v(A) := \int Q(A \mid x) dP_v(x)$. *The channel* Q *satisfies a* strong data processing *inequality with constant* $\alpha \le 1$ *for the given* f*-divergence*

$$D_f\left(M_0 \| M_1\right) \le \alpha D_f\left(P_0 \| P_1\right)$$

for any choice of P_0, P_1 *on* $\mathcal{X}$. *For any such* f, *we define the* f*-strong data processing constant*

$$\alpha_f(Q) := \sup_{P_0 \neq P_1} \frac{D_f\left(M_0 \| M_1\right)}{D_f\left(P_0 \| P_1\right)}.$$

These types of inequalities are common throughout information and probability theory. Perhaps their most frequent use is in the development conditions for the fast mixing of Markov chains. Indeed, suppose the Markov kernel Q satisfies a strong data processing inequality with constant α with respect to variation distance. If π denotes the stationary distribution of the Markov kernel Q and we use the operator $\circ$ to denote one step of the Markov kernel,[1]

$$Q \circ P := \int Q(\cdot \mid x) dP(x),$$

then for any initial distribution π_0 on the space $\mathcal{X}$ we have

$$\|\underbrace{Q \circ \cdots \circ Q}_{k \text{ times}} \pi_0 - \pi\|_{\mathrm{TV}} \le \alpha^k \left\|\pi_0 - \pi\right\|_{\mathrm{TV}}$$

because $Q \circ \pi = \pi$ by definition of the stationary distribution. Thus, the Markov chain enjoys geometric mixing.

To that end, a common quantity of interest is the *Dobrushin* coefficient, which immediately implies mixing rates.

Definition 9.2. *The* Dobrushin coefficient *of a channel or Markov kernel* Q *is*

$$\alpha_{\mathrm{TV}}(Q) := \sup_{x,y} \left\|Q(\cdot \mid x) - Q(\cdot \mid y)\right\|_{\mathrm{TV}}.$$

The Dobrushin coefficient satisfies many properties, some of which we discuss in the exercises and others of which we enumerate here. The first is that

Proposition 9.1.1. *The Dobrushin coefficient is the strong data processing constant for the variation distance, that is,*

$$\alpha_{\mathrm{TV}}(Q) = \sup_{P_0 \neq P_1} \frac{\|Q \circ P_0 - Q \circ P_1\|_{\mathrm{TV}}}{\|P_0 - P_1\|_{\mathrm{TV}}}.$$

Proof There are two directions to the proof; one easy and one more challenging. For the easy direction, we see immediately that if $\mathbf{1}_x$ and $\mathbf{1}_y$ denote point masses at x and y, then

$$\sup_{P_0 \neq P_1} \frac{\|Q \circ P_0 - Q \circ P_1\|_{\mathrm{TV}}}{\|P_0 - P_1\|_{\mathrm{TV}}} \ge \sup_{x,y} \left\|Q(\cdot \mid x) - Q(\cdot \mid y)\right\|_{\mathrm{TV}}$$

as $\|\mathbf{1}_x - \mathbf{1}_y\|_{\mathrm{TV}} = 1$ for $x \neq y$.

[1] The standard notation is usually to right-multiply the measure P, so that the marginal distribution $M = PQ$ means $M(A) = \int Q(A \mid x) dP(x)$; we find our notation more intuitive.

The other direction—that $\|Q \circ P_0 - Q \circ P_1\|_{\rm TV} \le \alpha_{\rm TV} \|P_0 - P_1\|_{\rm TV}$—is is more challenging. For this, recall Lemma 2.2.4 characterizing the variation distance, and let $Q_\star(A) := \inf_y Q(A \mid y)$. Then by definition of the Dobrushin coefficient $\alpha = \alpha_{\rm TV}(Q)$, we evidently have $|Q(A \mid x) - Q_\star(A)| \le \alpha$. Let $M_v = \int Q(\cdot \mid x) dP_v(x)$ for $v \in \{0, 1\}$. By expanding $dP_0 - dP_1$ into its positive and negative parts, we thus obtain

$$\begin{aligned}
M_0(A) - M_1(A) &= \int Q(A \mid x)(dP_0 - dP_1)(x) \\
&= \int Q(A \mid x) \left[dP_0(x) - dP_1(x)\right]_+ - \int Q(A \mid x) \left[dP_1(x) - dP_0(x)\right]_+ \\
&\le \int Q(A \mid x) \left[dP_0(x) - dP_1(x)\right]_+ - \int Q_\star(A) \left[dP_1(x) - dP_0(x)\right]_+ \\
&= \int Q(A \mid x) \left[dP_0(x) - dP_1(x)\right]_+ - \int Q_\star(A) \left[dP_0(x) - dP_1(x)\right]_+ ,
\end{aligned}$$

where the final equality uses Lemma 2.2.4. But of course we then obtain

$$M_0(A) - M_1(A) = \int \left(Q(A \mid x) - Q_\star(A)\right) \left[dP_0(x) - dP_1(x)\right]_+ \le \alpha \int \left[dP_0 - dP_1\right]_+ = \alpha \|P_0 - P_1\|_{\rm TV} ,$$

where the inequality follows as $0 \le Q(A \mid x) - Q_\star(A) \le \alpha$ and the equality is one of the characterizations of the total variation distance in Lemma 2.2.4. □

A more substantial fact is that the Dobrushin coefficient upper bounds *every* other strong data processing constant.

Theorem 9.1.2. *Let $f : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$ satisfy $f(1) = 0$. Then for any channel Q,*

$$\alpha_{\rm TV}(Q) \ge \alpha_f(Q).$$

The theorem is roughly a consequence of a few facts. First, Proposition 9.1.1 holds. Second, without loss of generality we may assume that $f \ge 0$; indeed, replace $f(t)$ with $h(t) = f(t) - f'(1)t$ for any $f'(1) \in \partial f(1)$, we have $h \ge 0$ as $0 \in \partial h(1)$ and $D_h = D_f$. Third, any $f \ge 0$ with $0 \in \partial f(1)$ can be approximated arbitrarily accurately with functions of the form $h(t) = \sum_{i=1}^k a_i \left[t - c_i\right]_+ + \sum_{i=1}^k b_i \left[d_i - t\right]_+$, where $c_i \ge 1$ and $d_i \le 1$. For such h, an argument shows that

$$D_h(Q \circ P_0 \| Q \circ P_1) \le \alpha_{\rm TV}(Q) D_h(P_0 \| P_1),$$

which follows from the similarities between variation distance, with $f(t) = \frac{1}{2}|t|$, and the positive part functions $[\cdot]_+$.

There is a related result, which we do not prove, that guarantees that strong data processing constants for χ^2-divergences are the "worst" constants. In particular, if $QP = \int Q(\cdot \mid x) dP(x)$ denotes the application of one step of a channel Q to $X \sim P$, then the χ^2 contraction coefficient is

$$\alpha_{\chi^2}(Q) = \sup_{P_0 \ne P_1} \frac{D_{\chi^2}\left(QP_0 \| QP_1\right)}{D_{\chi^2}\left(P_0 \| P_1\right)}.$$

Then it is possible to show that for any twice continuously differentiable f on $\mathbb{R}_{++}$ with $f''(1) > 0$,

$$\alpha_{\chi^2}(Q) \le \alpha_f(Q), \tag{9.1.1}$$

and we also have $\alpha_{\chi^2}(Q) = \alpha_{\rm kl}(Q)$, so that the strong data processing inequalities for KL-divergence and χ^2-divergence coincide.

In our context, that of (constrained) minimax lower bounds, such data processing inequalities immediately imply somewhat sharper lower bounds than the (unconstrained) applications in previous chapters. Indeed, let us revisit the situation present in the local Fano bound, where we the KL divergence has a Euclidean structure as in the bound (8.4.6), meaning that $D_{\rm kl}\left(P_0 \| P_1\right) \le \kappa^2 \delta^2$ when our parameters of interest $\theta_v = \theta(P_v)$ satisfy $\rho(\theta_0, \theta_1) \le \delta$. We assume that the constraints $\mathcal{C}$ impose that the data X_i is passed through a channel Q with KL-data processing constant $\alpha_{\rm KL}(Q) \le 1$. In this case, in the basic Le Cam's method (8.3.2), an application of Pinsker's inequality yields that whenever $\rho(\theta_0, \theta_1) \ge 2\delta$ then

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, \mathcal{C}) \ge \frac{\Phi(\delta)}{2}\left[1 - \sqrt{\frac{n}{2} D_{\rm kl}\left(M_0 \| M_1\right)}\right] \ge \frac{\Phi(\delta)}{2}\left[1 - \sqrt{n\kappa^2 \alpha_{\rm KL}(Q)\delta^2/2}\right],$$

and the "standard" choice of δ to make the probability of error constant results in $\delta^2 = (2n\kappa^2\alpha_{\rm KL}(Q))^{-1}$, or the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, \mathcal{C}) \ge \frac{1}{4}\Phi\left(\frac{1}{\sqrt{2n\kappa^2\alpha_{\rm KL}(Q)}}\right),$$

which suggests an effective sample size degradation of $n \mapsto n\alpha_{\rm KL}(Q)$. Similarly, in the local Fano method in Chapter 8.4.1, we see identical behavior and an effective sample size degradation of $n \mapsto n\alpha_{\rm KL}(Q)$, that is, if without constraints a sample size of $n(\epsilon)$ is required to achieve some desired accuracy ϵ, with the constraint a sample size of at least $n(\epsilon)/\alpha_{\rm KL}(Q)$ is necessary.

9.2 Local privacy

In Chapter 7 on differential privacy, we define *locally private mechanisms* (Definition 7.2) as those for which there is no trust: individuals randomize their own data, and no central curator collects or analyzes and then privatizes the resulting statistics. With such privacy mechanisms, we can directly develop strong data processing inequalities, after which we can prove strong lower bounds on estimation. In this section, we (more or less) focus on one-dimensional quantities and Le Cam's two-point method for lower bounds, as they allow the most direct application of the ideas. We will later develop more sophisticated techniques.

We begin with our setting. We have a ε-differentially private channel Q taking inputs $x \in \mathcal{X}$ and outputting Z. Here, we allow *sequential interactivity*, meaning that the ith private variable Z_i may depend on both X_i and Z_1^{i-1} (see the graphical model in Figure 9.1), so that instead of the basic constraint in Definition 7.2 that $Q(A \mid x) \le e^\varepsilon Q(A \mid x')$ for all x, x', local differential privacy instead means

$$\frac{Q(Z_i \in A \mid X_i = x, z_1^{i-1})}{Q(Z_i \in A \mid X_i = x', z_1^{i-1})} \le e^\varepsilon \tag{9.2.1}$$

for all (measurable) sets A and inputs x, x', z_1^{i-1}. The key result is the following contraction inequality on the space of probabilities.

Theorem 9.2.1. *Let Q be an ε-locally differentially private channel from $\mathcal{X}$ to $\mathcal{Z}$. Then for any distributions P_0, P_1 inducing marginal distributions $M_v(\cdot) = \int Q(\cdot \mid x) dP_v(x)$,*

$$D_{\rm kl}\left(M_0 \| M_1\right) + D_{\rm kl}\left(M_1 \| M_0\right) \le 4(e^\varepsilon - 1)^2 \left\|P_0 - P_1\right\|_{\rm TV}^2.$$

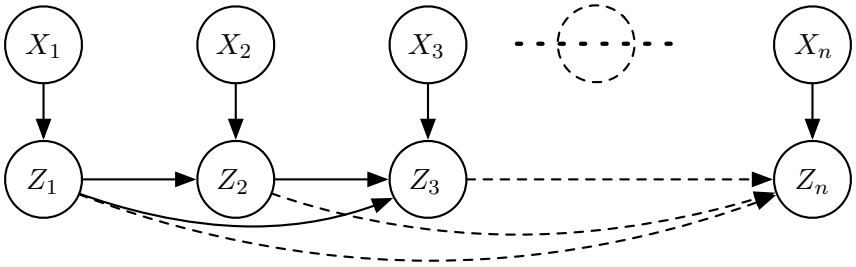

Figure 9.1. The sequentially interactive private observation model: the ith output Z_i may depend on X_i and the previously released Z_1^{i-1}.

Proof Without loss of generality, we assume that the output space $\mathcal{Z}$ is finite (by definition (2.2.3)), and let $m_v(z)$ and $q(z \mid x)$ be the p.m.f.s of M and Q, respectively, and let P_0 and P_1 have densities p_0 and p_1 with respect to a measure μ. Then

$$D_{\rm kl}\left(M_0 \| M_1\right) + D_{\rm kl}\left(M_1 \| M_0\right) = \sum_z (m_0(z) - m_1(z)) \log \frac{m_0(z)}{m_1(z)}$$

For any $a, b \geq 0$, we have $\log \frac{a}{b} = \log(1 + \frac{a}{b} - 1) \leq \frac{a}{b} - 1$, and similarly, $\log \frac{b}{a} \leq \frac{b}{a} - 1$. That is, $|\log \frac{a}{b}| \leq \frac{|a-b|}{\min\{a,b\}}$. Substituting above, we obtain

$$D_{\rm kl}\left(M_0 \| M_1\right) + D_{\rm kl}\left(M_1 \| M_0\right) \leq \sum_z \frac{(m_0(z) - m_1(z))^2}{\min\{m_0(z), m_1(z)\}}.$$

To control the difference $m_0(z) - m_1(z)$, note that for any fixed $x_0 \in \mathcal{X}$ we have

$$\int_{\mathcal{X}} q(z \mid x_0)(p_0(x) - p_1(x)) d\mu(x) = 0.$$

Thus

$$m_0(z) - m_1(z) = \int_{\mathcal{X}} (q(z \mid x) - q(z \mid x_0))(p_0(x) - p_1(x)) d\mu(x),$$

and so

$$\begin{aligned} |m_0(z) - m_1(z)| & \leq \sup_{x \in \mathcal{X}} |q(z \mid x) - q(z \mid x_0)| \int_{\mathcal{X}} |p_0(x) - p_1(x)| d\mu(x) \\ & = 2q(z \mid x_0) \sup_{x \in \mathcal{X}} \left(\frac{q(z \mid x)}{q(z \mid x_0)} - 1 \right) \left\| P_0 - P_1 \right\|_{\rm TV}. \end{aligned}$$

By definition of local differential privacy, $\frac{q(z|x)}{q(z|x_0)} - 1 \leq e^\varepsilon - 1$, and as x_0 was arbitrary we obtain

$$|m_0(z) - m_1(z)| \leq 2(e^\varepsilon - 1) \inf_x q(z \mid x) \left\| P_0 - P_1 \right\|_{\rm TV}.$$

Noting that $\inf_x q(z \mid x) \leq \min\{m_0(z), m_1(z)\}$ we obtain the theorem. □

To be able to apply this result to obtain minimax lower bounds for estimation as in Section 8.3, we need to address samples drawn from product distributions, even with the potential interaction (9.2.1). In this case, we consider sequential samples $Z_i \sim Q(\cdot \mid X_i, Z_1^{i-1})$ and define $M_v^n = \int Q(\cdot \mid x_1^n) dP_v(x_1^n)$ to be the marginal distribution over all the Z_1^n. Then we have the following corollary.

Corollary 9.2.2. *Assume that each channel $Q(\cdot \mid X_i, Z_1^{i-1})$ is ε_i-differentially private. Then*

$$D_{\rm kl}\left(M_0^n \| M_1^n\right) \le 4 \sum_{i=1}^n (e^{\varepsilon_i} - 1)^2 \left\|P_0 - P_1\right\|_{\rm TV}^2 .$$

Proof Recalling the chain rule (2.1.6) for the KL-divergence, we have

$$D_{\rm kl}\left(M_0^n \| M_1^n\right) = \sum_{i=1}^n \mathbb{E}_{M_0}\left[D_{\rm kl}\left(M_{0,i}(\cdot \mid Z_1^{i-1}) \| M_{1,i}(\cdot \mid Z_1^{i-1})\right)\right],$$

where the outer expectation is taken over Z_1^{i-1} drawn marginally from M_0^n, and $M_{v,i}(\cdot \mid z_1^{i-1})$ denotes the conditional distribution on Z_i given $Z_1^{i-1} = z_1^{i-1}$ when $X_1^n \stackrel{\rm iid}{\sim} P_v$. Writing this distribution out, we note that Z_i is conditionally independent of $X_{\setminus i}$ given X_i and Z_1^{i-1} by construction, so for any set A

$$\begin{aligned} M_{v,i}(A \mid z_1^{i-1}) &= \int Q(Z_i \in A \mid x_1^n, z_1^{i-1}) dP_v(x_1^n \mid z_1^{i-1}) = \int Q(Z_i \in A \mid x_i, z_1^{i-1}) dP_v(x_1^n \mid z_1^{i-1}) \\ &= \int Q(Z_i \in A \mid x_i, z_1^{i-1}) dP_v(x_i). \end{aligned}$$

Now we know that $Q(Z_i \in \cdot \mid x_i, z_1^{i-1})$ is ε_i-differentially private by assumption, so Theorem 9.2.1 gives

$$D_{\rm kl}\left(M_{0,i}(\cdot \mid z_1^{i-1}) \| M_{1,i}(\cdot \mid z_1^{i-1})\right) \le 4(e^{\varepsilon_i} - 1)^2 \left\|P_0 - P_1\right\|_{\rm TV}^2$$

for any realization z_1^{i-1} of Z_1^{i-1}. Iterating this gives the result. □

Local privacy is such a strong condition on the channel Q that it actually "transforms" the KL-divergence into a variation distance, so that even if two distributions P_0 and P_1 have infinite KL-divergence $D_{\rm kl}\left(P_0 \| P_1\right) = +\infty$—for example, if their supports are not completely overlapping—their induced marginals have the much smaller divergence $D_{\rm kl}\left(M_0 \| M_1\right) \le 4(e^\varepsilon - 1)^2 \left\|P_0 - P_1\right\|_{\rm TV}^2 \le 4(e^\varepsilon - 1)^2$. This transformation into a different metric means that even in estimation problems that should on their faces be easy become quite challenging under local privacy constraints; for example, minimax squared error for estimating the mean of a random variable with finite variance scales as $1/\sqrt{n}$ rather than the typical $1/n$ scaling in non-private cases (see Exercise 9.4).

Let us demonstrate how to apply Corollary 9.2.2 in a few applications. Our main object of interest is the private analogue of the minimax risk (8.1.1), where for a parameter $\theta : \mathcal{P} \to \Theta$, semimetric ρ, and loss Φ, for a family of channels $\mathcal{Q}$ we define the *channel-constrained minimax risk*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, \mathcal{Q}) := \inf_{\widehat{\theta}_n} \inf_{Q \in \mathcal{Q}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}\left[\Phi(\rho(\widehat{\theta}_n(Z_1^n), \theta(P)))\right]. \tag{9.2.2}$$

When we take $\mathcal{Q} = \mathcal{Q}_\varepsilon$ to be the collection of ε-locally differentially private (interactive) channels (9.2.1), we obtain the ε-locally private minimax risk.

A few examples showing lower (and upper) bounds for the private minimax risk (9.2.2) in mean estimation follow.

Example 9.2.3 (Bounded mean estimation)**:** Let $\mathcal{P}$ be the collection of distributions with supports on $[-b, b]$, where $0 < b < \infty$. Then for any $\varepsilon \ge 0$, the minimax squared error satisfies

$$\mathfrak{M}_n(\theta(\mathcal{P}), (\cdot)^2, \mathcal{Q}_\varepsilon) \gtrsim \frac{b^2}{(e^\varepsilon - 1)^2 n} + \frac{b^2}{n}.$$

The second term in the bound is the classic minimax rate for this collection of distributions. To see the first term, take Bernoulli distributions P_0 and $P_1 \in \mathcal{P}$, where for some $\delta \ge 0$ to be chosen, under P_0 we have $X = b$ with probability $\frac{1-\delta}{2}$ and $-b$ otherwise, while under P_1 we have $X = b$ with probability $\frac{1+\delta}{2}$ and $X = -b$ otherwise. Then $\|P_0 - P_1\|_{\rm TV} = \delta$, $\mathbb{E}_1[X] - \mathbb{E}_0[X] = 2b\delta$, and by Le Cam's method (8.3.3), for any ε-locally private channel Q and induced marginals M_0^n, M_1^n as in Corollary 9.2.2, we have

$$\begin{aligned}\mathfrak{M}_n(\theta(\mathcal{P}), (\cdot)^2, \{Q\}) \ge \frac{b^2\delta^2}{2}\left(1 - \sqrt{\frac{1}{2} D_{\rm kl}\left(M_0^n \| M_1^n\right)}\right) &\ge \frac{b^2\delta^2}{2}\left(1 - \sqrt{2(e^\varepsilon - 1)^2 n \left\|P_0 - P_1\right\|_{\rm TV}^2}\right) \\ &= \frac{b^2\delta^2}{2}\left(1 - \sqrt{2(e^\varepsilon - 1)^2 n \delta^2}\right).\end{aligned}$$

Setting $\delta^2 = \frac{1}{8n(e^\varepsilon - 1)^2}$ gives the claimed minimax bound. $\diamond$

Effectively, then, we see a reduction in the effective sample size: when ε is large, there is no change, but otherwise, the estimation error is similar to that when we observe a sample of size $n\varepsilon^2$.

Example 9.2.4 (Estimating the parameter of a uniform distribution)**:** In exercise 8.2, we show that estimating the parameter θ of a $\mathsf{Uniform}(\theta, \theta + 1)$ distribution has minimax squared error scaling as $1/n^2$. Under local differential privacy, this is impossible. Let $\mathcal{P} = \{\mathsf{Uniform}(\theta, \theta + 1), \theta \in [0, 1]\}$ be the collection of uniform distributions with the given supports. Letting P_0 and P_1 be $\mathsf{Uniform}(0, 1)$ and $\mathsf{Uniform}(\delta, 1 + \delta)$, respectively, where $\delta \ge 0$ is to be chosen, we have $\|P_0 - P_1\|_{\rm TV} = \delta$, while for any ε-differentially private channel Q and induced marginals M_0 and M_1,

$$D_{\rm kl}\left(M_0^n \| M_1^n\right) \le 4(e^\varepsilon - 1)^2 n \left\|P_0 - P_1\right\|_{\rm TV}^2 = 4(e^\varepsilon - 1)^2 n\delta^2.$$

Applying Le Cam's method (8.3.3) and taking $\delta \asymp \frac{1}{\sqrt{n(e^\varepsilon - 1)}}$, we thus have that if $\mathcal{Q}_\varepsilon$ denotes the collection of ε-locally differentially private channels,

$$\mathfrak{M}_n(\theta(\mathcal{P}), (\cdot)^2, \mathcal{Q}_\varepsilon) \gtrsim \frac{1}{(e^\varepsilon - 1)^2 n}.$$

When $\varepsilon \lesssim 1$, the best attainable rate thus scales as $\frac{1}{n\varepsilon^2}$. $\diamond$

In both the preceding examples, a number of simple estimators achieve the given minimax rates. The simplest is one based on the Laplace mechanism (Example 7.1.3): let $W_i \stackrel{\rm iid}{\sim} \mathsf{Laplace}(1)$, and set $Z_i = X_i + \frac{2b}{\varepsilon} W_i$ in Example 9.2.3 and $Z_i = X_i + \frac{2}{\varepsilon} W_i$ in Example 9.2.4. In the former, define $\widehat{\theta}_n = \overline{Z}_n$ to be the mean; in the latter, $\mathbb{E}[\overline{Z}_n] = \frac{\theta+1}{2}$, so $\widehat{\theta}_n = 2\overline{Z}_n - 1$ achieves the minimax rate.

More extreme examples are possible. Consider, for example, the problem of testing the support of a distribution, where we care only about distinguishing two distributions.

Example 9.2.5 (Support testing)**:** Consider the problem of testing between the support of two uniform distributions, that is, given n observations, we wish to test whether $P = P_0 = \mathsf{Uniform}[0, 1]$ or $P = P_1 = \mathsf{Uniform}[\theta, 1]$ for some $\theta \in (0, 1)$. We can ask the rate at which we may take $\theta \downarrow 0$ with n while still achieving non-trivial testing power. Without privacy, a simple (and optimal) test Ψ is to simply check whether any observation $X_i < \theta$, in which case we can trivially accept P_0 and reject P_1, otherwise accepting P_1. Then

$$P_0(X_i > \theta, \text{ all } i) = (1 - \theta)^n \text{ while } P_1(X_i > \theta, \text{ all } i) = 1.$$

So the summed probability of error

$$P_0(\Psi = 1) + P_1(\Psi = 0) = (1 - \theta)^n \le \exp(-\theta n),$$

and if $\theta \gg 1/n$ this tends to zero, while $\theta_n = \theta_0/n$ yields $\lim_n P_0(\Psi = 1) = e^{-\theta_0}$.

Consider now the private case. Then for any ε-differentially private channel Q and induced marginals M_0, M_1, we have $D_{\rm kl}\left(M_0^n \| M_1^n\right) \le 4n(e^\varepsilon - 1)^2 \left\|P_0 - P_1\right\|_{\rm TV}^2$ by Corollary 9.2.2 while $\|P_0 - P_1\|_{\rm TV} = \theta$. The Bretagnolle-Huber inequality (Proposition 2.2.8.(b)) thus guarantees that

$$\|M_0^n - M_1^n\|_{\rm TV}^2 \le 1 - \exp(-D_{\rm kl}\left(M_0^n \| M_1^n\right)) \le 1 - \exp(-4n(e^\varepsilon - 1)^2\theta^2).$$

Whenever $\theta \ll \frac{1}{\sqrt{n}}$, we have $\|M_0^n - M_1^n\|_{\rm TV} \to 0$, and so for *any* test based on the private data Z_1^n, the probabilities of error

$$\inf_\Psi \left\{P_0(\Psi(Z_1^n) \ne 0) + P_1(\Psi(Z_1^n) \ne 1)\right\} \ge 1 - \sqrt{1 - \exp(-c_\varepsilon n\theta^2)},$$

where $c_\varepsilon = 4(e^\varepsilon - 1)^2$. In the range that $\frac{1}{n} \ll \theta \ll \frac{1}{\sqrt{n}}$, then, there is an essentially exponential gap between the non-private and private cases. $\diamond$

9.3 Communication complexity

Communication complexity is a broad field, encompassing results establishing fundamental limits in streaming and online algorithms, memory-limited procedures, and (of course) in minimal communication in various fields. Recent connections between communication complexity and information-theoretic techniques have increased its applicability in statistical problems, which is our main motivation here, and to which we return in force in Section 9.4 to come. To motivate our approaches, however, we give a (necessarily limited) overview of communication complexity, along with some of the basic techniques and approaches, which then extend to statistical problems.

9.3.1 Classical communication complexity problems

The most basic problems in communication complexity are not really statistical, instead asking a simpler question: two entities (always named Alice and Bob) have inputs x, y and wish to jointly compute a function $f(x, y)$. The question is then how many bits—or other messages—Alice and Bob need to communicate to compute this value. Less abstractly, Alice and Bob have input domains $\mathcal{X}$ and $\mathcal{Y}$ (often, these are $\{0, 1\}^n$), and Alice receives a vector $x \in \mathcal{X}$ and Bob $y \in \mathcal{Y}$, each unknown to the other, and they jointly exchange messages until they can successfully evaluate $f(x, y)$. To abstract away any details of the computational model, we assume each has infinite computational power, which allows a focus on communication. To formulate this as communication, we consider a

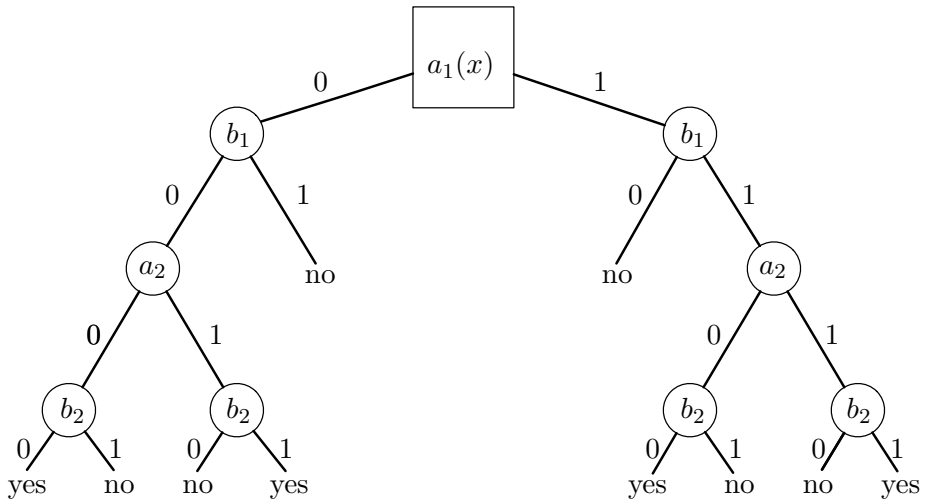

Figure 9.2. A communication tree representing testing equality for 2-dimensional bit strings $x, y \in \{0,1\}^2$. Internal nodes labeled a_j communicate the jth bit $a_j(x) = x_j$ of x, while internal nodes labeled b_j communicate the jth bit $b_j(y) = y_j$ of y. The maximum number of messages is 4. (A more efficient protocol is to have Alice send the entire string $x \in \{0,1\}^n$, then for Bob to check equality $x = y$ and output "Yes" or "No.")

protocol Π, which specifies the messages that each of Alice and Bob send to one another. We view this as a series of rounds, where at each round, the protocol allows one $\{0,1\}$-valued bit to be sent and determines who sends this bit, and, at termination time, can compute $f(x,y)$ based on the communicated message. Then the communication cost of Π is the maximum number of messages sent to (correctly) compute f over all inputs x, y.

A more convenient formulation for analysis is to consider a binary tree:

Definition 9.3. *A* protocol Π *over a domain* $\mathcal{X} \times \mathcal{Y}$ *with output space* $\mathcal{Z}$ *is a binary tree, where each internal node* v *is labeled with a mapping* $a_v : \mathcal{X} \to \{0,1\}$ *or* $b_v : \mathcal{Y} \to \{0,1\}$ *and each leaf is labeled with a value* $z \in \mathcal{Z}$*.*

Then to execute a communication protocol Π on input (x, y), we walk down the tree: beginning at the root node, for each internal node v labeled a_v (an Alice node) we walk left if $a_v(x) = 0$ and right if $a_v(x) = 1$, and each node v labeled b_v (a Bob node) we walk left if $b_v(y) = 0$ and right if $b_v(y) = 1$. Then the *communication cost* of the protocol Π is the height of the tree, which we denote by $\mathrm{depth}(\Pi)$. Figure 9.2 shows an example for testing the equality $x = y$ of two 2-dimensional bit strings $x, y \in \{0,1\}^2$.

In classical communication complexity, the main questions center around the *communication complexity* of a function $f : \mathcal{X} \to \mathcal{Y}$, which is the length of the shortest protocol that computes f correctly on all inputs: letting $\Pi_{\text{out}}(x, y)$ denote the final output of the protocol Π on inputs (x, y), this is

$$\mathsf{CC}(f) := \inf \left\{ \mathrm{depth}(\Pi) \mid \Pi_{\text{out}}(x,y) = f(x,y) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \right\}.$$

In many cases, it is useful to allow randomized communication protocols, which tolerate some probability of error; in this case, we let Alice and Bob each have access to (an arbitrary amount) of randomness, which we can identify without loss of generality with uniform random variables $U_a, U_b \overset{\text{iid}}{\sim} \mathsf{Uniform}[0,1]$, and the nodes a_v and b_v in Definition 9.3 are then mappings $a_v : \mathcal{X} \times [0,1] \to$

$\{0,1\}$ and $b_v : \mathcal{Y} \times [0,1] \to \{0,1\}$ and they calculate $a_v(\cdot, U_a)$ and $b_v(\cdot, U_b)$, respectively. Abusing notation slightly by leaving this randomness implicit, the *randomized communication complexity* for an accuracy δ is then the length of the shortest randomized protocol that calculates $f(x,y)$ correctly with probability at least $1-\delta$, that is,

$$\mathsf{RCC}_\delta(f) := \inf \left\{ \text{depth}(\Pi) \mid \mathbb{P}(\Pi_{\text{out}}(x,y) \neq f(x,y)) \leq \delta \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \right\}. \tag{9.3.1}$$

In the definition (9.3.1), we leave the randomization in Π implicit, and note that we require that the tree it induces still have a maximum length. We note that essentially any choice of $\delta > 0$ is immaterial, as we always have

$$\mathsf{RCC}_\delta(f) \leq O(1) \log \frac{1}{\delta} \cdot \mathsf{RCC}_{1/3}(f),$$

making all (constant) probability of error complexities essentially equivalent. (See Exercise 9.7.)

There are variants of randomized complexity that allow public randomness rather than private randomness, which can yield simpler algorithms and somewhat reduced complexity, but this improvement is limited, as Alice and Bob can always essentially simulate public randomness (see Exercise 9.8). Letting $\mathfrak{P}_{\text{pub}}$ be the collection of protocols in which both Alice and Bob have access to a shared random variable $U \sim \mathsf{Uniform}[0,1]$, we make the obvious extension

$$\mathsf{RCC}_\delta^{\text{pub}}(f) := \inf_{\Pi \in \mathfrak{P}_{\text{pub}}} \left\{ \text{depth}(\Pi) \mid \mathbb{P}(\Pi_{\text{out}}(x,y,U) \neq f(x,y)) \leq \delta \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \right\}.$$

Finally, we have *distributional* communication complexity, which for a probability measure μ on inputs $\mathcal{X} \times \mathcal{Y}$ is the depth of the shortest protocol that succeeds with a given μ-probability:

$$\mathsf{DCC}_\delta^\mu(f) := \inf \left\{ \text{depth}(\Pi) \mid \mu(\Pi_{\text{out}}(X,Y) \neq f(X,Y)) \leq \delta \right\}, \tag{9.3.2}$$

where the infimum is taken over *deterministic* protocols.

The final notion we consider is the *information complexity*. In this case, we require again that for each input pair x, y, the (potentially randomized) protocol $\Pi(x,y)$ still compute $f(x,y)$ correctly with probability at least $1-\delta$, but instead of measuring the depth of the tree, we let X, Y be drawn randomly from some distribution and measure the mutual information $I_2(X,Y;\Pi(X,Y))$. (We use base-2 logarithms to reflect bit communication.) In this case, we define

$$\mathsf{IC}_\delta(f) := \sup \inf_\Pi \left\{ I_2(X,Y;\Pi(X,Y)) \mid \mathbb{P}(\Pi_{\text{out}}(x,y) \neq f(x,y)) \leq \delta \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \right\}, \tag{9.3.3}$$

where the supremum is taken over joint distributions on (X,Y), the infimum over randomized protocols Π, and the right probability $\mathbb{P}$ is over any randomness in Π. There is a subtlety in this definition: we require Π to be accurate on *all* inputs (x,y), not just with probability over the distribution on (X,Y) in the information measure $I(X,Y;\Pi(X,Y))$. Relaxations to distributional variants of the information complexity (9.3.3) are also natural, as in the definition (9.3.2). Thus we sometimes consider the distributional information complexity

$$\mathsf{IC}_\delta^\mu(f) := \inf_\Pi \left\{ I_2(X,Y;\Pi(X,Y)) \mid \mu(\Pi_{\text{out}}(X,Y) \neq f(X,Y)) \leq \delta \right\},$$

where the infimum can be taken over deterministic or randomized protocols.

The different notions of communication complexity satisfy a natural ordering, making proving lower bounds for some notions (or conversely, developing low-communication methods for different protocols) much easier or harder than others. We record the standard inequalities in the coming proposition, which essentiall follows immediately from the operational interpretation of entropy as the average length of the best encoding of a signal (Section 2.4.1).

Proposition 9.3.1. *For any function f, $\delta \in (0,1)$, and probability measure μ on $\mathcal{X} \times \mathcal{Y}$,*

$$\mathsf{CC}(f) \ge \mathsf{RCC}_\delta(f) \ge \mathsf{RCC}_\delta^{\rm pub}(f) \ge \mathsf{DCC}_\delta^\mu(f) \ge \mathsf{IC}_\delta^\mu(f)$$

and

$$\mathsf{RCC}_\delta(f) \ge \mathsf{IC}_\delta(f).$$

Proof The first two inequalities are immediate. By Theorem 2.4.3, we have

$$\mathrm{depth}(\Pi) \ge H_2(\Pi) \ge H_2(\Pi) - H_2(\Pi \mid X, Y) = I_2(X, Y; \Pi(X, Y)),$$

and so for all $\delta \in (0, \frac{1}{2})$ we have both

$$\mathsf{RCC}_\delta(f) \ge \mathsf{IC}_\delta(f) \;\text{ and }\; \mathsf{DCC}_\delta^\mu(f) \ge \mathsf{IC}_\delta^\mu(f).$$

All that remains is to demonstrate $\mathsf{RCC}_\delta^{\rm pub}(f) \ge \mathsf{DCC}_\delta^\mu(f)$. For this, let Π be any protocol with public randomness U such that $\mathbb{P}(\Pi_{\rm out}(x, y, U) \ne f(x, y)) \le \delta$ for all x, y. Then by taking an expectation over $(X, Y) \sim \mu$, we obtain

$$\delta \ge \mathbb{E}_\mu \left[\mathbb{P}(\Pi_{\rm out}(X, Y, U) \ne f(X, Y) \mid X, Y)\right] \ge \inf_u \mu \left(\Pi_{\rm out}(X, Y, u) \ne f(X, Y)\right),$$

that is, there must be at least some u achieving the average error of Π, and the protocol Π is deterministic given u. So any protocol Π using public randomness to achieve probability of error δ can be modified into a deterministic protocol $\Pi(\cdot, \cdot, u)$ that achieves μ-probability of error δ.[2] □

Frequently, the first inequality in Proposition 9.3.1 is strict—even exponentially large—while the randomized complexity and information complexity end up being of roughly the same order. Understanding these differences is one of the major goals in communication complexity research.

9.3.2 Deterministic communication: lower bounds and structure

Deterministic communication complexity lower bounds often admit fairly elegant and somewhat elementary arguments, and the gaps between them and the randomized complexity highlight that we indeed expect providing lower bounds on randomized communication (9.3.1) or information (9.3.3) complexity to be quite challenging. The starting point, to which we will return when we consider randomized protocols, is to understand some structural aspects of the inputs and outputs of a protocol tree.

Recall that a set $R \subset \mathcal{X} \times \mathcal{Y}$ is a *rectangle* if it has the form $R = A \times B$ for some $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$. Equivalently, R is a rectangle if $(x_0, y_0) \in R$ and $(x_1, y_1) \in R$ imply that $(x_0, y_1) \in R$. As the next proposition shows, rectangular sets provide a key way to understand communication complexity.

Proposition 9.3.2. *Let v be a node in a deterministic protocol Π and R_v be those pairs (x, y) reaching node v. Then R_v is a rectangle.*

[2] This is one direction of Yao's minimax theorem [176], which states that communication complexity with public (shared) randomness and worst-case distributional complexity are identical: $\mathsf{RCC}_\delta^{\rm pub}(f) = \sup_\mu \mathsf{DCC}_\delta^\mu(f)$.

Proof We prove the result by induction. Certainly, for the root node v, we have $R_v = \mathcal{X} \times \mathcal{Y}$, which is a rectangle. Now, let v be an arbitrary (non-root) node in the tree and w its parent; assume w.l.o.g. that v is the left child of w and that in w, Alice speaks (that is, we use $a_w : \mathcal{X} \to \{0, 1\}$.) Then $R_w = A \times B$ by the inductive assumption. If $a_w(x) = 0$, then

$$R_v = \{\{x\} \times B \mid a_w(x) = 0, x \in A\} = \{\{x \mid a_w(x) = 0\} \cap A\} \cap B,$$

which is a rectangle. □

The structure of rectangles for correct protocols thus naturally determines the communication complexity of a function f. For a set $R \subset \mathcal{X} \times \mathcal{Y}$, we say R is f-constant if $f(x, y) = f(x', y')$ for all $(x, y) \in R$ and $(x', y') \in R$. Thus, any correct protocol Π necessarily partitions $\mathcal{X} \times \mathcal{Y}$ into a collection of f-constant rectangles, where we identify the rectangles with the leaves l of the protocol tree. In particular, Proposition 9.3.2 implies the following corollary.

Corollary 9.3.3. *Let N be the size of the minimal partition of $\mathcal{X} \times \mathcal{Y}$ into f-constant rectangles. Then $\mathsf{CC}(f) \geq \log_2 N$.*

Proof Any correct protocol Π partitions $\mathcal{X} \times \mathcal{Y}$ into the f-constant rectangles $\{R_l\}$ indexed by its leaves l. The minimal depth of a binary tree with at least N leaves is $\log_2 N$. □

A related corollary follows by considering *fooling sets*, which are basically sets that rectangles cannot contain.

Definition 9.4 (Fooling sets)**.** *A set $S \subset \mathcal{X} \times \mathcal{Y}$ is a* fooling set *for f if for any two pairs $(x_0, y_0) \in S$ and $(x_1, y_1) \in S$ satisfying $f(x_0, y_0) = f(x_1, y_1)$, at least one of the inequalities $f(x_0, y_1) \neq f(x_0, y_0)$ or $f(x_1, y_0) \neq f(x_0, y_0)$ holds.*

With this definition, the next corollary is almost immediate.

Corollary 9.3.4. *Let f have a fooling set S of size N. Then $\mathsf{CC}(f) \geq \log_2 N$.*

Proof By definition, no f-constant rectangle contains more than a single element of S. So the tree associated with any correct protocol Π has a single leaf for each element of S. □

An extension of the fooling set idea is the *rectangle measure* method, which proves that (for some probability measure P) the "size" of f-constant rectangles is small. By judicious choice of the probability, we can then demonstrate lower bounds.

Proposition 9.3.5. *Let P be a probability distribution on $\mathcal{X} \times \mathcal{Y}$. If all f-constant rectangles R have probability at most $P(R) \leq \delta$, Then $\mathsf{CC}(f) \geq \log_2 \frac{1}{\delta}$.*

Proof By the union bound, any f-constant partition of $\mathcal{X} \times \mathcal{Y}$ into rectangles $\{R_l\}_{l=1}^N$ satisfies $1 \leq \sum_{l=1}^N P(R_l) \leq N\delta$. So $N \geq \frac{1}{\delta}$, and the result follows by Corollary 9.3.3. □

With these results, we can provide lower bounds on two exemplar problems that will inform much of our coming development.

Example 9.3.6 (Equality)**:** Consider the problem of testing equality of two n-bit strings $x, y \in \{0,1\}^n$, letting $f = \mathrm{EQ}$ be $f(x,y) = 1$ if $x = y$ and 0 otherwise. Define the set $S = \{(x,x) \mid x \in \{0,1\}^n\}$, which has cardinality 2^n, and satisfies $f(x,x) = 1$ for all $(x,x) \in S$. That S is a fooling set is immediate: for any (x,x) and $(x',x') \in S$, if $x \neq x'$, then certainly $(x,x') \notin S$. So

$$n \leq \mathsf{CC}(\mathrm{EQ}) \leq n+1,$$

where the upper bound follows by letting Alice simply communicate the string x and Bob check if $x = y$, outputting 1 or 0 as $x = y$ or $x \neq y$. $\diamond$

The second example concerns inner products on $\mathbb{F}_2$, the field of arithmetic on the integers modulo 2 (that is, with bit strings); one could extend this to inner products in more complicated number systems (such as floating point), but the basic ideas are cleaner when we deal with bits.

Example 9.3.7 (Inner products on $\mathbb{F}_2$)**:** Consider computing the inner product $\mathrm{IP}_2(x,y) = \langle x, y \rangle \mod 2$ for n-bit strings $x, y \in \{0,1\}^n$, where addition is performed modulo 2. Rather than a constructing a fooling set directly, we use Proposition 9.3.5 and let P be the uniform distribution on $\{0,1\}^n \times \{0,1\}^n$. Let $R = A \times B$ be a rectangle with $\langle x, y \rangle = 0$ for all $x \in A$ and $y \in B$. The linearity of the inner product guarantees that $\langle x, y \rangle = 0$ for all $x \in \mathrm{span}(A)$ and $y \in \mathrm{span}(B)$, the (linear) spans of A and B in $\mathbb{F}_2^n$, respectively. Now recognize that $\mathrm{span}(A), \mathrm{span}(B) \subset \mathbb{F}_2^n$ are orthogonal subspaces of $\mathbb{F}_2^n$, and so their dimensions $d_0 = \dim(\mathrm{span}(A))$ and $d_1 = \dim(\mathrm{span}(B))$ satisfy $d_0 + d_1 \leq n$.

Noting that if $d_0 = \dim(A)$ then $|A| \leq 2^{d_0}$ in $\mathbb{F}_2^n$, we thus obtain $|R| \leq |A| \cdot |B| \leq 2^n$, which (under the uniform measure P) satisfies

$$P(R) \leq \frac{2^n}{2^{2n}} = 2^{-n}.$$

By Proposition 9.3.5, we thus have

$$n \leq \mathsf{CC}(\mathrm{IP}_2) \leq n+1,$$

where once again the upper bound follows by letting Alice simply communicate $x \in \{0,1\}^n$ and having Bob output $\langle x, y \rangle \mod 2$. $\diamond$

9.3.3 Randomization, information complexity, and direct sums

When we allow randomization, the complexity bounds can, in some cases, drastically change. Consider again the equality function in Example 9.3.6. When we allow randomization, we can achieve $O(\log n)$ complexity to check equality (with high probability).

Example 9.3.8 (Equality with randomization)**:** Let $x, y \in \{0,1\}^n$ and p be a prime number satisfying $n^2 \leq p \leq 2n^2$ (the Prime Number Theorem guarantees the existence of such a p). Let Alice choose a uniformly random number $U \in \{0, \ldots, p-1\}$ and compute the polynomial

$$a(U) = x_1 + x_2 U + x_3 U^2 + \cdots + x_n U^{n-1} \mod p.$$

Then Alice may communicate both U and $a(U)$ to Bob, which requires at most $2 \log_2 p \leq 4 \log_2 n + 2 \log 2$ bits. Then Bob checks whether

$$b(U) = y_1 + y_2 U + y_3 U^2 + \cdots + y_n U^{n-1} \mod p$$

satisfies $b(U) = a(U)$. If so, Bob outputs "Yes" (equality), and otherwise, Bob outputs "No." This protocol satisfies $\mathrm{depth}(\Pi) \le 4 \log_2 n + 1$. Moreover, if $x = y$, it is always correct, while if $x \neq y$, then the protocol is incorrect only if $a(U) = b(U)$, that is, U is a root of the polynomial

$$p(u) = \sum_{i=1}^{n} (x_i - y_i) u^{i-1}.$$

But this is a non-zero degree $n-1$ polynomial, which has at most $n-1$ roots (on the field $\mathbb{F}_p$; see Appendix A.1 for a brief review of polynomials). Thus for $x \neq y$ we have

$$\mathbb{P}(\Pi(x, y) \text{ fails}) = \mathbb{P}(a(U) = b(U)) \le \frac{n-1}{p} < \frac{1}{n},$$

and so $\mathsf{RCC}_{1/n}(\mathrm{EQ}) \le O(1) \log n$, exponentially improving over deterministic complexity.

In passing, we make two additional remarks. First, this protocol is one-way and non-interactive: Alice can simply send $O(\log n)$ bits. Second, we can achieve essentially any probability of success in the bound while still only paying logarithmically in communication, as taking $n^k \le p \le 2n^k$ for $k \ge 2$ yields $\mathsf{RCC}_{1/n^k}(\mathrm{EQ}) \le 2k \log_2 n + O(1)$. $\diamond$

Example 9.3.8 makes clear that any lower bounds on randomized communication complexity, or, relatedly, information complexity, will necessarily be somewhat more subtle than those we have presented for CC. We develop a few of the main ideas here. Because our focus is on information theoretic techniques, we pass over a few of the standard tools for proving lower bounds involving *discrepancy* and randomized inputs, touching on these in the bibliographic notes at the end of the chapter. One of our main goals will be to show that the information complexity of the inner product is indeed $\Omega(n)$, a much stronger result than Example 9.3.7. In contrast to the lower bounds we provide for minimax risk in most of this book, the focus in communication complexity is to take an *a priori* accurate estimator and demonstrate that it *requires* a certain amount of information to be communicated, rather than the contrapositive result that limited information yields inaccurate estimators. While these are clearly equivalent, it can be fruitful to use the perspective most relevant for the problem at hand.

Two main ideas form the basis for information complexity lower bounds: first, *direct sum* inequalitites, which show that computing a function on n inputs requires roughly order n more communication than computing it (or at least, one of the constituent functions making it up) on one. The second important insight is to provide lower bounds on the information necessary to compute different primitives, and the particular structure of even randomized communication protocols makes this possible. For the remainder of Section 9.3.3, we address the first of these, returning to the information complexity of primitives in Section 9.3.4.

Direct sum bounds and decomposition

To show direct sum inequalities, we demonstrate that computing some function on n inputs requires roughly n times the communication of single-input computation. In general, we consider functions f of the form

$$f(x_1^n, y_1^n) = g(h(x_1, y_1), h(x_2, y_2), \ldots, h(x_n, y_n)), \tag{9.3.4}$$

where g is the global function of the n primitives h, calling such functions decomposable with primitive h. Several problems have the decomposable structure (9.3.4); focusing on the case that the inputs $x, y \in \{0,1\}^n$ and $f(x, y) \in \{0, 1\}$, we have the following three immediate examples.

Example 9.3.9 (Composition in equality)**:** The equality function $f(x, y) = 1$ if $x \neq y$ and $f(x, y) = 0$ otherwise satisfies the decomposition (9.3.4), where $h(x_i, y_i) = \mathbf{1}\{x_i \neq y_i\}$ and g is the OR function $g(z) = \mathbf{1}\{\langle \mathbf{1}, z\rangle > 0\}$, which is 1 if any of $z_1, \ldots, z_n$ is non-zero, and 0 otherwise. $\diamond$

Example 9.3.10 (Decomposition of inner product)**:** The inner product in $\mathbb{F}_2$, $f(x, y) = \langle x, y\rangle \mod 2$, where $h(x_i, y_i) = x_i y_i$, and $g(z) = \langle \mathbf{1}, z\rangle \mod 2$, which satisfies $g(z) = 0$ if $\sum_{i=1}^n z_i$ is even and $g(z) = 1$ otherwise. $\diamond$

Example 9.3.11 (Decomposition of disjointness)**:** The set disjointness function $f(x, y) = \mathrm{DISJ}(x, y) := \mathbf{1}\{\langle x, y\rangle > 0\}$ arises when x, y are characteristic vectors of two subsets A, B of $[n]$, that is, $x_i = \mathbf{1}\{i \in A\}$ and $y_i = \mathbf{1}\{i \in B\}$. Then $f(x, y) = \mathbf{1}\{A \cap B \neq \varnothing\}$, which corresponds to g being the OR $g(z) = \mathbf{1}\{\langle \mathbf{1}, z\rangle > 0\}$ and h the AND function $h(x_i, y_i) = x_i y_i$. $\diamond$

While Example 9.3.8 makes clear that the decomposition (9.3.4) is not sufficient to guarantee a randomized complexity lower bound of order n, it will be useful.

To develop the main information complexity direct sum theorem showing that the information complexity of f is at least the sum of the complexities of its constituent primitives, we leverage what we term *plantable inputs*:

Definition 9.5. *Let $f : \mathcal{X}^n \times \mathcal{Y}^n \to \{0, 1\}$ have the decomposition (9.3.4), where the primitive h is $\{0, 1\}$-valued. The pair $(x, y) \in \mathcal{X}^n \times \mathcal{Y}^n$ admits a* planted solution *if for each $i \in \{1, \ldots, n\}$, all x_i', y_i', and vectors all*

$$x' = (x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \quad \text{and} \quad y' = (y_1, \ldots, y_{i-1}, y_i', y_{i+1}, \ldots, y_n),$$

we have $f(x', y') = h(x_i', y_i')$.

The binary inner product in Examples 9.3.7 and 9.3.10 has many plantable inputs: any of the 3^n pairs of vectors $x, y \in \{0, 1\}^n$ with $\langle x, y\rangle = 0$ admit planted solutions, as we have $x_i y_i = 0$ for each i. The set-disjointness problem, Example 9.3.11, has the same plantable inputs. For the equality function, only the 2^n pairs $x = y$ admit planted solutions.

We outline the key idea to our direct sum lower bounds. Because we define information complexity for protocols Π that are correct on all inputs with high probability, we can choose an arbitrary distribution on inputs $(x_1^n, y_1^n) \in \mathcal{X}^n \times \mathcal{Y}^n$. Thus we choose a *fooling distribution* μ for f, meaning that for $(X_i, Y_i) \stackrel{\text{iid}}{\sim} \mu$ the pair $(X_1^n, Y_1^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ always admits a planted solution (Definition 9.5). The next definition says this slightly differently.

Definition 9.6. *A distribution μ on $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a* fooling distribution *if all (x_1^n, y_1^n) in the support of the product μ^n admit planted solutions (Definition 9.5).*

Typically, fooling distributions μ require some dependence between X_i and Y_i—for example, in the inner product, we require $X_i Y_i = 0$, so that if $X_i = 1$ then $Y_i = 0$ and vice versa:

Example 9.3.12 (A fooling distribution for inner products and set disjointness)**:** Define the distribution μ on pairs $(x, y) \in \{0, 1\} \times \{0, 1\}$ as follows: let V be uniform on $\{0, 1\}$, and conditional on $V = 0$, set $X = 0$ and let $Y \sim \mathsf{Uniform}\{0, 1\}$; conditional on $V = 1$, set $Y = 0$ and let $X \sim \mathsf{Uniform}\{0, 1\}$. Then certainly $XY = 0$, and any set of pairs $(X_i, Y_i) \stackrel{\text{iid}}{\sim} \mu$ satisfy both that the binary inner product $\mathrm{IP}_2(X_1^n, Y_1^n) = \langle X_1^n, Y_1^n\rangle \mod 2 = 0$ and set disjointness $\mathrm{DISJ}(X_1^n, Y_1^n) = \mathbf{1}\{\langle X_1^n, Y_1^n\rangle > 0\} = 0$. $\diamond$

Fooling distributions, as in Example 9.3.12, make conditioning natural in information complexity. If $(X, Y) \sim \mu$, there is always a random variable V such that $X \perp\!\!\!\perp Y \mid V$, that is, X and Y are conditionally independent given V (trivially, we can take $V = X$). Thus, for function $h : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$, we define the *conditional information complexity*

$$\mathsf{CIC}_\delta^\mu(h) := \inf_\Pi \sup_V \left\{I(X, Y; \Pi(X, Y) \mid V) \text{ s.t. } \mathbb{P}(\Pi_{\text{out}}(x, y) \neq h(x, y)) \leq \delta \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}\right\},$$

where the infimum is over all (randomized) protocols and the supremum is over all random variables making X and Y conditionally independent with joint distribution $(X, Y) \sim \mu$. So if we can find a variable V making the mutual information $I(X, Y; \Pi(X, Y) \mid V)$ large for any correct protocol Π, the conditional information complexity of h is necessarily large.

With this, we obtain our main direct sum theorem for information complexity.

Theorem 9.3.13. *Let μ be a fooling distribution $\mathcal{X} \times \mathcal{Y}$ for a function f with primitive h. Then*

$$\mathsf{IC}_\delta(f) \geq n \cdot \mathsf{CIC}_\delta^\mu(h).$$

Proof Let $V = V_1^n \in \mathcal{V}^n$ be any random vector with i.i.d. entries making (X_i, Y_i) conditionally indpendent given V_i. Then for any protocol Π, we have

$$\begin{aligned} I(X_1^n, Y_1^n; \Pi) &= H(\Pi) - H(\Pi \mid X_1^n, Y_1^n) \\ &= H(\Pi) - H(\Pi \mid X_1^n, Y_1^n, V) \geq H(\Pi \mid V) - H(\Pi \mid X_1^n, Y_1^n, V) = I(X_1^n, Y_1^n; \Pi \mid V) \end{aligned}$$

because we have the Markov chain $V \to (X_1^n, Y_1^n) \to \Pi$. Using the chain rule for mutual information, where we recognize that X_1^n and Y_1^n are independent given V, we have

$$\begin{aligned} I(X_1^n, Y_1^n; \Pi \mid V) &= \sum_{i=1}^n I(X_i, Y_i; \Pi \mid V, X_1^{i-1}, Y_1^{i-1}) \\ &= \sum_{i=1}^n H(X_i, Y_i \mid V, X_1^{i-1}, Y_1^{i-1}) - H(X_i, Y_i \mid V, \Pi, X_1^{i-1}, Y_1^{i-1}) \\ &\geq \sum_{i=1}^n H(X_i, Y_i \mid V) - H(X_i, Y_i \mid V, \Pi) = \sum_{i=1}^n I(X_i, Y_i; \Pi \mid V) \end{aligned} \tag{9.3.5}$$

because conditioning reduces entropy and (X_i, Y_i) are independent of X_1^{i-1}, Y_1^{i-1} given V.

Now we come to the key reduction from the global protocol Π to one solving individual primitives. On inputs $(x, y) \in \mathcal{X} \times \mathcal{Y}$, define the simulated protocol $\Pi_{i,v}(x, y)$ so that given the vector $v_{\setminus i} \in \mathcal{V}^{n-1}$, Alice and Bob independently generate $(X_j^*, Y_j^*) \stackrel{\text{iid}}{\sim} \mu(\cdot \mid V_j = v_j)$ for $j \neq i$, which is possible because of the assumed conditional independence given V, yielding $X_{\setminus i}^* \in \mathcal{X}^{n-1}$ and $Y_{\setminus i}^* \in \mathcal{Y}^{n-1}$, respectively. They then execute the protocol $\Pi((X_{\setminus i}^*, x), (Y_{\setminus i}^*, y))$ (where we substitute x and y into input position i for each). Two key consequences of this simulation follow: that $\Pi_{i,v}$ is a δ-error protocol for the primitive h and that we have the distributional equality

$$(X_i, Y_i, V_i, \Pi_{i,v}(X_i, Y_i)) \stackrel{\text{dist}}{=} (X_i, Y_i, V_i, \Pi(X_1^n, Y_1^n)) \mid V_{\setminus i} = v_{\setminus i}, \tag{9.3.6}$$

that is, the joint over the simulated protocol is equal to that over the original protocol Π conditional on $V_{\setminus i} = v_{\setminus i}$. The latter claim (9.3.6) is essentially definitional; the former requires a bit more work.

To see that $\Pi_{i,v}$ is a δ-error protocol for the primitive h, note that by construction, $X^*_{\setminus i}$ and $Y^*_{\setminus i}$ are in the support of μ, and so admit planted solutions. In particular, $f((X^*_{\setminus i}, x), (Y^*_{\setminus i}, y)) = h(x, y)$, and so $\Pi_{i,v}$ is necessarily a δ-error protocol.

The distributional equality (9.3.6) guarantees that for any v we have

$$I(X_i, Y_i; \Pi(X_1^n, Y_1^n) \mid V_i, V_{\setminus i} = v_{\setminus i}) = I(X_i, Y_i; \Pi_{i,v}(X_i, Y_i) \mid V_i),$$

and as $\Pi_{i,v}$ is a δ-error protocol for h, we have

$$\inf_v I(X_i, Y_i; \Pi_{i,v}(X_i, Y_i) \mid V_i) \geq \mathsf{CIC}^\mu_\delta(h).$$

Substituting in the bound (9.3.5), we obtain

$$I(X_1^n, Y_1^n; \Pi) \geq \sum_{i=1}^n I(X_i, Y_i; \Pi \mid V) \geq \sum_{i=1}^n \inf_v I(X_i, Y_i; \Pi_{i,v}(X_i, Y_i) \mid V_i) \geq n\mathsf{CIC}^\mu_\delta(h),$$

as desired. □

With Theorem 9.3.13 in hand, we have our desired direct sum result, so that proving information complexity lower bounds reduces to providing lower bounds on the (conditional) information complexity of various 1-bit primitives. The following corollary highlights the theorem's applications to inner product and set disjointness (Examples 9.3.10 and 9.3.11).

Corollary 9.3.14. *Let f be the binary inner product $f(x, y) = \langle x, y \rangle \mod 2$ or the disjointness function $f(x, y) = \mathbf{1}\{\langle x, y \rangle > 0\}$. Let μ be the fooling distribution in Example 9.3.12. Then*

$$\mathsf{IC}_\delta(f) \geq n \cdot \mathsf{CIC}^\mu_\delta(h)$$

where $h(a, b) = ab$ is the product (or AND) function.

Exercise 9.10 explores similar techniques for the entrywise lesser than or equal function, showing similar complexity lower bounds.

9.3.4 The structure of randomized communication and communication complexity of primitives

Theorem 9.3.13 provides a powerful direct sum result that demonstrates that, at least if a problem admits planted solutions for (nearly) i.i.d. sampling, then the information complexity must scale at least linearly in the complexity of the primitives making up the function f. Thus, we turn to providing information lower bounds for computing different primitive functions. Our main tool will be to show that even randomized communication protocols essentially partition the input space $\mathcal{X} \times \mathcal{Y}$ into rectangles—in analogy with Proposition 9.3.2 in the deterministic case—which allows us to provide lower bounds. The broad idea is simple: if we have an accurate protocol for computing a certain function h, we must necessarily be able to distinguish between the distribution of Π on different inputs (x, y), as the fundamental connection between tests and variation distance (Proposition 2.3.1) reveals.

Our main goal now is to prove the following proposition, which gives a lower bound on the (conditional) information complexity of computing the AND of two bits.

Proposition 9.3.15. *Let $h(x, y) = xy$ for inputs $x, y \in \{0, 1\}$. Let μ be the fooling distribution in Example 9.3.12. Then*

$$\mathsf{CIC}_\delta^\mu(h) \ge \frac{1}{4}\left(1 - 2\sqrt{\delta(1-\delta)}\right).$$

We prove this proposition in the remainder of this section, noting that as an immediate corollary, we obtain the following lower bounds on the communication complexity of set disjointness and binary inner product.

Corollary 9.3.16. *Let f be the binary inner product $f(x, y) = \langle x, y\rangle \mod 2$ or the disjointness function $f(x, y) = \mathbf{1}\{\langle x, y\rangle > 0\}$. Then*

$$\mathsf{IC}_\delta(f) \ge \frac{n}{4}(1 - 2\sqrt{\delta(1-\delta)}).$$

To control the complexity of computing individual primitives, it proves easier to use metrics tied more directly to testing. To that end, we recall the connection between Hellinger distance and the mutual information, or Jensen-Shannon divergence, between a variable X and a single bit $B \in \{0, 1\}$ in Proposition 2.2.10, which gives that if $B \to Z$, where $Z \sim P_b$ conditional on $B = b$, then

$$I_2(Z; B) \ge d_{\rm hel}^2(P_0, P_1).$$

To apply this inequality, recall the fooling distribution μ for inner products in Example 9.3.12, where $V \sim \mathsf{Uniform}\{0, 1\}$ and conditional on $V = 0$ we set $X = 0$ and draw $Y \sim \mathsf{Uniform}\{0, 1\}$, and otherwise $Y = 0$ and $X \sim \mathsf{Uniform}\{0, 1\}$. Then for $V \to (X, Y)$ from this distribution, we have

$$I_2(X, Y; \Pi(X, Y) \mid V) = \frac{1}{2} I_2(Y; \Pi(0, Y) \mid V = 0) + \frac{1}{2} I_2(X; \Pi(X, 0) \mid V = 1).$$

Letting Q_{xy} denote the (conditional) distribution over Π on input bits $x, y \in \{0, 1\}$ and noting that X and Y above are each uniform on $\{0, 1\}$, we see that Proposition 2.2.10 applies and so

$$I_2(X, Y; \Pi(X, Y) \mid V) \ge \frac{1}{2} d_{\rm hel}^2(Q_{01}, Q_{00}) + \frac{1}{2} d_{\rm hel}^2(Q_{10}, Q_{00}).$$

Applying the triangle inequality that $(a-b)^2 \le (|a-c| + |c-b|)^2 \le 2(a-c)^2 + 2(b-c)^2$, we obtain the following lemma.

Lemma 9.3.17. *Let Π be any protocol acting on two bit inputs $x, y \in \{0, 1\}$, and let μ be the fooling distribution in Example 9.3.12. Let Q_{xy} be the distribution of $\Pi(x, y)$ on inputs x, y. Then*

$$I_2(X, Y; \Pi(X, Y) \mid V) \ge \frac{1}{4} d_{\rm hel}^2(Q_{01}, Q_{10}).$$

The last step in the proof of Proposition 9.3.15 is to demonstrate a property of (randomized) protocols Π analogous to the rectangular property of deterministic communcation that Propositions 9.3.2 and 9.3.5 demonstrate. In analogy with the output leaf in the tree for determinstc communication complexity, let τ be the *transcript* of the communication protocol, that is, its entire communication trace. Then we claim the following analog of Proposition 9.3.2 that the set of inputs resulting in a particular output in deterministic complexity is a rectangle in $\mathcal{X} \times \mathcal{Y}$.

Lemma 9.3.18. *Let Π be any randomized protocol with inputs in $\mathcal{X} \times \mathcal{Y}$. Then there exist functions q_x and q_y such that for any transcript τ,*

$$\mathbb{P}(\Pi(x, y) = \tau) = q_x(\tau) \cdot q_y(\tau).$$

Proof We may view any randomized protocol as a particular instatiation of a deterministic protocol $\Pi(\cdot, \cdot, u_a, u_b)$, where $u_a, u_b \in [0, 1]$ are realizations of the randomness available to Alice and Bob, respectively, inducing a particular binary communication tree. By Proposition 9.3.2, for any leaf l, the set

$$R_l(u_a, u_b) = \{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid \Pi(x, y, u_a, u_b) \text{ reaches } l\}$$

is a rectangle, that is, $R_l(u_a, u_b) = A_l(u_a) \times B_l(u_b)$ for sets $A_l(u) \subset \mathcal{X}$ and $B_l(u) \subset \mathcal{Y}$. Of course, the leaves l of the tree are in bijection with the entire transcript τ, so that if τ ends in leaf l, then

$$\mathbb{P}(\Pi(x, y) = \tau) = \mathbb{P}((x, y) \in R_l(U_a, U_b)) = \mathbb{P}(x \in A_l(U_a), y \in B_l(U_b))$$

where $U_a, U_b \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0, 1]$ are the the randomness Alice and Bob use, respectively.

Expanding this as an integral gives

$$\begin{aligned}\mathbb{P}(x \in A_l(U_a), y \in B_l(U_b)) &= \int_0^1 \int_0^1 \mathbf{1}\left\{x \in A_l(u_a)\right\} \mathbf{1}\left\{y \in B_l(u_b)\right\} du_a du_b \\ &= \mathbb{P}(x \in A_l(U_a))\mathbb{P}(y \in B_l(U_b)).\end{aligned}$$

Set $q_x(\tau) = \mathbb{P}(x \in A_l(U_a))$ and $q_y(\tau) = \mathbb{P}(y \in B_l(U_b))$. □

We thus have the following key *cut and paste* property, which shows that in some sense, Hellinger distances respect the "rectangular" structure of communication protocols.

Lemma 9.3.19. *Let Π be any protocol acting on inputs in $\mathcal{X} \times \mathcal{Y}$ and let $Q_{x,y}$ be the distribution of $\Pi(x, y)$ on inputs x, y. Then*

$$d_{\text{hel}}(Q_{x,y}, Q_{x',y'}) = d_{\text{hel}}(Q_{x,y'}, Q_{x',y}).$$

Proof Let $\mathcal{T}$ be the collection of all possible transcripts the protocol outputs. By Lemma 9.3.18 we have

$$\begin{aligned}d_{\text{hel}}^2(Q_{x,y}, Q_{x',y'}) &= \frac{1}{2} \sum_{\tau \in \mathcal{T}} \left(\sqrt{Q_{x,y}(\tau)} - \sqrt{Q_{x',y'}(\tau)}\right)^2 \\ &= \frac{1}{2} \sum_{\tau \in \mathcal{T}} \left(\sqrt{q_x(\tau) q_y(\tau)} - \sqrt{q_{x'}(\tau) q_{y'}(\tau)}\right)^2 = 1 - \sum_{\tau} \sqrt{q_x(\tau) q_y(\tau) q_{x'}(\tau) q_{y'}(\tau)}.\end{aligned}$$

Rearranging by the trivial modification $q_x q_y q_{x'} q_{y'} = q_x q_{y'} q_{x'} q_y$, we have the result. □

We now finalize the proof of Proposition 9.3.15. Substituting this cutting and pasting in Lemma 9.3.17 we have

$$I_2(X, Y; \Pi(X, Y) \mid V) \geq \frac{1}{4} d_{\text{hel}}^2(Q_{01}, Q_{10}) = \frac{1}{4} d_{\text{hel}}^2(Q_{00}, Q_{11}).$$

Then a simple lemma recalling the testing inequalities in Chapter 2.3.1 completes the proof of the proposition, because it guarantees that $4I_2(X, Y; \Pi(X, Y) \mid V) \geq 1 - 2\sqrt{\delta(1 - \delta)}$ no matter the choice of protocol Π, and so

$$\mathsf{CIC}_\delta^\mu(h) \geq \inf_\Pi I_2(X, Y; \Pi(X, Y) \mid V) \geq \frac{1}{4}\left(1 - 2\sqrt{\delta(1 - \delta)}\right).$$

Lemma 9.3.20. *Let Π be any δ-accurate protocol for computing $h(x, y) = xy$ and Q_{xy} be its distribution on inputs (x, y). Then $d_{\rm hel}^2(Q_{00}, Q_{11}) \ge 1 - 2\sqrt{\delta(1-\delta)}$.*

Proof Assume that Π computes the product $xy \in \{0, 1\}$ correctly with probability at least $1 - \delta$, that is, $\mathbb{P}(\Pi_{\rm out}(x, y) \neq xy) \le \delta$ for all $x, y \in \{0, 1\}$. By Le Cam's testing lower bounds (Proposition 2.3.1), we know that

$$\begin{aligned} 2\delta \ge \mathbb{P}(\Pi_{\rm out}(0,0) \neq 0) + \mathbb{P}(\Pi_{\rm out}(1,1) \neq 1) &\ge 1 - \|Q_{00} - Q_{11}\|_{\rm TV} \\ &\overset{(\star)}{\ge} 1 - d_{\rm hel}(Q_{00}, Q_{11})\sqrt{2 - d_{\rm hel}^2(Q_{00}, Q_{11})}, \end{aligned}$$

where inequality $(\star)$ follows from the inequalities in Proposition 2.2.7 relating Hellinger and total-variation distance. Let $d = d_{\rm hel}^2(Q_{00}, Q_{11})$ for shorthand. Then rearranging gives $d(2 - d) \ge (1 - 2\delta)^2$. Solving for d in $0 \ge d^2 - 2d + (1 - 2\delta)^2$ yields $d \ge 1 - \sqrt{1 - (1 - 2\delta)^2}$. Recognize that $1 - (1 - 2\delta)^2 = 4(\delta - \delta^2)$. □

9.4 Communication complexity in estimation

A major application combining strong data processing inequalities and communication is in the communication and information complexity of statistical estimation itself. In this context, we limit the amount of information—or perhaps bits—that a procedure may send about individual examples, and then ask to what extent this constrains the estimator. This has applications in situations in which the memory available to an estimator is limited, in situations with privacy—as we shall see—and of course, when we restrict the number of bits different machines storing distributed data may send.

We consider the following setting: m machines, or agents, have data X_i, $i = 1, \ldots, m$. Communication proceeds in rounds $t = 1, 2, \ldots, T$, where in each round t machine i sends datum $Z_i^{(t)}$. To allow for powerful protocols—with little restriction except that each machine i may send only a certain amount of information—we allow $Z_i^{(t)}$ to depend arbitrarily on the previous messages $Z_1^{(t)}, \ldots, Z_{i-1}^{(t)}$ as well as $Z_k^{(\tau)}$ for all $k \in \{1, \ldots, m\}$ and $\tau < t$. We visualize this as a public blackboard B, where in each round t each $Z_i^{(t)}$ is collected into $B^{(t)}$, along with the previous public blackboards $B^{(\tau)}$ for $\tau < t$, and all machines may read these public blackboards. Thus, in round t, individual i generates the communicated variable $Z_i^{(t)}$ according to the channel

$$Q_{Z_i^{(t)}}(\cdot \mid X_i, Z_{<i}^{(t)}, B^{(t-1)}) = Q_{Z_i^{(t)}}(\cdot \mid X_i, Z_{\to i}^{(t)}).$$

Here we have used the notation $Z_{<i} := (Z_1, \ldots, Z_{i-1})$, and we will use $Z_{\le i} := (Z_1, \ldots, Z_i)$ and similarly for superscripts throughout. We will also use the notation $Z_{\to i}^{(t)} = (B^{(1)}, Z_{<i}^{(t)})$ to denote all the messages coming into communication of $Z_i^{(t)}$. Figure 9.3 illustrates two rounds of this communication scheme.

We can provide lower bounds on the minimax risk of communication-constrained estimators by extending the data processing inequality approach we have developed. Our approach to the lower bounds, which we provide in Sections 9.4.1 and 9.4.2 to follow, is roughly as follows. First, we develop another *direct sum* bound, in analogy with Theorem 9.3.13, meaning that the difficulty of

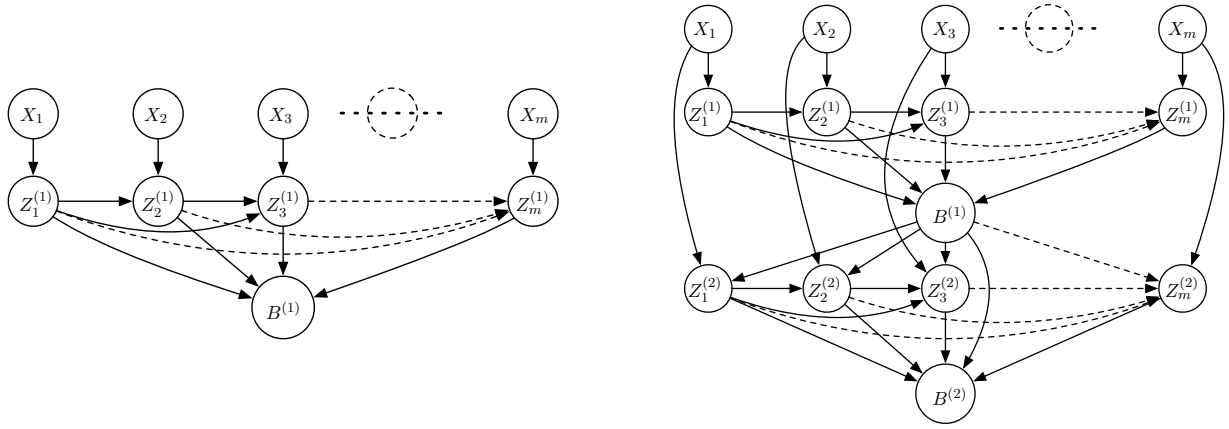

Figure 9.3. Left: single round of communication of variables, writing to public blackboard $B^{(1)}$. Right: two rounds of communication of variables, writing to public blackboards $B^{(1)}$ and $B^{(2)}$.

solving a d-dimensional problem is roughly d-times that of solving a 1-dimensional version of the problem; thus, any lower bounds on the error in 1-dimensional problems imply lower bounds for d-dimensional problems. Second, we provide an extension of the data processing inequalities we have developed thus far to apply to particular communication scenarios.

The key to our reductions is that we consider families of distributions where the coordinates of X are independent, which dovetails with Assouad's method. We thus index our distributions by $v \in \{0,1\}^d$, and in proving our lower bounds, we assume the typical Markov structure

$$V \to (X_1, \ldots, X_m) \to \Pi(X_1^m),$$

where V is chosen uniformly at random from $\{-1,1\}^d$, and $\Pi = \Pi(X_1^m)$ denotes the protocol of the entire communication—in this context, this is the entire set of blackboard messages

$$\Pi = (B^{(1)}, \ldots, B^{(T)}),$$

(which also encodes the message order). We assume that X follows a d-dimensional product distribution, so that conditional on $V = v$ we have

$$X \stackrel{\text{iid}}{\sim} P_v = P_{v_1} \otimes P_{v_2} \otimes \cdots \otimes P_{v_d}. \tag{9.4.1}$$

The generation strategy (9.4.1) guarantees that conditional on the jth coordinate $V_j = v_j$, the coordinates $X_{i,j}$ are i.i.d. and independent of $V_{\setminus j} = (V_1, \ldots, V_{j-1}, V_{j+1}, \ldots, V_d)$ as well as independent of $X_{i',j}$ for data points $i' \neq i$.

9.4.1 Direct sum communication bounds

Our first step is to argue that, if we can prove a lower bound on the information complexity of one-dimensional estimation, we can prove a lower bound on d-dimensional problems that scales with the dimension. To accomplish this reduction, let $X_{\le m,j} = (X_{i,j})_{i=1}^m$ be the jth coordinate of the data, and let $X_{\le m,\setminus j}$ be the remaining $d-1$ coordinates across all $i = 1, \ldots, m$. Then by the construction (9.4.1), we have the Markov structure

$$V_j \to X_{\le m,j} \to \Pi(X_1^m) \leftarrow X_{\le m,\setminus j} \leftarrow V_{\setminus j}.$$

In particular, viewing $X_{\le m,\setminus j}$ as extraneous randomness, we have the simpler Markovian structure

$$V_j \to X_{\le m,j} \to \Pi, \tag{9.4.2}$$

so that we may think of the communication $\Pi = \Pi(X_{\le m,j})$ as acting only on $X_{\le m,j}$. Now, define M_{-j} and M_j to be the marginal distributions over the total communication protocol Π conditional on $V_j = \pm j$, the one-variable model (9.4.2). Then Le Cam's testing equality (Proposition 2.3.1), and the equivalence between Hellinger and variation distance (Proposition 2.2.7) imply that

$$\begin{aligned}\inf_{\widehat{V}} 2\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(\Pi) \neq V_j) \ge \sum_{j=1}^d (1 - \|M_{-j} - M_{+j}\|_{\rm TV}) &\ge \sum_{j=1}^d (1 - \sqrt{2} d_{\rm hel}(M_{-j}, M_{+j})) \\ &\ge d\left(1 - \sqrt{\frac{2}{d}\sum_{j=1}^d d_{\rm hel}^2(M_{-j}, M_{+j})}\right)\end{aligned}$$

by Cauchy-Schwarz. Summarizing, we have the following

Proposition 9.4.1 (Assouad's method in communication)**.** *Let M_{+j} be the marginal distribution over Π conditional on $V_j = 1$ and M_{-j} be the marginal distribution of Π conditional on $V_j = -1$ in Markov structure (9.4.2) and assume X_i follow the product distribution (9.4.1). Then*

$$\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(\tau) \neq V_j) \ge \frac{d}{2}\left(1 - \sqrt{\frac{2}{d}\sum_{j=1}^d d_{\rm hel}^2(M_{-j}, M_{+j})}\right).$$

Recalling Assouad's method (Lemma 8.5.2) of Chapter 8.5, we see that any time we have a problem with separation with respect to the Hamming metric (8.5.1), we have a lower bound on its error in estimation problems. This proposition analogizes Theorem 9.3.13, in that small Hellinger distance between the individual marginals $M_{\pm j}$ necessarily makes the testing and estimation problems hard.

9.4.2 Communication data processing

We now revisit the data processing inequalities in Section 9.1, where we consider a variant that allows us to prove lower bounds for estimation problems with limited communication. It will be more notationally convenient in this section to use $V \in \{0, 1\}$ rather than $\{-1, 1\}$, so we do so without comment. Our starting point is a revised strong data processing inequality.

Definition 9.7. *Let P_0, P_1 be arbitrary distributions on a space $\mathcal{X}$, let $V \in \{0, 1\}$ uniformly at random, and conditional on $V = v$, draw $X \sim P_v$. Consider the Markov chain $V \to X \to Z$. The* mutual information strong data processing constant *$\beta(P_0, P_1)$ is*

$$\beta(P_0, P_1) := \sup_{X \to Z} \frac{I(V; Z)}{I(X; Z)},$$

where the supremum is taken over all conditional distributions (Markov kernels) from X to Z.

In contrast to Definition 9.1, in this definition we have a contraction over the "beginning" of the chain $V \to X$ rather than the distribution $X \to Z$. Identifying Z with a communication protocol $\Pi(X_1^m)$, this makes it possible to develop lower bounds on estimation and testing that then depend on the information $I(X; \Pi)$.

Distributions with bounded likelihood ratios provide one way to demonstrate a strong data processing inequality of the form in Definition 9.7, where in analogy with Theorem 9.2.1 we obtain a contraction inequality involving the total variation distance.

Proposition 9.4.2. *Let $V \to X \to Z$, where $X \sim P_v$ conditional on $V = v$. Let P_X and $P_X(\cdot \mid Z)$ denote the marginal and conditional distributions on X given Z, respectively. If $|\log \frac{dP_v}{dP_{v'}}| \le \alpha$ for all v, v', then*

$$I(V; Z) \le 4(e^\alpha - 1)^2 \mathbb{E}_Z \left[\|P_X(\cdot \mid Z) - P_X\|_{\rm TV}^2 \right] \le 2(e^\alpha - 1)^2 I(X; Z).$$

We leave the proof of this proposition as Exercise 9.12, as it follows by adapting the techniques we use to prove Theorem 9.2.1, with the main difference being the random variables with bounded likelihood ratios ($X \to Z$ versus $V \to X$). A brief example illustrates Proposition 9.4.2.

Example 9.4.3 (Bernoulli distributions)**:** Let $P_v = \mathsf{Bernoulli}(\frac{1+v\delta}{2})$ for $v \in \{-1, 1\}$. Then we have likelihood ratio bound

$$\left| \log \frac{dP_1}{dP_{-1}} \right| \le \log \frac{1+\delta}{1-\delta}$$

and so under the conditions of Proposition 9.4.2, for any Z we have

$$I(V; Z) \le 2 \left(\frac{1+\delta}{1-\delta} - 1 \right)^2 I(X; Z) = 2 \left(\frac{2\delta}{1-\delta} \right)^2 I(X; Z) \stackrel{(i)}{\le} 10\delta^2 I(X; Z),$$

where inequality (i) holds for $\delta \in [0, 1/10]$. $\diamond$

We now give the two main results connecting mutual information and the contraction-type bounds in Definition 9.7. To provide bounds using Proposition 9.4.1, we wish to control the Hellinger distance between individual marginals $M_{\pm j}$, so we consider single variables in the Markov chain

$$V \to (X_1, \ldots, X_m) \to \Pi,$$

where $V \in \{0, 1\}$. To state the coming theorems, we make a restriction on the data generation $V \to X$, calling distributions P_0 and P_1 (c, β)-*contractive* if

$$\beta(P_0, P_1) \le \beta \le 1 \quad \text{and} \quad \max \left\{ D_\infty \left(P_0 \| P_1 \right), D_\infty \left(P_1 \| P_0 \right) \right\} \le \log c, \tag{9.4.3}$$

where $D_\infty \left(\cdot \| \cdot \right)$ denotes the Rényi-∞-divergence. Proposition 9.4.2 shows that whenever such a c exists we certainly have $\beta(P_0, P_1) \le 2(c-1)^2$.

The next theorem then provides the basic information contraction inequality for single-variable communication.

Theorem 9.4.4. *Let $1 \le c < \infty$ and $\beta \le 1$. Let P_0 and P_1 be (c, β)-contractive (9.4.3) distributions on $\mathcal{X}$ and M_v, $v \in \{0, 1\}$ be the marginal distribution of the protocol Π conditional on $V = v$. Then*

$$d_{\rm hel}^2(M_0, M_1) \le \frac{7}{2}(c+1)\beta \cdot \min \left\{ I(X_1^m; \Pi(X_1^m) \mid V = 0), I(X_1^m; \Pi(X_1^m) \mid V = 1) \right\}.$$

The proof of Theorem 9.4.4 is quite complicated, so we defer it to Section 9.5.

We can use Theorem 9.4.4 to obtain bounds on the probability of error—detection of d-dimensional signals—in higher dimensional problems based on mutual information alone. Because the theorem provides a bound involving the minimum of the conditional mutual informations, we have substantial freedom to combine the direct-sum lower bounds in Section 9.4.1 to massage it into the mutual information between the data X_1^m and the protocol $\Pi(X_1^m)$.

We thus recall the definition (9.4.1) of our product distribution signals, where we assume that each individual datum $X_i = (X_{i,1}, \ldots, X_{i,d}) = (X_{i,j})_{j=1}^d$ belongs to a d-dimensional set and conditional on $V = v \in \{-1, 1\}^d$ has independent coordinates distributed as $X_{i,j} \sim P_{v_j}$. With this, we have the following theorem, which follows by a combination of Assouad's method (in the context of communication bounds, i.e. Proposition 9.4.1) and Theorem 9.4.4.

Theorem 9.4.5. *Let Π the entire communication protocol in Figure 9.3, $V \in \{-1,1\}^d$ be uniform, and generate $X_i \stackrel{\text{iid}}{\sim} P_v$, $i = 1, \ldots, m$ according to the independent coordinate distribution (9.4.1). Assume additionally that for each coordinate $j = 1, \ldots, d$, the coordinate distributions $P_{\pm v_j}$ are (c, β)-contractive (9.4.3). Then for any estimator $\widehat{V}$,*

$$\sum_{j=1}^{d} \mathbb{P}(\widehat{V}_j(\Pi) \neq V_j) \geq \frac{d}{2}\left(1 - \sqrt{7(c+1)\frac{\beta}{d} \cdot I(X_1, \ldots, X_m; \Pi \mid V)}\right).$$

Proof Under the given conditions, Proposition 9.4.1 and Theorem 9.4.4 immediately combine to give

$$\sum_{j=1}^{d} \mathbb{P}(\widehat{V}_j(\Pi) \neq V_j) \geq \frac{d}{2}\left(1 - \sqrt{7(c+1)\frac{\beta}{d}\sum_{j=1}^{d} \min_{v \in \{-1,1\}} I(X_{1,j}, \ldots, X_{m,j}; \Pi \mid V_j = v)}\right).$$

Certainly

$$\min_{v \in \{-1,1\}} I(X_{1,j}, \ldots, X_{m,j}; \Pi \mid V_j = v) \leq I(X_{1,j}, \ldots, X_{m,j}; \Pi \mid V_j).$$

Then, using that w.l.o.g. we may assume the $X_{i,j}$ are discrete, we obtain

$$\begin{aligned}
\sum_{j=1}^{d} I((X_{i,j})_{i=1}^m; \Pi \mid V_j) &= \sum_{j=1}^{d} \left[H((X_{i,j})_{i=1}^m \mid V_j) - H((X_{i,j})_{i=1}^m \mid \Pi, V_j)\right] \\
&\stackrel{(i)}{=} \sum_{j=1}^{d} \left[H((X_{i,j})_{i=1}^m \mid (X_{i,j'})_{i \leq m, j' < j}, V) - H((X_{i,j})_{i=1}^m \mid \Pi, V_j)\right] \\
&\leq \sum_{j=1}^{d} \left[H((X_{i,j})_{i=1}^m \mid (X_{i,j'})_{i \leq m, j' < j}, V) - H((X_{i,j})_{i=1}^m \mid (X_{i,j'})_{i \leq m, j' < j}, \Pi, V)\right] \\
&= \sum_{j=1}^{d} I((X_{i,j})_{i=1}^m; \Pi \mid V, (X_{i,j'})_{i \leq m, j' < j}) = I(X_1, \ldots, X_m; \Pi \mid V),
\end{aligned}$$

where equality (i) used the independence of $X_{i,j}$ from $V_{\setminus j}$ and $X_{i,j'}$ for $j' \neq j$ given V_j, and the inequality that conditioning reduces entropy. This gives the theorem. □

9.4.3 Applications: communication and privacy lower bounds

Let us now turn to a few different applications of our lower bounds on communication-constrained estimators. We evidently require two conditions: first, we must show that the distributions our data follows satisfy a strong (mutual information) data processing inequality (Definition 9.7). Second, we must provide a (good enough) upper bound on the mutual information $I(X_1, \ldots, X_m; \Pi \mid V)$ between the data points X_i and communication protocol. While there are many strategies to providing bounds and strong data processing inequalities, we focus mainly on situations with bounded likelihood ratio, where Proposition 9.4.2 directly provides the type of strong data processing inequality we require.

Communication lower bounds

Our first set of examples consideres direct communication bounds, where controlling $I(X_1^m; \Pi)$ is relatively straightforward. Assume the setting in the introduction to Section 9.4, where to establish our communication bounds we assume each machine $i = 1, \ldots, m$ may send at most B_i total bits of information throughout the entire communication protocol—that is, for each pair i, t, we have a bound

$$H(Z_i^{(t)} \mid Z_{\to i}^{(t)}) \le B_{i,t} \quad \text{and} \quad \sum_t B_{i,t} \le B_i \tag{9.4.4}$$

on the message from X_i in round t. (This is a weaker condition that $H(Z_i^{(t)}) \le B_{i,t}$ for each i, t.) With this bound, we can provide minimax lower bounds on communication-constrained estimator.

For our first collection, we consider estimating the parameters of d independent Bernoulli distributions in squared error. Let $\mathcal{P}_d$ be the family of d-dimensional Bernoulli distributions, where we let the parameter $\theta \in [0,1]^d$ be such that $P_\theta(X_j = 1) = \theta_j$. Then we have the following result.

Proposition 9.4.6. *Let $\mathfrak{M}_m(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \{B_i\}_{i=1}^m)$ denote the minimax mean-square error for estimation of a d-dimensional Bernoulli under the information constraint (9.4.4). Then*

$$\mathfrak{M}_m(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \{B_i\}_{i=1}^m) \ge c \min\left\{\frac{d}{m} \frac{d}{\frac{1}{m}\sum_{i=1}^m B_i}, d\right\},$$

where $c > 0$ is a numerical constant.

Proof By the standard Assouad reduction (Section 8.5), when we take coordinate distributions $P_{v_j} = \mathsf{Bernoulli}(\frac{1+\delta v_j}{2})$, we have a $c\delta^2$-separation in Hamming metric. Applying Theorem 9.4.5 and Example 9.4.3, we obtain the minimax lower bound, valid for $0 \le \delta \le \frac{1}{10}$, of

$$\mathfrak{M}_m(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \{B_i\}_{i=1}^m) \ge c\delta^2 d \left(1 - \sqrt{C\frac{\delta^2}{d} I(X_1, \ldots, X_m; \Pi \mid V)}\right).$$

Now, we note that for any Markov chain $V \to X \to Z$,

$$I(X; Z \mid V) = H(Z \mid V) - H(Z \mid X, V) = H(Z \mid V) - H(Z \mid X) \le H(Z) - H(Z \mid X) = I(X; Z).$$

Thus we obtain

$$\begin{aligned} I(X_1, \ldots, X_m; \Pi \mid V) &\le I(X_1, \ldots, X_m; \Pi) \\ &= \sum_{i=1}^m \sum_{t=1}^T I(X_1, \ldots, X_m; Z_i^{(t)} \mid Z_{\to i}^{(t)}). \end{aligned}$$

As the message $Z_i^{(t)}$ satisfies the conditional independence $Z_i^{(t)} \perp\!\!\!\perp X_{\setminus i} \mid Z_{\to i}^{(t)}, X_i$, this final quantity equals $\sum_{i,t} I(X_i; Z_i^{(t)} \mid Z_{\to i}^{(t)})$. But of course $I(X_i; Z_i^{(t)} \mid Z_{\to i}^{(t)}) \le H(Z_i^{(t)} \mid Z_{\to i}^{(t)}) \le B_{i,t}$, and so

$$\mathfrak{M}_m(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \{B_i\}_{i=1}^m) \ge c\delta^2 d \left(1 - \sqrt{C\frac{\delta^2}{d} \sum_{i,t} B_{i,t}}\right).$$

Choosing $\delta = \min\{1/10, \frac{d}{2C\sum_i B_i}\}$ gives the result. □

This result deserves some discussion. It is sharp in the case that the number of bits is of order d or less from each machine: when we set $B_i = d$, the lower bound becomes

$$\sup_\theta \mathbb{E}_\theta[\|\widehat{\theta}(\Pi) - \theta\|_2^2] \gtrsim \min\left\{\frac{d}{m} \cdot \frac{d}{d}, d\right\} = \frac{d}{m},$$

which is certainly achievable (each machine simply sends its entire vector $X_i \in \{0,1\}^d$). When machines communicate fewer than d bits, we have a tighter result; for example, if only k/m machines send d bits, and the rest communicate little, we obtain

$$\sup_\theta \mathbb{E}_\theta[\|\widehat{\theta}(\Pi) - \theta\|_2^2] \gtrsim \min\left\{\frac{d}{m} \cdot \frac{md}{kd}, d\right\} = \frac{d}{k},$$

which is similarly intuitive. The extension of these ideas to the case when each machine has an individual sample of size n is more challenging, as it requires tensorized variants of the strong data processing inequality in Definition 9.7; we provide remarks in the bibliographical section.

Lower bounds in locally private estimation

We return to the local privacy setting we consider in Section 9.2, except now we allow substantially more interaction. We treat local differential privacy in the communication model of Figure 9.3, where n individuals have data X_i which they wish to privatize, and proceed in rounds, releasing data $Z_i^{(t)}$ from individual i in round t. A natural setting is to assume each data release $Z_i^{(t)}$ is $\varepsilon_{i,t}$-differentially private: instead of the sequentially interactive model (9.2.1), we have

$$Q(Z_i^{(t)} \in A \mid X_i = x, Z_{\to i}^{(t)} = z_{\to i}^{(t)}) \le \exp(\varepsilon_{i,t}) \cdot Q(Z_i^{(t)} \in A \mid X_i = x', Z_{\to i}^{(t)} = z_{\to i}^{(t)}) \tag{9.4.5}$$

for each i, t and all possible $x, x', z_{\to i}^{(t)}$. At a more abstract level, rather than a particular privacy guarantee on each individual data release $Z_i^{(t)}$, we can assume a more global stability guarantee akin to the (average) KL-stability in interactive data analysis (Definition 5.1). Thus, let $\Pi(x_1^n)$ be the entire collection of communicated information in the protocol in Figure 9.3 on input data $x_1, \ldots, x_n$. Abusing notation to let $D_{\rm kl}\left(Z_0 \| Z_1\right)$ be the KL-divergence between the distributions of Z_0 and Z_1, as in Definition 5.1, we make the following definition to capture arbitrary interactions.

Definition 9.8 (Average KL-privacy)**.** *Let the samples $x_{\le n} \in \mathcal{X}^n$ and $x_{\le n}^{(i)} \in \mathcal{X}^n$ differ only in example i. Then the data release Π is $\varepsilon_{\rm kl}$-*KL-locally-private on average *if*

$$\frac{1}{n}\sum_{i=1}^n D_{\rm kl}\left(\Pi(x_{\le n}^{(i)}) \| \Pi(x_{\le n})\right) \le \varepsilon_{\rm kl}.$$

The following observation shows that for appropriate choices of $\varepsilon_{\rm kl}$, this is indeed weaker than the interactive guarantee (9.4.5).

Lemma 9.4.7. *Let the communication Q satisfy the interactive privacy guarantee (9.4.5) and Π be the induced communication protocol over rounds $t \le T$. Then*

$$\frac{1}{n}\sum_{i=1}^n D_{\rm kl}\left(\Pi(x_{\le n}^{(i)}) \| \Pi(x_{\le n})\right) \le \frac{1}{n}\sum_{i=1}^n \sum_{t=1}^T \min\left\{\varepsilon_{i,t}, \frac{3}{2}\varepsilon_{i,t}^2\right\}.$$

Proof Using the chain rule for the KL-divergence, we have for any j that

$$\begin{aligned} D_{\rm kl}\left(\Pi(x_{\le n}^{(j)})\|\Pi(x_{\le n})\right) &= \sum_{i=1}^n \sum_{t=1}^T \mathbb{E}\left[D_{\rm kl}\left(Q(Z_i^t \in \cdot \mid x_i^{(j)}, Z_{\to i}^{(t)})\|Q(Z_i^t \in \cdot \mid x_i, Z_{\to i}^{(t)})\right)\right] \\ &= \sum_{t=1}^T \mathbb{E}\left[D_{\rm kl}\left(Q(Z_i^t \in \cdot \mid x_j^{(j)}, Z_{\to i}^{(t)})\|Q(Z_i^t \in \cdot \mid x_j, Z_{\to i}^{(t)})\right)\right], \end{aligned}$$

where the expectation is taken over $Z_{\to i}^{(t)}$ in the protocol $\Pi(x_{\le n}^{(j)})$, and the second equality follows because $x_j^{(i)} = x_j$ for all j except index i. Now let P_0 and P_1 be arbitrary distributions whose densities satisfy $p_0(z)/p_1(z) \le e^\varepsilon$. Then

$$D_{\rm kl}\left(P_0\|P_1\right) \le \varepsilon \quad \text{and} \quad D_{\rm kl}\left(P_0\|P_1\right) \le \log\left(1 + D_{\chi^2}\left(P_0\|P_1\right)\right) \le \log\left(1 + (e^\varepsilon - 1)^2\right)$$

by Proposition 2.2.9. Then by inspection $\min\{\varepsilon, \log(1 + (e^\varepsilon - 1)^2)\} \le \min\{\varepsilon, \frac{3}{2}\varepsilon^2\}$ for all $\varepsilon \ge 0$. Returning to the initial KL-divergence sum, we thus obtain

$$\sum_{i=1}^n D_{\rm kl}\left(\Pi(x_{\le n}^{(i)})\|\Pi(x_{\le n})\right) \le \sum_{i=1}^n \sum_{t=1}^T \mathbb{E}\left[\min\left\{\varepsilon_{i,t}, \frac{3}{2}\varepsilon_{i,t}^2\right\}\right],$$

as desired. □

The key is that the average KL-local privacy guarantee is sufficient to provide a mutual information bound, thus allowing us to apply Theorem 9.4.5 as in the proof of Proposition 9.4.6.

Proposition 9.4.8. *Let Π be any $\varepsilon_{\rm kl}$-KL-locally-private on average protocol and assume that $X_1, \ldots, X_n$ are independent conditional on V. Then Then*

$$I(X_1, \ldots, X_n; \Pi(X_1^n) \mid V) \le n\varepsilon_{\rm kl}.$$

Proof The conditional independence of the X_i guarantees that

$$\begin{aligned} I(X_1^n; \Pi(X_1^n) \mid V) &= \sum_{i=1}^n H(X_i \mid X_1^{i-1}, V) - H(X_i \mid \Pi, X_1^{i-1}, V) \\ &\le \sum_{i=1}^n H(X_i \mid X_{\setminus i}, V) - H(X_i \mid \Pi, X_{\setminus i}, V) = \sum_{i=1}^n I(X_i; \Pi(X_1^n) \mid V, X_{\setminus i}). \end{aligned}$$

We abuse notation to let $\Pi^*(X_{\setminus i})$ be the marginal protocol (marginalizing over X_i). Then

$$I(X_i; \Pi(X_1^n) \mid V, X_{\setminus i}) = \mathbb{E}\left[D_{\rm kl}\left(\Pi(X_{\setminus i}, X_i)\|\Pi^*(X_{\setminus i})\right)\right] \le \mathbb{E}\left[D_{\rm kl}\left(\Pi(X_{\setminus i}, X_i)\|\Pi(X_{\setminus i}, X_i')\right)\right]$$

where the first expectation is taken over V and $X_j \stackrel{\rm iid}{\sim} P_v$ conditional on $V = v$ and the inequality uses convexity and draws X_i' independently. Summing over $i = 1, \ldots, n$, Definition 9.8 gives the result. □

Applying Theorem 9.4.5, we then obtain the following corollary.

Corollary 9.4.9. *Let the conditions of Theorem 9.4.5 hold. If the data release* Π *is* $\varepsilon_{\rm kl}$*-private on average, then*

$$\sum_{j=1}^{d} \mathbb{P}(\widehat{V}_j(\Pi) \neq V_j) \geq \frac{d}{2}\left(1 - \sqrt{7(c+1)\frac{\beta}{d}n\varepsilon_{\rm kl}}\right).$$

Specializing to the case that we wish to estimate a d-dimensional Bernoulli vector, where $X \in \{\pm 1\}$ has coordinates with $\mathbb{P}(X_j = 1) = \theta_j$, Example 9.4.3 gives the following minimax lower bound.

Corollary 9.4.10. *Let* $\mathfrak{M}_n(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \varepsilon_{\rm kl})$ *denote the minimax mean-square error for estimation of a d-dimensional Bernoulli under the* $\varepsilon_{\rm kl}$*-KL-locally-private-on-average constraint in Definition 9.8. Then*

$$\mathfrak{M}_n(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \varepsilon_{\rm kl}) \geq c \min\left\{d, \frac{d^2}{n\varepsilon_{\rm kl}}\right\}.$$

Proof By Corollary 9.4.9 and Example 9.4.3, we have minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \varepsilon_{\rm kl}) \gtrsim d\delta^2\left(1 - \sqrt{C\frac{\delta^2}{d}n\varepsilon_{\rm kl}}\right)$$

for a numerical constant C, which is valid for $\delta \lesssim 1$. Choose δ^2 to scale as $\min\{1, \frac{d}{n\varepsilon_{\rm kl}}\}$. □

When instead of the average KL-privacy we use the pure local differential privacy constraint (9.4.5), Lemma 9.4.7 implies the following.

Corollary 9.4.11. *Let* $\mathfrak{M}_n(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \varepsilon)$ *denote the minimax mean-square error for estimation of a d-dimensional Bernoulli where each data release is* $\varepsilon_{i,t}$*-locally differentially private (9.4.5), and* $\sum_{t=1}^{\infty} \varepsilon_{i,t} \leq \varepsilon$*. Then*

$$\mathfrak{M}_n(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \varepsilon) \geq c \min\left\{d, \frac{d^2}{n(\varepsilon \wedge \varepsilon^2)}\right\}.$$

9.5 Proof of Theorem 9.4.4

The proof proceeds in stages. The basic ideas are as follows:

1. Relate the Hellinger distance between the marginal distributions M_0 and M_1 of Π conditional on $V = 0$ or 1 to a sum of Hellinger distances between the marginal M_0 and an alternative M_i' where $X_i \sim P_1$ and $X_{\setminus i} \stackrel{\rm iid}{\sim} P_0$.
2. Provide a data processing inequality to relate $d_{\rm hel}(M_0, M_i')$ and the mutual information $I(X_i; \Pi)$ between the individual observation X_i and the protocol Π.
3. Use the standard chain rules for mutual information to finalize the theorem.

Step 1: sequential modification of marginals

We begin by relating the marginal distributions M_0 and M_1 by a sequence of one-variable changes. To that end, for bit vectors $b \in \{0,1\}^m$ define M_b to be the marginal distribution over the protocol $\Pi(X_1^m)$ generated from $(X_1, \ldots, X_m)$, where for each i we generate X_i by indpendently sampling

$$X_i \mid b \sim P_{b_i}. \tag{9.5.1}$$

For the standard basis vectors $e_1, \ldots, e_m$, we expect M_0 to be close to M_{e_l}, and thus hope for some type of tensorization behavior, where we can relate M_0 and M_1 via one-step changes from M_0 to M_{e_l}. The next lemma realizes this promise.

Lemma 9.5.1. *Let M_0, M_1, and M_{e_l} be as above. Then*

$$d_{\rm hel}^2(M_0, M_1) \le 7 \sum_{l=1}^m d_{\rm hel}^2(M_0, M_{e_l}). \tag{9.5.2}$$

Proof The proof crucially relies on the Euclidean structures that the Hellinger distance induces along with analogues of the cut-and-paste (the "rectangular" structure of inputs in communication protocols) properties from deterministic and randomized two-player communication. We assume without loss of generality that Π is discrete, as the Hellinger distance is an f-divergence and so can be arbitrarily approximated by discrete random variables.

First, we analogize the "rectangular" probabilistic structure of two-player communication protocols in Lemmas 9.3.18 and 9.3.19, which yields a multi-player cut-and-paste lemma.

Lemma 9.5.2 (cutting and pasting)**.** *Let $a, b, c, d \in \{0,1\}^m$ be bit vectors satisfying $a_i + b_i = c_i + d_i$ for each $i = 1, \ldots, m$. Then*

$$d_{\rm hel}^2(M_a, M_b) = d_{\rm hel}^2(M_c, M_d).$$

Proof We claim the following analogue of Lemma 9.3.18: for any $X_1^m = x_1^m$ and any communication transcript τ, we may write

$$Q(\Pi(x_1^m) = \tau \mid x_1^m) = \prod_{i=1}^m f_{i,x_i}(\tau) \tag{9.5.3}$$

for some functions f_{i,x_i}. Indeed, letting $\tau = \{z_i^{(t)}\}_{i \le n, t \le T}$ we have

$$Q(\Pi(x_1^m) = \tau \mid x_1^m) = \prod_{i,t} Q(z_i^{(t)} \mid x_1^m, z_{\to i}^{(t)}) = \prod_{i=1}^m \underbrace{\prod_{t=1}^T Q(z_i^{(t)} \mid x_i, z_{\to i}^{(t)})}_{=: f_{i,x_i}(\tau)}$$

where we use that message $z_i^{(t)}$ depends only on x_i and $z_{\to i}^{(t)}$. Then we can write $M_b(\Pi(X_1^m) = \tau)$ as a product using Eq. (9.5.3): integrating over independent $X_i \sim P_{b_i}$, we have

$$M_b(\Pi(X_1^m) = \tau) = \int Q(\tau \mid x_1^m) dP_{b_1}(x_1) \cdots dP_{b_m}(x_m) = \prod_{i=1}^m \underbrace{\int f_{i,\tau}(x_i) dP_{b_i}(x_i)}_{:= g_{i,b_i}(\tau)} = \prod_{i=1}^m g_{i,b_i}(\tau).$$

Taking M_a, M_b, M_c, M_d as in the statement of the lemma,

$$d_{\rm hel}^2(M_a, M_b) = 1 - \sum_\tau \sqrt{\prod_{i=1}^m g_{i,a_i}(\tau) g_{i,b_i}(\tau)}.$$

But as $a_i + b_i = c_i + d_i$ and each is $\{0,1\}$-valued, we certainly have $g_{i,a_i} g_{i,b_i} = g_{i,c_i} g_{i,d_i}$, and so the lemma follows. □

The second result we require is due to Jayram [115], and is the following:

Lemma 9.5.3. *Let $\{P_b\}_{b\in\{0,1\}^m}$ be any collection of distributions satisfying the cutting and pasting property $d_{\rm hel}^2(P_a, P_b) = d_{\rm hel}^2(P_c, P_d)$ whenever $a, b, c, d \in \{0,1\}^m$ satisfy $a + b = c + d$. Let $N = 2^k$ for some $k \in \mathbb{N}$. Then for any collection of bit vectors $\{b^{(i)}\}_{i=1}^N \subset \{0,1\}^m$ with $\langle b^{(i)}, b^{(j)}\rangle = 0$ for all $i \neq j$ and $b = \sum_i b^{(i)}$,*

$$\prod_{l=1}^{k}(1 - 2^{-l}) d_{\rm hel}^2(P_{\mathbf{0}}, P_b) \le \sum_{i=1}^{m} d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)}}).$$

We defer the technical proof to Section 9.5.1.

A computation shows that $\prod_{l=1}^{k}(1 - 2^{-l}) > \frac{2}{7}$. Lemma 9.5.3 nearly gives us our desired result (9.5.2), except that it requires a power of 2. To that end, let k_0 be the largest $k \in \mathbb{N}$ such that $2^{k_0} \le m$, and construct bit vectors $b^{(1)}, \ldots, b^{(2^{k_0})}$ satisfying $\sum_i b^{(i)} = \mathbf{1}$ and $1 \le \left\|b^{(i)}\right\|_0 \le 2$ for each i. Then Lemma 9.5.3, via the cutting-pasting property of the marginals M, implies

$$\frac{2}{7} d_{\rm hel}^2(M_0, M_1) \le \sum_{i=1}^{2^{k_0}} d_{\rm hel}^2(M_0, M_{b^{(i)}}) \le 2 \sum_{i=1}^{m} d_{\rm hel}^2(M_0, M_{e_i}),$$

where the second inequality again follows from Lemma 9.5.3 as $b^{(i)} = e_j$ or $e_j + e_{j'}$ for some basis vectors e_j, $e_{j'}$. This gives Lemma 9.5.1. □

Step 2: from Hellinger to Shannon information

Now we relate the strong data processing processing constants for mutual information in Definition 9.7 to compare Hellinger distances with mutual information. We claim the following lemma.

Lemma 9.5.4. *Let the conditions of Theorem 9.4.4 hold. Let M_0 and M_{e_l} be the marginal distributions over Π when X_i have the sampling distribution (9.5.1). Then for $l \in \{1, \ldots, m\}$,*

$$d_{\rm hel}^2(M_{e_l}, M_0) \le \frac{c+1}{2} \beta I(X_l; \Pi(X_1^m) \mid V = 0).$$

Proof Consider the following alternative distributions. Let $W \sim \mathsf{Uniform}\{0,1\}$, and draw $X' \in \mathcal{X}^m$ with independent coordinates according to

$$X_i' \stackrel{\rm iid}{\sim} P_0 \text{ if } W = 0 \quad \text{or} \quad X_i' \sim \begin{cases} P_0 & \text{if } i \neq l \\ P_1 & \text{if } i = l \end{cases} \text{ if } W = 1.$$

Then we have the Markov chain $W \to X' \to \Pi(X')$, and moreover,

$$W \to X_l' \to \Pi(X') \leftarrow X_{\setminus l}',$$

so that additionally $W \to X_l' \to \Pi(X')$ is a Markov chain. As a consequence, Definition 9.7 of the strong data processing inequality gives

$$I(W; \Pi(X')) \le \beta I(X_l'; \Pi(X')).$$

Using Proposition 2.2.10, we thus have

$$d_{\rm hel}^2(M_{e_l}, M_0) \le I(W; \Pi(X')) \le \beta I(X_l'; \Pi(X')). \tag{9.5.4}$$

It remains to relate $I(X_l'; \Pi(X'))$ to $I(X_l; \Pi(X) \mid V = 0)$. Here we bounded likelihood ratio between P_0 by P_1. Indeed, we have by the condition (9.4.3) that

$$P_0 \geq \frac{1}{c} P_1 \;\; \text{so} \;\; (c+1)P_0 \geq P_0 + P_1 \;\; \text{or} \;\; P_0 \geq \frac{2}{c+1} \frac{P_0 + P_1}{2}.$$

As a consequence, we have

$$\begin{aligned} I(X_l; \Pi(X_1^m) \mid V = 0) &= \int D_{\rm kl}\left(Q(\cdot \mid X_l = x) \| M_0\right) dP_0(x) \\ &\geq \frac{2}{c+1} \int D_{\rm kl}\left(Q(\cdot \mid X_l = x) \| M_0\right) \frac{dP_0(x) + dP_1(x)}{2} \\ &\geq \frac{2}{c+1} \int D_{\rm kl}\left(Q(\cdot \mid X_l = x) \| \overline{M}\right) \frac{dP_0(x) + dP_1(x)}{2} \\ &= \frac{2}{c+1} I(X_l'; \Pi(X')), \end{aligned}$$

where the second inequality uses that $\overline{M} = \int Q(\cdot \mid X_l = x) \frac{dP_0(x) + dP_1(x)}{2}$ minimizes the integrated KL-divergence (recall inequality (8.7.3)). Returning to inequality (9.5.4), we evidently have the result of the lemma. □

Step 3: Completing the proof of Theorem 9.4.4

By combining the tensorization Lemma 9.5.1 with the information bound in Lemma 9.5.4, we obtain

$$d_{\rm hel}^2(M_0, M_1) \leq 7 \sum_{i=1}^m d_{\rm hel}^2(M_0, M_{e_i}) \leq \frac{7}{2}(c+1)\beta \sum_{i=1}^m I(X_i; \Pi \mid V = 0).$$

By symmetry, we also have

$$d_{\rm hel}^2(M_0, M_1) \leq 7 \sum_{i=1}^m d_{\rm hel}^2(M_0, M_{e_i}) \leq \frac{7}{2}(c+1)\beta \sum_{i=1}^m I(X_i; \Pi \mid V = 1).$$

Now, we note that as the X_i are independent conditional on V (and w.l.o.g. for the purposes of mutual information, we may assume they are discrete), for any $v \in \{0, 1\}$ we have

$$\begin{aligned} \sum_{i=1}^m I(X_i; \Pi \mid V = v) &= \sum_{i=1}^m \left[H(X_i \mid V = v) - H(X_i \mid \Pi, V = v)\right] \\ &= \sum_{i=1}^m \left[H(X_i \mid X_1^{i-1}, V = v) - H(X_i \mid \Pi, V = v)\right] \\ &\leq \sum_{i=1}^m \left[H(X_i \mid X_1^{i-1}, V = v) - H(X_i \mid X_1^{i-1}, \Pi, V = v)\right] \\ &= \sum_{i=1}^m I(X_i; \Pi \mid X_1^{i-1}, V = v) = I(X_1, \ldots, X_m; \Pi \mid V = v), \end{aligned}$$

where the inequality used that conditioning decreases entropy. We thus obtain

$$d_{\rm hel}^2(M_0, M_1) \leq \frac{7}{2}(c+1)\beta \min_{v \in \{0,1\}} I(X_1, \ldots, X_m; \Pi \mid V = v)$$

as desired.

9.5.1 Proof of Lemma 9.5.3

We prove the result by induction. It is trivially true for $m = 1$, that is, $k = 0$, so now we consider the inductive case, that is, it holds for $m = 1, \ldots, 2^{k-1}$ and we consider $m = 2^k$.

First, we observe that if $\{u_i\}_{i=1}^N$ are arbitrary vectors and $\overline{u} = \frac{1}{N}\sum_{i=1}^N u_i$ is their mean, then

$$\sum_{i,j} \|u_i - u_j\|_2^2 = \sum_{i,j} \|u_i - \overline{u} + \overline{u} - u_j\|_2^2 = \sum_{i,j} \|u_i - \overline{u}\|_2^2 + \sum_{i,j} \|\overline{u} - u_j\|_2^2 = 2N \sum_{i=1}^N \|\overline{u} - u_i\|_2^2 .$$

Thus, if u_0 is any other vector, that $\overline{u}$ minimizes $\sum_i \|u_i - u\|_2^2$ over all u gives

$$\frac{1}{N} \sum_{1 \le i < j \le N} \|u_i - u_j\|_2^2 \le \sum_{i=1}^N \|u_i - \overline{u}\|_2^2 \le \sum_{i=1}^N \|u_i - u_0\|_2^2 . \tag{9.5.5}$$

Now, we return to the Hellinger distances. Evidently $2d_{\rm hel}^2(P_a, P_b) = \left\|\sqrt{p_a}(\cdot) - \sqrt{p_b}(\cdot)\right\|_2^2$, so that it is a Euclidean distance. As a consequence, for any pairwise disjoint collection of N bit vectors $b^{(i)}$, we have

$$\sum_{i=1}^N d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)}}) \ge \frac{1}{N} \sum_{1 \le i < j \le N} d_{\rm hel}^2(P_{b^{(i)}}, P_{b^{(j)}}) = \frac{1}{N} \sum_{1 \le i < j \le N} d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)} + b^{(j)}}) \tag{9.5.6}$$

where the inequality follows from (9.5.5) and the equality by the assumed cut-and-paste property. Now, we apply Baranyai's theorem, which says that we may decompose any complete graph K_N, where N is even, into $N - 1$ perfect matchings $\mathcal{M}_i$ with $N/2$ edges—necessarily, as they form a perfect matching—where each $\mathcal{M}_i$ is edge disjoint. Identifying the pairs $i < j$ with the complete graph, we thus obtain

$$\sum_{1 \le i < j \le N} d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)} + b^{(j)}}) = \sum_{l=1}^{N-1} \sum_{(i,j) \in \mathcal{M}_l} d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)} + b^{(j)}}). \tag{9.5.7}$$

Now fix $n \in \{1, \ldots, N-1\}$ and a matching $\mathcal{M}_n$. By assumption we have $\langle b^{(i)} + b^{(j)}, b^{(i')} + b^{(j')} \rangle = 0$ for any distinct pairs $(i, j), (i', j') \in \mathcal{M}_n$, and moreover, $\sum_{(i,j) \in \mathcal{M}_n} (b^{(i)} + b^{(j)}) = b$. Thus, our induction hypothesis gives that for any $l \in \{1, \ldots, N-1\}$ and any of our matchings $\mathcal{M}_n$, we have

$$\sum_{(i,j) \in \mathcal{M}_n} d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)} + b^{(j)}}) \ge d_{\rm hel}^2(P_{\mathbf{0}}, P_b) \prod_{l=1}^{k-1} (1 - 2^{-l}).$$

Substituting this lower bound into inequality (9.5.7) and using inequality (9.5.6), we obtain

$$\sum_{i=1}^N d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)}}) \ge \frac{1}{N} \cdot (N-1) d_{\rm hel}^2(P_{\mathbf{0}}, P_b) \prod_{l=1}^{k-1} (1 - 2^{-l}) = d_{\rm hel}^2(P_{\mathbf{0}}, P_b) \prod_{l=1}^{k} (1 - 2^{-l}),$$

where we have used $N = 2^k$.

9.6 Bibliography

Data processing inequalities originate with Dobrushin's study of central limit theorems for Markov chains [62, 63]; Dobrushin first proved Proposition 9.1.1 (see [63, Sec. 3.1]). Cohen et al. [50] show that the strong data processing constant for variation distance is the largest of the strong data processing constants (Theorem 9.1.2) for finite state spaces using careful linear algebraic techniques, also showing the opposite extremality (inequality (9.1.1)) of the χ^2 contraction coefficient [50, Proposition II.6.15] for finite state spaces. Del Moral et al. [61] and Polyanskiy and Wu [146] give related and approachable treatments for general alphabets, and Exercises 9.1 and 9.2 follow [61]. More broadly, strong data processing inequalities arise in many applications in communication, estimation, and some functional analysis [147, 146].

Communication complexity begins with Yao [176], which introduces the communication complexity setting we discuss in Section 9.3, making the connections between randomized complexities and public (shared) randomness. The standard classical reference for the subject is Kushilevitz and Nisan's book [124]. There are numerous techniques that we do not discuss, including so-called discrepancy lower bounds, which address both randomized and deterministic communication complexity; for example, these give the stronger lower bound that $\mathsf{DCC}_\delta(\mathrm{IP}_2) \geq n - O(1)$ [124, Example 3.29 and Exercise 3.30]. Communication complexity has uses far beyond the "standard" communication setting we have outlined, with more recent research showing how to use the techniques to provide lower bounds on the performance of algorithms in many computational models, such as streaming models and memory-limited computation [141, 148]. Our information complexity approach follows Bar-Yossef et al. [15]. Recent work has shown how communication lower bounds and strong data processing inequalities can be used to show the necessity of "memorization" in some natural problems in machine learning, where any learning procedure with good enough performance necessarily encodes substantial irrelevant information about a dataset [38].

Our treatment of communication complexity and its applications in estimation follows an approach Zhang et al. [178] originate. The particular techniques we adapt, involving direct sums and strong data processing in communication, we adapt from Braverman et al. [37] and Garg et al. [90]. Our results apply most easily to scenarios in which each machine or agent owns only a single data item, which allows application of Proposition 9.4.2; tensorizing this to multiple observations requires some care, but can be done with a truncation argument [178, 37] or more careful Sobolev inequalities [147]. Our extension to private estimation scenarios follows the paper [65], which also shows how to generalize to other variants of privacy.

9.7 Exercises

Exercise 9.1 (Approximating nonnegative convex functions)**:** Let $f : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ be a closed, nonnegative convex function.

(a) Show that there exists a sequence of piecewise linear functions f_n satisfying $f_{n-1} \leq f_n \leq f$ for all n and for which $f_n(x) \uparrow f(x)$ pointwise for all x s.t. $f(x) < \infty$, and $f_n(x) \uparrow \infty$ otherwise. *Hint:* Let $\mathcal{L}$ be the collection of linear functions below f, that is $\mathcal{L} = \{l \mid l(x) = a + bx, l(x) \leq f(x) \text{ for all } x\}$, and note that $f(x) = \sup\{l(x) \mid l \in \mathcal{L}\}$. (See Appendix C.2.) You may replace $\mathcal{L}$ with functions of the form $l(x) = f(x_0) + g(x - x_0)$, where $g \in \partial f(x_0)$ is a subderivative of f at x_0.

(b) Show that if for some $z_0 \in \mathbb{R}$ we have $f(z_0) = 0$, then one may take the functions f_n to be of the form $f_n(x) = \sum_{i=1}^n a_i [b_i - x]_+ + \sum_{i=1}^n c_i [x - d_i]_+$, where $b_i \leq z_0$, $d_i \geq z_0$, and $a_i, c_i \geq 0$.

(c) Conclude that for any measure μ on $\mathbb{R}_+$, $\int f_n d\mu \uparrow \int f d\mu$.

Exercise 9.2 (Proving Theorem 9.1.2)**:** In this question, we formalize the sketched proof of Theorem 9.1.2 by filling in details of the following steps. Let $\alpha = \alpha_{\rm TV}(Q)$ be the Dobrushin coefficient of the channel Q and $f : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ be a closed convex function.

(a) There exists a nondecreasing sequence f_n of piecewise linear functions, each of the form $f_n(x) = \sum_{i=1}^n a_i \left[b_i - x\right]_+ + \sum_{i=1}^n c_i \left[x - d_i\right]_+$, where $b_i \le 1$, $d_i \ge 1$, and $a_i, c_i \ge 0$. *Hint:* Exercise 9.1.

(b) Let $M_v(A) = \int Q(A \mid x) dP_v(x)$ for $v \in \{0, 1\}$ be the induced marginal distributions. Show that for any function of the form $h(t) = \left[t - \Delta\right]_+$, where $\Delta > 1$,

$$D_h\left(M_0 \| M_1\right) \le \alpha D_h\left(P_0 \| P_1\right) \tag{9.7.1}$$

by the following steps:

i. Define the set $\mathcal{X}(\Delta) := \{x \mid dP_0(x) \le \Delta dP_1(x)\}$. Argue that $\mathcal{X}(\Delta)$ must be non-null (i.e., have positive measure).

ii. Define the probability distribution P_Δ with density

$$dP_\Delta(x) = \frac{\Delta dP_1(x) - dP_0(x)}{\int \left[\Delta dP_1(x) - dP_0(x)\right]_+} \mathbf{1}\left\{x \in \mathcal{X}(\Delta)\right\}.$$

Argue that the measure

$$G = \Delta P_1 - (\Delta - 1) P_\Delta$$

is a probability distribution.

iii. Show that

$$D_h\left(P_0 \| P_1\right) = \left\|P_0 - G\right\|_{\rm TV}.$$

It may be useful to show that $dP_0 - dG \le 0$ on $\mathcal{X}(\Delta)$.

iv. Conclude that

$$D_h\left(P_0 \| P_1\right) \ge \frac{1}{\alpha} \left\|Q \circ P_0 - Q \circ G\right\|_{\rm TV} \ge \frac{1}{\alpha} D_h\left(Q \circ P_0 \| Q \circ P_1\right).$$

(c) Using the monotone convergence theorem, show that $D_f\left(M_0 \| M_1\right) \le \alpha D_f\left(P_0 \| P_1\right)$.

Exercise 9.3 (Markov chain mixing)**:** Consider a Markov chain $X_1, X_2, \ldots$ with transition distribution $P(\cdot \mid x)$ and stationary distribution π. Let $P^k(\cdot \mid x)$ denote the distribution of the Markov chain initialized in state x after k steps. Assume there exists some (finite) positive integer $k \in \mathbb{N}$ such that for any two initial states x_0, x_1, the Markov chain satisfies

$$\left\|P^k(\cdot \mid x_0) - P^k(\cdot \mid x_1)\right\|_{\rm TV} \le \beta < 1.$$

Show that the Markov chain enjoys *fast mixing* for any f divergence: if there is any n such that $D_f\left(P^n(\cdot \mid x) \| \pi\right) < \infty$, the Markov chain mixes exponentially quickly in that it satisfies

$$\limsup_n \frac{1}{n} \log D_f\left(P^n(\cdot \mid x) \| \pi\right) \le \frac{1}{k} \log \beta < 0.$$

In brief, as soon as one can demonstrate a constant gap in variation distance, one is guaranteed a Markov chain mixes geometrically.

Exercise 9.4: For $k \in [1, \infty]$, we consider the collection of distributions

$$\mathcal{P}_k := \{P : \mathbb{E}_P[|X|^k]^{1/k} \le 1\},$$

that is, distributions P supported on $\mathbb{R}$ with kth moment bounded by 1. We consider minimax estimation of the mean $\mathbb{E}[X]$ for these families under ε-local differential privacy, meaning that for each observation X_i, we observe a private realization Z_i (which may depend on Z_1^{i-1}) where Z_i is an ε-differentially private view of X_i. Let $\mathcal{Q}_\varepsilon$ denote the collection of all ε-differentially private channels, and define the (locally) private minimax risk

$$\mathfrak{M}_n(\theta(\mathcal{P}), (\cdot)^2, \varepsilon) := \inf_{\widehat{\theta}_n} \inf_{Q \in \mathcal{Q}_\varepsilon} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}[(\widehat{\theta}_n(Z_1^n) - \theta(P))^2].$$

(a) Assume that $\varepsilon \le 1$. For $k \in [1, \infty]$, show that there exists a constant $c > 0$ such that

$$\mathfrak{M}_n(\theta(\mathcal{P}_k), (\cdot)^2, \varepsilon) \ge c \left(\frac{1}{n\varepsilon^2}\right)^{\frac{k-1}{k}}.$$

(b) Give an ε-locally differentially private estimator achieving the minimax rate in part (a).

Exercise 9.5: Show that strong data processing inequality in Theorem 9.2.1 is sharp in the following sense. There exist ε-differentially private channels Q_ε such that for any Bernoulli distributions P_0 and P_1 and induced marginal distributions $M_{v,\varepsilon} = Q(\cdot \mid X = 1)P_v(X = 1) + Q(\cdot \mid X = 0)P_v(X = 0)$,

$$\frac{D_{\rm kl}\left(M_{0,\varepsilon} \| M_{1,\varepsilon}\right)}{\|P_0 - P_1\|_{\rm TV}^2} = \frac{\varepsilon^2}{2} + O(\varepsilon^3)$$

as $\varepsilon \downarrow 0$.

Exercise 9.6: We apply the results of Exercise 9.4 to a problem of estimation of drug use. Assume we interview a series of individuals $i = 1, \ldots, n$, asking whether each takes illicit drugs. Let $X_i \in \{0, 1\}$ be 1 if person i uses drugs, 0 otherwise, and define $\theta^* = \mathbb{E}[X] = \mathbb{E}[X_i] = P(X = 1)$. Instead of X_i we observe answers Z_i under differential privacy,

$$Z_i \mid X_i = x \sim Q(\cdot \mid X_i = x)$$

for a ε-differentially private Q with $\varepsilon < \frac{1}{2}$ (so that $(e^\varepsilon - 1)^2 \le 2\varepsilon^2$). Let $\mathcal{Q}_\varepsilon$ denote the family of all ε-differentially private channels, and let $\mathcal{P}$ denote the Bernoulli distributions with parameter $\theta(P) = P(X_i = 1) \in [0, 1]$ for $P \in \mathcal{P}$.

(a) Use Le Cam's method and the strong data processing inequality in Theorem 9.2.1 to show that the minimax rate for estimation of the proportion θ^* in absolute value satisfies

$$\mathfrak{M}_n(\theta(\mathcal{P}), |\cdot|, \varepsilon) := \inf_{Q \in \mathcal{Q}_\varepsilon} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}\left[|\widehat{\theta}(Z_1, \ldots, Z_n) - \theta(P)|\right] \ge c\frac{1}{\sqrt{n\varepsilon^2}},$$

where $c > 0$ is a universal constant.

(b) Give a rate-optimal estimator for this problem. That is, define an ε-differentially private channel Q and an estimator $\widehat{\theta}$ such that $\mathbb{E}[|\widehat{\theta}(Z_1^n) - \theta|] \le C/\sqrt{n\varepsilon^2}$, where C is a universal constant.

(c) Download the dataset at `http://web.stanford.edu/class/stats311/Data/drugs.txt`, which consists of a sample of 100,000 hospital admissions and whether the patient was abusing drugs (a 1 indicates abuse, 0 no abuse). Use your estimator from part (b) to estimate the population proportion of drug abusers: give an estimated number of users for $\varepsilon \in \{2^{-k}, k = 0, 1, \ldots, 10\}$. Perform each experiment several times. Assuming that the proportion of users in the dataset is the true population proportion, how accurate is your estimator?

Exercise 9.7: Show that the randomized communication complexity (9.3.1) satifies $\mathsf{RCC}_\delta(f) \le O(1) \log \frac{1}{\delta} \mathsf{RCC}_{1/3}(f)$ for any f and any $\delta < 1$.

Exercise 9.8 (From public to private randomness)**:** Consider the randomized complexity (9.3.1) and associated public-randomness complexity $\mathsf{RCC}_\delta^{\text{pub}}$. Let $\mathcal{X} = \mathcal{Y} = \{0,1\}^n$ and $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$, and let Π be a protocol using public randomness U such that $\max_{x,y} \mathbb{P}(\Pi(x, y, U) \ne f(x, y)) \le \epsilon$.

(a) Use Hoeffding's inequality to show that there are $k = \frac{\log 2}{\delta^2} n$ points $u_1, \ldots, u_k$ such that if $I \in [k]$ is chosen uniformly at random, then $\mathbb{P}(\Pi(x, y, u_I) \ne f(x, y)) \le \epsilon + \delta$.

(b) Give a protocol that uses no public randomness but whose communication complexity is at most $\text{depth}(\Pi) + O(1) \log \frac{n}{\delta}$.

(c) Conclude that $\mathsf{RCC}_\delta(f) \le \mathsf{RCC}_\delta^{\text{pub}}(f) + O(1) \log \frac{n}{\delta}$.

Exercise 9.9 (An information lower bound for indexing)**:** In the indexing problem in communication complexity, Alice receives an n-bit string $x \in \{0,1\}^n$ and Bob an index $y \in [n] = \{1, \ldots, n\}$, and the two communicate to evaluate x_y; set $f(x, y) = x_y$.

(a) Show that if Bob can send messages, the communication complexity of indexing satisfies $\mathsf{CC}(f) \le O(1) \log n$.

In the *one way* communication model, only Alice can send messages. Let μ be the uniform distribution on $(X, Y) \in \{0,1\}^n \times [n]$. We will show that $\mathsf{DCC}_\delta^\mu(f) \ge (1 - h_2(\delta))n$, where $h_2(p) = -p \log_2 p - (1-p) \log_2(1-p)$ is the binary entropy.

(b) Fix the index $Y = i$ and let $p_i = \mathbb{P}(\widehat{X}_i = X_i \mid Y = i)$ based on a protocol Π. Use Fano's inequality (Proposition 8.4.1) to argue that $h_2(p_i) \ge H_2(X_i \mid \Pi)$.

(c) Show that if Π is a δ-error one-way protocol under μ, then

$$I(X_1^n; \Pi) \ge (1 - h_2(\delta))n.$$

Exercise 9.10 (Information complexity for entrywise less or equal)**:** Consider the entrywise less than or equal to function $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ with $f(x, y) = \mathbf{1}\{x \preceq y\}$, so that $f(x, y) = 1$ if $x_i \le y_i$ for each i and 0 if there exists i such that $x_i > y_i$.

(a) Show that f has the decompositional structure (9.3.4). Give the functions g and h.

(b) Give a fooling distribution μ on $\mathcal{X} \times \mathcal{Y}$ for f.

(c) Use Theorem 9.3.13 and a modification of the proof of Proposition 9.3.15 to show that $\mathsf{IC}_\delta(f) \geq \frac{n}{4}(1 - 2\sqrt{\delta(1-\delta)})$. (This is order optimal, because $\mathsf{IC}_\delta(f) \leq \mathsf{CC}(f) \leq n + 1$ trivially.)

Exercise 9.11 (Lower bounds for private logistic regression)**:** This question is (likely) challenging. Consider the logistic regression model for $y \in \{\pm 1\}$, $x \in \mathbb{R}^d$, that

$$p_\theta(y \mid x) = \frac{1}{1 + \exp(-y\langle \theta, x \rangle)}.$$

For a distribution P on $(X, Y) \in \mathbb{R}^d \times \{\pm 1\}$, where $Y \mid X = x$ has logistic distribution, define the excess risk

$$L(\theta, P) := \mathbb{E}_P[\ell(\theta; X, Y)] - \inf_\theta \mathbb{E}_P[\ell(\theta; X, Y)]$$

where $\ell(\theta; x, y) = \log(1 + \exp(-y\langle x, \theta \rangle))$ is the logistic loss. Let $\mathcal{P}$ be the collection of such distributions, where X is supported on $\{-1, 1\}^d$. Following the notation of Exercise 8.4, for a channel Q mapping $(X, Y) \to Z$, define

$$\mathfrak{M}_n(\mathcal{P}, L, Q) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}[L(\widehat{\theta}(Z_1^n), P)],$$

where the expectation is taken over $Z_i \sim Q(\cdot \mid X_i, Z_1^{i-1})$. Assume that the channel releases are all (locally) ε-differentially private.

(a) Show that for all n large enough,

$$\mathfrak{M}_n(\mathcal{P}, L, Q) \geq c \cdot \frac{d}{n} \cdot \frac{d}{\varepsilon \wedge \varepsilon^2}$$

for some (numerical) constant $c > 0$.

(b) Suppose we allow additional passes through the dataset (i.e. multiple rounds of communication), but still require that all data Z_i released from X_i be ε-differentially private. That is, assume we have the (sequential and interactive) release schemes of Fig. 9.3, and we guarantee that

$$Z_i^{(t)} \sim Q(\cdot \mid X_i, B^{(1)}, \ldots, B^{(t)}, Z_1^{(t)}, \ldots, Z_{i-1}^{(t)})$$

is $\varepsilon_{i,t}$-differentially private, where $\sum_t \varepsilon_{i,t} \leq \varepsilon$ for all i. Does the lower bound of part (a) change?

Exercise 9.12: In this question, we prove Proposition 9.4.2.

(a) Show that if $p(v)$ and $p(v \mid x)$ denote the p.m.f.s of V and V conditional on $X = x$, then

$$e^{-\alpha} p(v) \leq p(v \mid x) \leq e^\alpha p(v).$$

(b) Show that

$$|p(v \mid z) - p(v)| \leq 2(e^\alpha - 1) \left\| P_X(\cdot \mid z) - P_X(\cdot) \right\|_{\mathrm{TV}}.$$

(c) Complete the proof of the proposition.

JCD Comment: A few additional exercises to add:

1. Prove Yao's minimax theorem.
2. Is there a clean "memorization" phenomenon to cover?

Chapter 10

Testing and functional estimation

When we wish to estimate a complete "object," such as the parameter θ in a linear regression $Y = X\theta + \varepsilon$, or a density when we observe $X_1, \ldots, X_n$ i.i.d. with a density f, the previous chapters give a number of approaches to proving fundamental optimality results and limits. In many cases, however, we wish to estimate *functionals* of a distribution or larger parameter, rather than the entire distribution or a high-dimensional parameter. Suppose we wish to estimate some statistic $T(P) \in \mathbb{R}$ of a probability distribution P. Then a naive estimator is to construct an estimate $\widehat{P}$ of P, and simply plug it in: use $\widehat{T} = T(\widehat{P})$. But frequently—and as we have seen in the preceding chapters—our ability to estimate $\widehat{P}$ may be limited, while various statistics of P may be easier to estimate. As a trivial example of this phenomenon, suppose we have an unknown distribution P supported on $[-1, 1]$, and we wish to estimate the statistic $T(P) = \mathbb{E}_P[X]$, its expectation. Then the trivial sample mean estimator

$$T_n := \overline{X}_n$$

satisfies $\mathbb{E}[(T_n - \mathbb{E}[X])^2] \le \frac{1}{n}$. But an estimator that first attempts to approximate the full distribution P via some $\widehat{P}$ and then estimate $\int x d\widehat{P}(x)$ is likely to incur substantial additional error.

Alternatively, we might wish to test different properties of distributions. In *goodness of fit testing*, we are given a sample $X_1, \ldots, X_n$ i.i.d. from a distribution Q, and we wish to distinguish whether $Q = P$ or Q is far from P. In related *two-sample tests*, we are given samples $X_1^n \stackrel{\text{iid}}{\sim} P$ and $Y_1^m \stackrel{\text{iid}}{\sim} Q$, and again wish to test whether $Q = P$ or Q and P are far from one another. For example, in a medical study, we may wish to distinguish whether there are significant differences between a treated population Q and control population P.

More broadly, we wish to develop tools to understand the optimality of different estimators and tests of *functionals*, by which we mean scalar valued parameters of a distribution P. Such parameters could include the norm $\|\theta\|_2$ of a regression vector, an estimate of the best possible expected loss $\inf_f \mathbb{E}_P[\ell(f(X), Y)]$ in a prediction problem, the distance $\|P - P_0\|_{\text{TV}}$ of a sampled population P from a reference P_0, or the probability mass of outcomes we have not observed in a study. This chapter develops a few of the tools to understand these problems.

10.1 Le Cam's convex hull method

Our starting point is to revisit Le Cam's method from Chapter 8.3, which focused on "two-point" methods to provide a lower bound on estimation error. We can substantially generalize this by instead comparing families of distributions that all induce separations between statistics of one

another, and then computing the distance between the convex hulls of the families. This leads to Le Cam's *convex hull method*, which we state abstractly and specialize later to different scenarios of interest. Let $\mathcal{P}$ be a collection of distributions on an underlying space $\mathcal{X}$, and let $\theta : \mathcal{P} \to \mathbb{R}^d$ be a parameter of interest. We say that two subsets $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_1$ are δ-*separated* in $\|\cdot\|$ if

$$\|\theta(P_0) - \theta(P_1)\| \geq \delta \text{ for all } P_0 \in \mathcal{P}_0 \text{ and } P_1 \in \mathcal{P}_1. \tag{10.1.1}$$

We do not require that all of $\mathcal{P}_0$ be somehow on one side or the other of the collection $\{\theta(P_1) \mid P_1 \in \mathcal{P}_1\}$ of parameters associated with $\mathcal{P}_1$, just that they be pairwise separate.

Let $\mathrm{Conv}(\mathcal{P})$ be the collection of mixtures of elements of $\mathcal{P}$, that is,

$$\mathrm{Conv}(\mathcal{P}) = \left\{ \sum_{i=1}^m \lambda_i P_i \mid m \in \mathbb{N}, \ \lambda \succeq 0, \ \langle \lambda, \mathbf{1} \rangle = 1, \ P_i \in \mathcal{P} \right\}.$$

Defining the minimax risk

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|) = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\|\widehat{\theta} - \theta(P)\| \right]$$

(note the temporary lack of sample size n), we then have the following generalization of inequality (8.3.3).

Theorem 10.1.1 (Le Cam's Convex Hull Lower Bound)**.** *Let $\mathcal{P}_0$ and $\mathcal{P}_1 \subset \mathcal{P}$ be δ-separated in $\|\cdot\|$. Then*

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|) \geq \frac{\delta}{2} \sup \left\{ \left[1 - \left\| \overline{P}_0 - \overline{P}_1 \right\|_{\mathrm{TV}} \right] \mid \overline{P}_0 \in \mathrm{Conv}(\mathcal{P}_0), \overline{P}_1 \in \mathrm{Conv}(\mathcal{P}_1) \right\}$$

Proof For any parameter θ, the separation $\|\theta(P_0) - \theta(P_1)\| \geq \delta$ and the triangle inequality guarantees that at least one of $\|\theta - \theta(P_0)\| \geq \delta/2$ or $\|\theta - \theta(P_1)\| \geq \delta/2$ holds for all pairs $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_1$. Let $\overline{P}_0 = \sum_{j=1}^m \alpha_j P_j$ and $\overline{P}_1 = \sum_{j=1}^m \beta_j Q_j$ for $P_j \in \mathcal{P}_0$ and $Q_j \in \mathcal{P}_1$, respectively, where α, β are convex combinations. Then by Markov's inequality,

$$\begin{aligned}
\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|) &\geq \frac{1}{2} \sum_{j=1}^m \alpha_j \mathbb{E}_{P_j} \left[\|\widehat{\theta} - \theta(P_j)\| \right] + \frac{1}{2} \sum_{j=1}^m \beta_j \mathbb{E}_{Q_j} \left[\|\widehat{\theta} - \theta(Q_j)\| \right] \\
&\geq \frac{\delta}{2} \left[\sum_{j=1}^m \alpha_j \mathbb{E}_{P_j} \left[\mathbf{1}\{\|\widehat{\theta} - \theta(P_j)\| \geq \delta/2\} \right] + \sum_{j=1}^m \beta_j \mathbb{E}_{Q_j} \left[\mathbf{1}\{\|\widehat{\theta} - \theta(P_j)\| \geq \delta/2\} \right] \right] \\
&\geq \delta \sum_{j=1}^m \left(\alpha_j \mathbb{E}_{P_j} \left[\inf_{P_0 \in \mathcal{P}_0} \mathbf{1}\{\|\widehat{\theta} - \theta(P_0)\| \geq \delta/2\} \right] + \beta_j \mathbb{E}_{Q_j} \left[\inf_{P_1 \in \mathcal{P}_1} \mathbf{1}\{\|\widehat{\theta} - \theta(P_1)\| \geq \delta/2\} \right] \right) \\
&= \frac{\delta}{2} \left(\mathbb{E}_{\overline{P}_0} \left[\inf_{P_0 \in \mathcal{P}_0} \mathbf{1}\{\|\widehat{\theta} - \theta(P_0)\| \geq \delta/2\} \right] + \mathbb{E}_{\overline{P}_1} \left[\inf_{P_1 \in \mathcal{P}_1} \mathbf{1}\{\|\widehat{\theta} - \theta(P_1)\| \geq \delta/2\} \right] \right).
\end{aligned}$$

Note that if we define $f_v(x) = \inf_{P \in \mathcal{P}_v} \mathbf{1}\{\|\widehat{\theta}(x) - \theta(P)\| \geq \delta/2\}$ for $v = 0, 1$, then $f_0 + f_1 \geq 1$. We claim the following lemma, which extends Le Cam's lemma (Proposition 2.3.1) to give

Lemma 10.1.2. *For any two distributions P_0 and P_1,*

$$\inf_{f_0 + f_1 \geq 1} \mathbb{E}_{P_0}[f_0] + \mathbb{E}_{P_1}[f_1] \geq 1 - \|P_0 - P_1\|_{\mathrm{TV}}.$$

We leave this form of total variation distance as an exercise (see Exercise 2.1). Substituting it into the display above, we find that for any $\overline{P}_v \in \mathrm{Conv}(\mathcal{P}_v)$, we have

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|) \geq \frac{\delta}{2}\left[1 - \left\|\overline{P}_0 - \overline{P}_1\right\|_{\mathrm{TV}}\right].$$

Taking a supremum over the $\overline{P}_v$ gives the theorem. □

10.1.1 The χ^2-mixture bound

Theorem 10.1.1 provides a powerful tool for developing lower bounds between collections of well-separated distributions. The most typical approach is to take the class $\mathcal{P}_0$ to consist of a single "base" distribution P_0, and then let $\mathcal{P}_1$ vary around P_0 in some prescribed way, so that for an index set $\mathcal{V}$, we let $\overline{P} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v$. Even so, when we have a sample of size n from one of the distributions, this results in a total variation quantity of the form

$$\left\|P_0^n - \overline{P^n}\right\|_{\mathrm{TV}} \quad \text{where} \quad \overline{P^n} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v^n,$$

yielding a mixture of product distributions—something frequently quite challenging to control.

The key technique here is to leverage the inequalities relating divergences from Chapter 2, which allows us to replace the variation distance with something more convenient. In previous chapters, this was the KL-divergence; now, instead, we use a χ^2-divergence, as it interacts much more nicely with the mixture product structure. Essentially, we replace an expectation over $X \sim P$ with two expectations: one over $X \sim P$ and another over independent samples $V, V' \sim \mathsf{Uniform}(\mathcal{V})$. To obtain the bound, first note that

$$2\left\|P_0 - \overline{P}\right\|_{\mathrm{TV}}^2 \leq D_{\mathrm{kl}}\left(\overline{P}\|P_0\right) \leq \log(1 + D_{\chi^2}\left(\overline{P}\|P_0\right)) \leq D_{\chi^2}\left(\overline{P}\|P_0\right)$$

by Propositions 2.2.8 and 2.2.9.

We then have the following technical lemma.

Lemma 10.1.3. *Let $\overline{P} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v$ and P_v and P_0 have densities p_v, p_0 with respect to some base measure μ on a set $\mathcal{X}$. Then*

$$D_{\chi^2}\left(\overline{P}\|P_0\right) = \frac{1}{|\mathcal{V}|^2}\sum_{v,v'\in\mathcal{V}}\int \frac{p_v(x)p_{v'}(x)}{p_0(x)}d\mu(x) - 1 = \frac{1}{|\mathcal{V}|^2}\sum_{v,v'\in\mathcal{V}} \mathbb{E}_0\left[\frac{p_v(X)p_{v'}(X)}{p_0^2(X)}\right] - 1,$$

where the expectation is taken with respect to $X \sim P_0$. More generally, let $V \in \mathcal{V}$ be a random variable distributed according to π and conditional on $V = v$, let $X \mid V = v \sim P_v$. Then for the paired likelihood ratio $l(x \mid v, v') = \frac{p_v(x)p_{v'}(x)}{p_0^2(x)}$, the marginal distribution $\overline{P}$ of X satisfies

$$D_{\chi^2}\left(\overline{P}\|P_0\right) = \mathbb{E}_0\left[l(X \mid V, V')\right] - 1,$$

where the expectation is taken jointly over $X \sim P_0$ and $V, V' \stackrel{\mathrm{iid}}{\sim} \pi$.

Proof The starting point is to notice that for any two distributions P and Q we have $D_{\chi^2}\left(P\|Q\right) = \int (dP/dQ - 1)^2 dQ = \int \frac{dP^2}{dQ} - 2\int \frac{dP}{dQ} dQ + \int dQ = \int \frac{dP^2}{dQ} - 1$. Then we proceed by recognizing that $(\frac{1}{N}\sum_{i=1}^N x_i)^2 = \frac{1}{N^2}\sum_{i,j} x_i x_j$ for any sequence x_i, and so

$$D_{\chi^2}\left(\overline{P}\|P_0\right) + 1 = \int \frac{((1/|\mathcal{V}|)\sum_{v\in\mathcal{V}} dP_v)^2}{dP_0} = \frac{1}{|\mathcal{V}|^2}\sum_{v,v'\in\mathcal{V}} \int \frac{dP_v dP_{v'}}{dP_0}$$

as desired. The second statement has identical proof to the first except that we replace $\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}$ with expectations according to π. □

The applications of this lemma are many, and going through a few examples will best show how to leverage it. Roughly, our typical approach is the following: we identify $\mathcal{V}$ with $\{\pm 1\}^d$ or some other suitably nice collection of vectors. We then choose distributions P_v and P_0 with densities suitably nice that the ratios p_v/p_0 "act" like exponentials involving inner products of $v \in \mathcal{V}$ with some other quantity; then, because v is uniform in $\mathcal{V}$ in Lemma 10.1.3, we can leverage all the tools we have developed to control moment generating functions and concentration inequalities in Chapter 4 to bound the χ^2-divergence and then apply Theorem 10.1.1.

Let us give one example of this approach, where we see the technique we use to prove the lemma arises frequently. Let $P_0 = \mathsf{N}(0, \sigma^2 I_d)$ be the standard normal distribution on $\mathbb{R}^d$, and for $\mathcal{V} = \{-1, 1\}^d$ and some $\delta \geq 0$ to be chosen, let $P_v = \mathsf{N}(\delta v, \sigma^2 I_d)$. Then we have the following lemma, which shows that while $D_{\rm kl}\left(P_v\|P_0\right) = \frac{d\delta^2}{2\sigma^2}$ for each individual P_v, the divergence for the average can be much smaller (even quadratically so in the ratio δ^2/σ^2).

Lemma 10.1.4. *Let P_0 and P_v be Gaussian distributions as above, and define the mixture $\overline{P} = \frac{1}{2^d}\sum_{v\in\{\pm1\}^d} P_v$. Then*

$$2\left\|P_0 - \overline{P}\right\|_{\rm TV}^2 \leq \log(1 + D_{\chi^2}\left(\overline{P}\|P_0\right)) \leq \frac{d\delta^4}{2\sigma^4}.$$

Proof The first inequality combines Pinsker's inequality (Proposition 2.2.8) with the bound $D_{\rm kl}\left(P\|Q\right) \leq \log(1 + D_{\chi^2}\left(P\|Q\right))$ in Proposition 2.2.9. Now we expand the χ^2-divergence, yielding

$$1 + D_{\chi^2}\left(\overline{P}\|P_0\right) = \mathbb{E}\left[\exp\left(-\frac{1}{2\sigma^2}\left\|Y - \delta V\right\|_2^2 - \frac{1}{2\sigma^2}\left\|Y - \delta V'\right\|_2^2 + \frac{1}{\sigma^2}\left\|Y\right\|_2^2\right)\right],$$

where the expectation is over $Y \sim \mathsf{N}(0, \sigma^2 I_n)$ and $V, V' \stackrel{\rm iid}{\sim} \mathsf{Uniform}(\mathcal{V})$. Taking the expectation over Y first, before averaging over the packing elements, allows more careful control. Indeed, expanding the squares and recognizing that $\|v\|_2^2 = d$ for each $v \in \{\pm 1\}^d$, we have

$$\begin{aligned}
1 + D_{\chi^2}\left(\overline{P}\|P_0\right) = \mathbb{E}\left[\exp\left(\frac{\delta}{\sigma^2}\langle Y, V + V'\rangle - \frac{n\delta^2}{\sigma^2}\right)\right] &= \mathbb{E}\left[\exp\left(\frac{\delta^2}{2\sigma^2}\left\|V + V'\right\|_2^2 - \frac{n\delta^2}{\sigma^2}\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\delta^2}{\sigma^2}\langle V, V'\rangle\right)\right] \\
&\leq \exp\left(\frac{d\delta^4}{2\sigma^4}\right),
\end{aligned}$$

where the final key inequality follows because an individual $U \sim \mathsf{Uniform}(\{\pm 1\})$ is 1-sub-Gaussian, and $\langle V, V'\rangle$ is thus d-sub-Gaussian. □

10.1.2 Estimating errors and the norm of a Gaussian vector

JCD Comment: It would probably be good to connect this to some other literatures and motivate things, e.g.,

1. Signal detection: is there something to discover?
2. Multiple testing: say we have d distinct p-values U_j. Then set $Z_j = \Phi^{-1}(U_j)$. Under the null that $U_j \sim \mathsf{Uniform}[0,1]$ these are i.i.d. $\mathsf{N}(0,1)$. Alternatives then deviate from this. Often interesting to consider other alternatives (sparse/dense/etc.)

JCD Comment: Clean this up now, because I moved Lemma 10.1.4 up.

Let us give one example to show how the mixture approach suggested by Lemma 10.1.3 works, along with showing that a more naive approach using the two point method of Chapter 8.3 fails to provide the correct bounds. After this we will further develop the techniques. We motivate the example by considering regression problems, then simplify it to a more stylized and easily workable form. Suppose we wish to estimate the best possible loss achievable in a regression problem,

$$\inf_\theta \mathbb{E}[(X^\top \theta - Y)^2].$$

For simplicity, assume that $X \sim \mathsf{N}(0, I_d)$, and that "base" distribution P_0 is simply that $Y \sim \mathsf{N}(0,1)$, while the alternatives are that $Y = X^\top \theta^\star + (1 - \|\theta^\star\|_2^2)\varepsilon$, where $\varepsilon \sim \mathsf{N}(0,1)$ and $\|\theta^\star\|_2^2 \le 1$. In either case we have $Y \sim \mathsf{N}(0,1)$ marginally, while

$$\inf_\theta \mathbb{E}_0[(X^\top \theta - Y)^2] = 1 \quad \text{and} \quad \inf_\theta \mathbb{E}_{\theta^\star}[(X^\top \theta - Y)^2] = 1 - \|\theta^\star\|_2^2,$$

so that estimating the final risk is equivalent to estimating the ℓ_2-norm $\|\theta^\star\|_2^2$.

To make the calculations more palatable, let us assume the simpler *Gaussian sequence model*

$$Y = \theta^\star + \varepsilon, \quad \varepsilon \sim \mathsf{N}(0, \sigma^2 I_n) \tag{10.1.2}$$

where $\theta^\star \in \mathbb{R}^n$ satisfies $\|\theta^\star\|_2 \le r$ for some radius r, and we wish to estimate the statistic

$$T(P) := \|\theta^\star\|_2^2.$$

Note that $\mathbb{E}[\|Y\|_2^2] = \|\theta^\star\|_2^2 + n\sigma^2$, so that a natural estimator is the debiased quantity

$$T_n := \|Y\|_2^2 - n\sigma^2.$$

Using that $\mathbb{E}[\varepsilon_j^2] = 1$ and $\mathbb{E}[\varepsilon_j^4] = 3$, we then obtain

$$\mathbb{E}\left[\left|T_n - \|\theta^\star\|_2^2\right|^2\right] = \sum_{j=1}^n \mathrm{Var}\left((\theta_j^\star + \sigma\varepsilon_j)^2\right) = 2n\sigma^4 + \|\theta^\star\|_2^2\,\sigma^2 \le 2n\sigma^4 + r^2\sigma^2.$$

That is, the family $\mathcal{P}_{\sigma,r}$ defined as Gaussian sequence models (10.1.2) with variance σ^2 and $\|\theta^\star\|_2^2 \le r^2$ satisfies

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \le \sqrt{2n\sigma^4 + r^2\sigma^2} \le \sqrt{2n}\sigma^2 + r\sigma. \tag{10.1.3}$$

We first provide the more naive approach. Suppose that we were to use Le Cam's two-point method to achieve a lower bound in this case. The minimax risk from inequality (8.3.3) shows that

(for a numerical constant $c > 0$), if P_0 and P_1 are (respectively) $\mathsf{N}(\theta_0, \sigma^2 I_n)$ and $\mathsf{N}(\theta_1, \sigma^2 I_n)$, then for any choice of θ_0, θ_1 we have

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \ge \frac{1}{4}\left\{\left|\|\theta_0\|_2^2 - \|\theta_1\|_2^2\right| \cdot [1 - \|P_0 - P_1\|_{\rm TV}]\right\}. \tag{10.1.4}$$

Recalling Pinsker's inequality (Proposition 2.2.8), we have

$$1 - \|P_0 - P_1\|_{\rm TV} \ge 1 - \frac{1}{\sqrt{2}}\sqrt{D_{\rm kl}\left(P_0 \| P_1\right)} = 1 - \frac{1}{2}\frac{\|\theta_0 - \theta_1\|_2}{\sigma}.$$

So whenever $\|\theta_0 - \theta_1\|_2 \le \sigma$, we have

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \ge \frac{1}{8}\left|\|\theta_0\|_2^2 - \|\theta_1\|_2^2\right|.$$

Take any θ_0 such that $\|\theta_0\|_2 = r$ and $\theta_1 = (1-t)\theta_0$, then choose the largest $t \in [0,1]$ such that $\|\theta_0 - \theta_1\|_2 = tr \le \sigma$. The choice $t = \min\{1, \frac{\sigma}{r}\}$ then gives that

$$\|\theta_0\|_2^2 - \|\theta_1\|_2^2 = r^2(1 - (1-t)^2) = r^2(2t - t^2) = 2\min\left\{r^2, r\sigma\right\} - \min\left\{r^2, \sigma^2\right\} \ge \min\{r^2, \sigma r\}.$$

In particular, this application of the two-point approach yields

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \ge \frac{1}{4}\min\left\{r^2, \sigma r\right\}. \tag{10.1.5}$$

(A careful inspection of the argument, potentially replacing the application of Pinsker with KL with a Hellinger distance bound, as in Proposition 2.2.8 shows that this is, essentially, the "best possible" bound achievable by the two-point approach.) While this bound *does* capture the second term in the upper bound (10.1.3) whenever $\sigma r \le r^2$, that is, $r \ge \sigma$, we require more sophisticated techniques to address the scaling with dimension n in the problem.

We therefore turn to using the mixture approach. Let $P_0 = \mathsf{N}(0, \sigma^2 I_n)$, and for $\mathcal{V} = \{\pm 1\}^n$ define $P_v = \mathsf{N}(\delta v, \sigma^2 I_n)$. It is immediate that $T(P_0) = 0$ while $T(P_v) = \delta^2 n$, so we have separation in the values of the statistic. In this case, we apply Theorem 10.1.1 and to obtain

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \ge \frac{\delta^2 n}{2}\left\{1 - \sqrt{\frac{1}{2}\log(1 + D_{\chi^2}\left(\overline{P} \| P_0\right))}\right\}$$

for $\overline{P} = \frac{1}{2^n}\sum_{v \in \mathcal{V}} P_v$. Substituting the result of Lemma 10.1.4 into the minimax lower bound, we obtain

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \ge \frac{\delta^2 n}{2}\left(1 - \sqrt{\frac{n\delta^4}{4\sigma^4}}\right).$$

We choose δ so that the (implied) probability of error in the hypothesis test from which our reduction follows is at least $\frac{1}{2}$, for which it evidently suffices to take $\delta = \frac{\sigma}{n^{1/4}}$. Putting all the pieces together, we achieve the minimax lower bound

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \ge \frac{\delta^2 n}{4} = \frac{\sigma^2\sqrt{n}}{4}. \tag{10.1.6}$$

Comparing the result from the upper bound (10.1.3), we see that at least in the regime that the radius r scales at most as $\sigma\sqrt{n}$, the mixture Le Cam method allows us to characterize the minimax risk of estimation of $\|\theta\|_2^2$ in a Gaussian sequence model.

By combining the result (10.1.3) with the more naive two-point lower bound (10.1.5), which is valid in "large radius" regimes, we have actually characterized the minimax risk.

Corollary 10.1.5. *Let $\mathcal{P}_{\sigma,r}$ be the Gaussian sequence model family $\{\mathsf{N}(\theta, \sigma^2 I_n) \mid \|\theta\|_2 \le r\}$, and $T(\theta) = \|\theta\|_2^2$. Then there is a numerical constant $c > 0$ such that the minimax absolute error satisfies*

$$c\left(\sigma^2\sqrt{n} + r\sigma\right) \le \mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \le \sqrt{2n\sigma^4 + r^2\sigma^2}.$$

Proof The only thing to recognize is that $r\sigma \ge \sigma^2\sqrt{n}$ whenever $r \ge \sigma\sqrt{n}$, in which case $\min\{r^2, \sigma r\} = \sigma r$ in the bound (10.1.5). □

10.2 Minimax hypothesis testing

In the general hypothesis testing problem, we have a family of potential distributions $\mathcal{P}$, and we are given a sample $X \sim P$ for some $P \in \mathcal{P}$. Then we wish to distinguish between two disjoint hypotheses H_0 and H_1:

$$\begin{aligned} H_0 : &\quad P \in \mathcal{P}_0 \\ H_1 : &\quad P \in \mathcal{P}_1, \end{aligned} \tag{10.2.1}$$

where the collections $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_1 \subset \mathcal{P}$ are disjoint. Then for a given test statistic $\Psi : \mathcal{X} \to \{0, 1\}$, we define the *risk* of the test to be

$$R(\Psi \mid \mathcal{P}_0, \mathcal{P}_1) := \sup_{P \in \mathcal{P}_0} P(\Psi \neq 0) + \sup_{P \in \mathcal{P}_1} P(\Psi \neq 1),$$

that is, the sum of the worst-case probabilities that the test is correct. (We also use the notation $R(\Psi \mid H_0, H_1)$ to denote the same quantity.) In the scenarios we consider, we will assume a metric ρ on the family of distributions $\mathcal{P}$, and instead of the general hypothesis test (10.2.1), we will consider testing whether $P \in \mathcal{P}_0$ or $\rho(P, P_0) \ge \epsilon$ for all $P_0 \in \mathcal{P}$, giving the variant

$$\begin{aligned} H_0 : &\quad P \in \mathcal{P}_0 \\ H_1 : &\quad P \in \mathcal{P}_1(\epsilon) := \{P \in \mathcal{P} \text{ s.t. } \rho(P, P_0) \ge \epsilon \text{ all } P_0 \in \mathcal{P}_0\} \end{aligned} \tag{10.2.2}$$

In this case, we can define the *risk at distance* ϵ for a sample of size n by

$$R_n(\Psi, \epsilon) := \sup_{P \in \mathcal{P}_0} P(\Psi(X_1^n) \neq 0) + \sup_{P \in \mathcal{P}_1(\epsilon)} P(\Psi(X_1^n) \neq 1), \tag{10.2.3}$$

leaving $\mathcal{P}_0$ and $\mathcal{P}$ implicit in the definition, and where we let $X_1^n \stackrel{\text{iid}}{\sim} P$. From this, we can define the minimax test risk

$$\inf_{\Psi} R_n(\Psi, \epsilon).$$

We then ask for the particular thresholds ϵ at which the minimax test risk becomes small or large. Thus, while the coming definition allows some ambiguity, we say that a sequence ϵ_n is a *minimax threshold* or *critical testing radius* for the testing problem (10.2.2) if there exist numerical constants $0 < c \le C < \infty$ such that

$$\inf_{\Psi} R_n(\Psi, C\epsilon_n) \le \frac{1}{3} \quad \text{and} \quad \inf_{\Psi} R_n(\Psi, c\epsilon_n) \ge \frac{2}{3}. \tag{10.2.4}$$

Here, the constants $\frac{1}{3}$ and $\frac{2}{3}$ are unimportant, the point being that for separation at most $c\epsilon_n$, *no* hypothesis test can test whether the distribution P satisfies $P \in \mathcal{P}_0$ or $\inf_{P_0 \in \mathcal{P}_0} \rho(P, P_0) \ge c\epsilon_n$

with reasonable accuracy. But it *is* possible to test whether $P \in \mathcal{P}_0$ or $\inf_{P_0 \in \mathcal{P}_0} \rho(P, P_0) \geq C\epsilon_n$ with reasonable accuracy. Moreover, we can make the probability of error exponentially small by increasing the sample size by a constant factor, as Exercise 10.2 explores.

Conveniently, the minimax test risk has a precise divergence-based form, to which we can apply the techniques comparing different divergences we have developed. In particular, we have the following analogue of Le Cam's convex hull lower bound in Theorem 10.1.1, which provides the same fundamental quantity (the variation distance between convex hulls of $\mathcal{P}_0$ and $\mathcal{P}_1$) for lower bounds, except that it applies for testing.

Proposition 10.2.1 (Convex hull lower bounds in testing)**.** *For any classes $\mathcal{P}_0$ and $\mathcal{P}_1$, the minimax test risk satisfies*

$$\inf_{\Psi} R(\Psi \mid \mathcal{P}_0, \mathcal{P}_1) \geq 1 - \sup \left\{ \left\| \overline{P}_0 - \overline{P}_1 \right\|_{\rm TV} \mid \overline{P}_0 \in {\rm Conv}(\mathcal{P}_0), \overline{P}_1 \in {\rm Conv}(\mathcal{P}_1) \right\}.$$

Proof Let $\overline{P}_0 \in {\rm Conv}(\mathcal{P}_0)$ and $\overline{P}_1 \in {\rm Conv}(\mathcal{P}_1)$. Then for any test Ψ,

$$R(\Psi \mid \mathcal{P}_0, \mathcal{P}_1) \geq \overline{P}_0(\Psi \neq 0) + \overline{P}_1(\Psi \neq 1)$$

because suprema are always at least as large as averages. Now note that the set $A = \{x \mid \Psi(x) = 0\}$ satisfies

$$\overline{P}_0(\Psi \neq 0) + \overline{P}_1(\Psi \neq 1) = \overline{P}_0(A^c) + \overline{P}_1(A) = 1 - (\overline{P}_0(A) - \overline{P}_1(A)),$$

and take an infimum over regions A. □

In fact, equality typically holds in Proposition 10.2.1, but this requires the application of (infinite dimensional) convex duality, which is beyond our scope here.

10.2.1 Detecting a difference in populations

With the generic worst-case hypothesis testing setup in place, we can give a general recipe for developing tests. We specialize this recipe in the next few sections to different problems, including signal detection in a Gaussian model, two-sample tests in multinomials, and goodness of fit testing. The basic approach in all of these problems is frequently the following: to demonstrate achievability and testability, we develop an estimator T_n of the distance $\rho(P_0, P_1)$, or some other function of the distance, where T_n has reasonable properties. We then develop a test Ψ by thresholding this estimator. For the converse results that no test can distinguish the families $\mathcal{P}_0$ and $\mathcal{P}_1$ at a particular distance, we use the mixture χ^2 approaches we have outlined.

Let us give the general recipe first. Suppose that we have a statistic T designed to separate the classes $\mathcal{P}_0$ and $\mathcal{P}_1$. Such a statistic should assign large values for samples $X \sim P_1$ for $P_1 \in \mathcal{P}_1$ and small values for samples $X \sim P_0$. A more quantitative version of this, where the separation $\mathbb{E}_1[T] - \mathbb{E}_0[T]$ is commensurate with the variance of T, is sufficient to test between $\mathcal{P}_0$ and $\mathcal{P}_1$ with high accuracy. To that end, we say that the statistic T *robustly C-separates* $\mathcal{P}_0$ and $\mathcal{P}_1$ if

$$\mathbb{E}_{P_1}[T] - \sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[T] \geq C \left(\sup_{P_0 \in \mathcal{P}_0} \sqrt{{\rm Var}_{P_0}(T)} + \sqrt{{\rm Var}_{P_1}(T)} \right). \tag{10.2.5}$$

for each $P_1 \in \mathcal{P}_1$. Typically, we choose statistics T so that $\mathbb{E}_{P_0}[T] = 0$ for each P_0 in the null $\mathcal{P}_0$ (though this is not always possible). The next proposition shows how to define a test that leverages this to achieve small worst-case test error.

Proposition 10.2.2. *Let the statistic $T : \mathcal{X} \to \mathbb{R}$ robustly C-separate P_0 from $\mathcal{P}_1$. Then for the threshold $\tau = \sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[T] + \sup_{P_0 \in \mathcal{P}_0} \sqrt{\mathrm{Var}_{P_0}(T)}$, the test*

$$\Psi(X) := \mathbf{1}\{T \geq \tau\}$$

satisfies

$$R(\Psi \mid \{P_0\}, \mathcal{P}_1) \leq \frac{2}{C^2}.$$

Proof Without loss of generality we assume $\sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[T] = 0$, as the test is invariant to shifts, so that $\tau = \sup_{P_0 \in \mathcal{P}_0} \sqrt{\mathrm{Var}_{P_0}(T)}$. We can also assume that $C \geq 1$, as otherwise the proposition is vacuous. We control the test error in each case. Under any null P_0, we have

$$P_0(\Psi \neq 0) = P_0(T \geq \tau) \leq \frac{\mathrm{Var}_0(T)}{C^2 \tau^2} = \frac{1}{C^2}.$$

For the alternatives under $P_1 \in \mathcal{P}_1$, we have

$$P_1(\Psi \neq 1) = P_1(T \leq \tau) = P_1(T - \mathbb{E}_1[T] \leq \tau - \mathbb{E}_1[T]) \leq \frac{\mathrm{Var}_1(T)}{[\mathbb{E}_1[T] - \tau]_+^2}.$$

But of course,

$$\mathbb{E}_1[T] - \tau = \mathbb{E}_1[T] - \sup_{P_0} \mathbb{E}_{P_0}[T] - \sup_{P_0} \sqrt{\mathrm{Var}_{P_0}(T)} \geq C\sqrt{\mathrm{Var}_1(T)} + (C-1) \sup_{P_0} \sqrt{\mathrm{Var}_{P_0}(T)}$$

by the robust C-separation. As we have assumed w.l.o.g. that $C \geq 1$, this yields

$$P_1(\Psi \neq 1) \leq \frac{\mathrm{Var}_1(T)}{C^2 \mathrm{Var}_1(T)} = \frac{1}{C^2}$$

as desired. □

10.2.2 Signal detection and testing a Gaussian mean

A common problem in statistics, communication, and information theory is the *signal detection* problem, where we observe $X \sim P$ from an unknown distribution P, and wish to detect if there is some "signal" present in P. To study such a problem, we typically formulate a null model—indicating absence of signal—and a set of alternatives for which there is *some* signal, though we only care to test its existence. The existence of a signal can then justify further investigation or data collection to actually estimate the signal.

Let us give a few variants of this problem, for which a substantial literature exists.

Example 10.2.3 (Dense Gaussian signal detection)**:** We consider testing the null H_0 and alternative H_1 given by

$$\begin{aligned} H_0 : &\quad P = P_0 = \mathsf{N}(0, I_d) \\ H_1 : &\quad P \in \mathcal{P}_1(r) := \{\mathsf{N}(\theta, I_d) \mid \|\theta\|_2 \geq r\}. \end{aligned} \tag{10.2.6}$$

That is, we are interested in whether $X \sim P$ has a mean θ separated by at least r from the all-zeros vector. The problem is to find the critical radius r at which testing between $\mathcal{P}_0 = \{P_0\}$ and $\mathcal{P}_1$ becomes feasible (or infeasible). ◇

Example 10.2.4 (A global null in multiple hypothesis testing)**:** Consider the problem of testing d distinct null hypotheses $H_{0,j}$, $j = 1, \ldots, d$, where for each we have a p-value Y_j and reject $H_{0,j}$ if $Y_{0,j} \le \tau$ for a threshold τ. (Recall that a p value is a random variable Y that is *sub-uniform*, meaning that $P(Y \le u) \le P(U \le u)$ for $U \sim \mathsf{Uniform}[0,1]$, so we are less likely to reject at threshold τ than a uniform would be.) If we assume the Y_j are exact p-values, that is, $P(Y_j \le u) = u$ for $u \in [0,1]$, then testing the global independent null

$$H_0 := \bigcap_{j=1}^{d} H_{0,j} = \text{each } Y_j \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0,1]$$

is equivalent to Gaussian signal detection. Indeed, let $Z_j = \Phi^{-1}(Y_j)$, where Φ denotes the standard Gaussian cumulative distribution. Then under the global null H_0, we have

$$Z \sim \mathsf{N}(0, I_d).$$

The question of which alternative class $\mathcal{P}_1$ to consider is then frequently a matter of applications. For example, we might be curious about alternatives for which a few nulls $H_{0,j}$ are false, that is, *sparse* alternatives. Example 10.2.3 corresponds to something like dense alternatives. $\diamond$

With these as motivation, let us consider Example 10.2.3 more carefully, in effort to find the critical radius r at which minimax testing becomes feasible (or infeasible). While our standard techniques for estimation tell us that the minimax rate for estimating θ in a normal location family $\mathcal{P} = \{\mathsf{N}(\theta, \sigma^2 I_d)\}_{\theta \in \mathbb{R}^d}$ (say, in mean squared error) necessarily scale as

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2) = \frac{d\sigma^2}{n},$$

we can *test* whether the mean of a Gaussian is zero at a smaller dimensionality—effectively, while $\mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \to 0$ as $n \to \infty$ if and only if $d/n \to 0$, in the testing case, we can save a dimension-dependent factor $\sqrt{d}$. In particular, the next two examples—one addressing achievability and one the fundamental limit—show that in the dense Gaussian signal detection problem of Example 10.2.3, the critical test radius (10.2.4) at which testing is feasible or infeasible scales as

$$r_n := \frac{d^{1/4}}{\sqrt{n}}.$$

We can achieve (asymptotically) accurate testing in the dense signal detection problem (10.2.6) if and only if $\sqrt{d}/n \to 0$ as $n \to \infty$.

We first demonstrate achievability in Example 10.2.3, leveraging Proposition 10.2.2.

Example 10.2.5 (Achievability in Gaussian mean testing)**:** We wish to test the alternatives (10.2.6). We use the approach of Proposition 10.2.2: find an estimator of $\|\theta\|_2^2$, and then threshold it for our test. The discussion preceding Corollary 10.1.5 (specifically equation (10.1.3)) shows that given a sample of size n, the estimator $T_n = \|\overline{X}_n\|_2^2 - d/n$ is unbiased for $\|\theta\|_2^2$ and satisfies

$$\mathbb{E}_\theta\left[(T_n - \|\theta\|_2^2)^2\right] = \mathrm{Var}_\theta(T_n) \le \frac{2d}{n^2} + \frac{\|\theta\|_2^2}{n}. \tag{10.2.7}$$

Note that $\mathbb{E}_0[T_n] = 0$, and so because

$$\mathbb{E}_\theta[T_n] - \mathbb{E}_0[T_n] = \|\theta\|_2^2,$$

the statistic T_n robustly 2-separates P_0 from $\mathcal{P}_1(r)$ (recall definition (10.2.5)) whenever

$$\frac{1}{2}\|\theta\|_2^2 \geq \left(\frac{\sqrt{2d}}{n} + \sqrt{\frac{2d}{n^2} + \frac{1}{n}\|\theta\|_2^2}\right)$$

for all θ with $\|\theta\|_2 \geq r$. Immediately we see that if we take radius $r^2 = C\frac{\sqrt{d}}{n}$ for some $C > 0$, then this separation occurs if $C\sqrt{d} \geq 2(\sqrt{2d} + \sqrt{2d + C\sqrt{d}})$, which of course happens for large constant C. Applying Proposition 10.2.2, we thus see that the test $\Psi(X_1^n) = \mathbf{1}\left\{T_n \geq \sqrt{2d/n^2}\right\}$ satisfies

$$R_n(\Psi, Cr_n) \leq \frac{1}{3} \quad \text{for} \quad r_n = \frac{d^{1/4}}{\sqrt{n}},$$

which gives the achievability required for the critical test radius (10.2.4). $\diamond$

Example 10.2.5 shows that at the critical radius $r_n = \frac{d^{1/4}}{\sqrt{n}}$, it is possible (in a worst-case sense) to test between the null $H_0 : \mathsf{N}(0, I_d)$ and alternatives $H_1 : \mathsf{N}(\theta, I_d)$ for $\|\theta\|_2 \geq Cr_n$, where C is a numerical constant. We can also provide the converse.

Example 10.2.6 (Lower bounds in Gaussian mean testing)**:** Let $\mathcal{P}_1(r) = \{\mathsf{N}(\theta, I_d) \mid \|\theta\|_2 \geq r\}$ be a collection of Gaussians with means r away from the origin in ℓ_2-norm. We seek the critical radius r below which it is impossible to distinguish between $P_0 = \mathsf{N}(0, I_d)$ and $P_1 \in \mathcal{P}_1(r)$ given an i.i.d. sample X_1^n. Lemma 10.1.4 and Proposition 10.2.1 combine (set $\sigma^2 = \frac{1}{n}$ and $\delta^2 = \frac{r^2}{d}$ in Lemma 10.1.4) to give

$$\inf_\Psi R_n(\Psi \mid \mathcal{P}_0, \mathcal{P}_1(r)) \geq 1 - \frac{1}{\sqrt{2}}\left[\exp\left(\frac{n^2r^4}{2d}\right) - 1\right].$$

In particular, the threshold $r^2 = \sqrt{d}/n$ means that there is necessarily constant test error probability $R_n \geq 1 - \frac{1}{\sqrt{2}}(\sqrt{e} - 1) > .54$. Combining the estimation guarantee with this lower bound shows that the critical radius (10.2.4) for testing $H_0 : \mathsf{N}(0, I_d)$ against the family of alternatives $H_1 : \mathsf{N}(\theta, I_d)$ with $\|\theta\|_2^2 \geq r^2$ is precisely $r^2 = \sqrt{d}/n$. $\diamond$

10.2.3 Goodness of fit and two-sample tests for multinomials

The basic question in goodness of fit testing—called property testing in the theoretical computer science literature—is the following. Given a sample $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P$, we wish to test whether $P = P_0$ for a prescribed base distribution P_0 or P is far from P_0. The related two-sample testing problem generalizes this, where we assume samples $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P$ and $Y_1, \ldots, Y_m \stackrel{\text{iid}}{\sim} Q$, and wish to test whether $P = Q$. Each of these falls into the class of hypothesis tests (10.2.2), where the choice of the metric ρ can change the character of upper and lower bounds somewhat dramatically. General methods for developing goodness of fit and two-sample tests typically take the broad approach in Section 10.2.1, defining a statistic T that separates the distribution P_0 (or the joint that X_i and Y_j have the same distribution) from the alternatives about which we are curious, then thresholding that statistic.

It turns out that even in what might appear to be a particularly simple case—that of multinomial distributions, where we identify the distribution P with a probability mass function (p.m.f.) $p \in \Delta_d$—a surprising amount of complexity arises. We thus work through two examples on testing distance between discrete distributions by considering two metrics on the probability mass functions: the ℓ_2-metric and the total variation distance (or ℓ_1 metric). Then $\rho(p, q) = \|p - q\|$ for $\|\cdot\| = \|\cdot\|_2$ or $\|\cdot\| = \|\cdot\|_1$. In the uniformity testing case, we let $p_0 = \frac{1}{d}\mathbf{1}$ be the uniform distribution on $[d]$, and we seek the critical threshold ϵ at which testing

$$\|p - p_0\| = 0 \quad \text{versus} \quad \|p - p_0\| \geq \epsilon$$

from n i.i.d. observations $X_i \stackrel{\text{iid}}{\sim} p$ becomes feasible or infeasible.

It is simpler (for analyzing procedures) to consider a slight variant of this problem, which uses the *Poissonization* trick. To motivate the idea, identify the observations X_i with the basis vectors (so that observing item $j \in \{1, \ldots, d\}$ corresponds to $X_i = e_j$). Then that the sample mean $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ is unbiased, but its coordinates exhibit dependence in that $\langle \mathbf{1}, \widehat{p} \rangle = 1$—an annoyance for analyses. Thus, we consider an alternative approach, where we assume a two-stage sampling procedure: we first drawn $N \sim \mathsf{Poi}(n)$, and then conditional on $N = m$, draw $X_i \stackrel{\text{iid}}{\sim} p$, $i = 1, \ldots, m$. As $\mathbb{E}[N] = n$ and N concentrates around its mean, this is nearly equivalent to simply observing $X_i \stackrel{\text{iid}}{\sim} p$ for $i = 1, \ldots, n$, and a standard probabilistic calculation shows that the distribution of $\{X_i\}_{i=1}^N$ conditional on $N = m$ is identical to the distribution of $X_i \stackrel{\text{iid}}{\sim} p$, $i = 1, \ldots, m$.

Even more, the minimax risk for estimation in this Poissonized sampling scheme is similar to that for estimation in the original multinomial setting. Indeed, suppose that we wish to estimate an abstract statistic $T(p)$ of $p \in \Delta_d$, and assume for simplicity that $T(p) \in [-r, r]$ for some fixed r. Define the minimax and Poissonized minimax risks

$$\mathfrak{M}_n := \inf_{T_n} \sup_{p \in \Delta_d} \mathbb{E}_p \left[(T_n(X_1^n) - T(p))^2 \right]$$

and

$$\mathfrak{M}_{\mathsf{Poi}(n)} := \inf_{\{T_m\}} \sup_{p \in \Delta_d} \mathbb{E}_p \left[(T_N(X_1^N) - T(p))^2 \right],$$

where the latter expectation is taken over the sample size $N \sim \mathsf{Poi}(n)$, and $\{T_m\}$ denotes a sequence of estimators (defined for all sample sizes m). We have the following proposition, which shows that if we can provide procedures that work in the poissonized (independent sampling) setting, then the standard multinomial sampling setting is similarly easy (or challenging).

Proposition 10.2.7. *There exist numerical constants $0 < c, C < \infty$ such that*

$$\mathfrak{M}_{\mathsf{Poi}(2n)} - Cr^2 \exp(-cn) \leq \mathfrak{M}_n \leq 2 \cdot \mathfrak{M}_{\mathsf{Poi}(n/2)}. \tag{10.2.8}$$

For a proof, see Exercises 10.3 and 10.4.

Let us leverage these ideas to construct an estimator for the ℓ_2-distance between two multinomial distributions. In this case, suppose we have $X_i \stackrel{\text{iid}}{\sim} p$ and $Y_i \stackrel{\text{iid}}{\sim} q$, where $p, q \in \Delta_d$, both for $i = 1, \ldots, N$ and $N \sim \mathsf{Poi}(n)$, and we define

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^N X_i, \quad \widehat{q} = \frac{1}{n}\sum_{i=1}^N Y_i. \tag{10.2.9}$$

This is equivalent to sampling $n\widehat{p}_j \stackrel{\text{ind}}{\sim} \mathsf{Poi}(np_j)$ and $n\widehat{q}_j \stackrel{\text{ind}}{\sim} \mathsf{Poi}(nq_j)$, $j = 1, \ldots, d$, and so we use the quantities (10.2.9) to define an estimator we can threshold using Proposition 10.2.1. We work through this in the next (somewhat complicated) example.

Example 10.2.8 (Estimating the ℓ_2-distance between multinomials)**:** For the estimators (10.2.9), define the quantity

$$Z_j := (n\widehat{p}_j - n\widehat{q}_j)^2 - n\widehat{p}_j - n\widehat{q}_j.$$

Recalling that if $W \sim \mathsf{Poi}(\lambda)$ then $\mathbb{E}[W] = \mathrm{Var}(W) = \lambda$, we have $\mathbb{E}[n\widehat{p}_j] = p_j$ and $\mathrm{Var}(n\widehat{p}_j) = np_j$, so

$$\begin{aligned}\mathbb{E}[Z_j] &= \mathbb{E}[(n\widehat{p}_j)^2] + \mathbb{E}[(n\widehat{q}_j)^2] - 2n^2 p_j q_j - np_j - nq_j \\ &= \mathrm{Var}(n\widehat{p}_j) + \mathrm{Var}(n\widehat{q}_j) + (np_j)^2 + (nq_j)^2 - 2n^2 p_j q_j - np_j - nq_j = n^2 \left\|p - q\right\|_2^2.\end{aligned}$$

In particular, the statistic

$$T_n := \frac{1}{n^2}\langle \mathbf{1}, Z\rangle$$

satisfies $\mathbb{E}[T_n] = \|p - q\|_2^2$.
To be able to test whether p and q are identical using Proposition 10.2.2, we must compute the variance of $\langle \mathbf{1}, Z\rangle$, which—conveniently, by the independence our Poisson sampling gives—is $\sum_{j=1}^d \mathrm{Var}(Z_j)$. Leveraging that for a Poisson $W \sim \mathsf{Poi}(\lambda)$ we have (by tedious calculation) that

$$\mathbb{E}[W] = \lambda,\ \mathbb{E}[W^2] = \lambda(1+\lambda),\ \mathbb{E}[W^3] = \lambda + 3\lambda^2 + \lambda^3,\ \mathbb{E}[W^4] = \lambda + 7\lambda^2 + 6\lambda^3 + \lambda^4,$$

we obtain (see Exercise 10.7)

$$\mathrm{Var}(Z_j) = 4n^3(p_j - q_j)^2(p_j + q_j) + 2(p_j + q_j)^2 n^2 \tag{10.2.10}$$

and

$$\mathrm{Var}(\langle \mathbf{1}, Z\rangle) \le 4n^3 \left\|p - q\right\|_4^2 \left\|p + q\right\|_2 + 2n^2 \left\|p + q\right\|_2^2.$$

Under the (non-point) null $H_0 : p = q$, $\mathrm{Var}(\langle \mathbf{1}, Z\rangle) = 2n^2 \left\|p + q\right\|_2^2 \le 8n^2$, as $\sup_{p,q} \|p + q\|_2 = 2$. Proposition 10.2.2 thus shows that if

$$\|p - q\|_2^2 \ge C\left(\sqrt{\frac{8}{n^2}} + \sqrt{\frac{16\left\|p - q\right\|_4^2}{n} + \frac{8}{n^2}}\right), \tag{10.2.11}$$

then the test

$$\Psi := \mathbf{1}\left\{T_n \ge \sqrt{8}/n\right\}$$

satisfies $P_0(\Psi \ne 0) + P_1(\Psi \ne 1) \le \frac{2}{C^2}$, where P_0 is any distribution with $p = q$ and P_1 is any distribution with $\|p - q\|_2$ satisfying the separation (10.2.11). As $\|p - q\|_2 \ge \|p - q\|_4$, inequality (10.2.11) a necessary and sufficient condition for inequality (10.2.11) to hold is that $\|p - q\|_2 \gtrsim 1/\sqrt{n}$. $\diamond$

Summarizing, we see that if we wish to test whether two multinomials are identical or separated in ℓ_2, the critical threshold for the hypothesis test

$$\begin{aligned} H_0 :&\quad p = q \\ H_1 :&\quad \|p - q\|_2 \ge \delta \end{aligned} \tag{10.2.12}$$

satisfies $\delta \le \frac{1}{\sqrt{n}}$: we can test between H_0 and H_1 at separations that are essentially "independent" of the dimension or number of categories d. This is in fact sharp, as a relatively straightforward argument with Le Cam's two-point lemma demonstrates (see Exercise 10.9). However, if we change the norm $\|\cdot\|_2$ into the ℓ_1-norm $\|\cdot\|_1$, the story changes significantly.

Let us change the hypothesis test (10.2.12) to simpler looking—in that we only test goodness of fit—ℓ_1-based variant. Identifying distributions P on $\{1, \ldots, d\}$ with their p.m.f.s $p \in \Delta_d$, let P_0 be the uniform distribution on $\{1, \ldots, d\}$, with p.m.f. $p_0 = \frac{1}{d}\mathbf{1}$. Then we consider the testing problem

$$\begin{aligned} H_0 : & \quad p = p_0 \\ H_1 : & \quad \|p - p_0\|_1 \ge \delta, \end{aligned} \tag{10.2.13}$$

which tests the ℓ_1-distance to uniformity. In this case, developing a test that distinguishes these hypotheses at the optimal rate is quite sophisticated, though we outline an approach to it in the exercises. To develop the correct order of lower bound—that is, a threshold δ for which no test can reliably distinguish H_0 from H_1—is possible via the mixture of χ^2-distributions approach we have developed in Lemma 10.1.3.

JCD Comment: Should I just do these as lemmas / propositions rather than examples? They're a bit involved for examples!

Proposition 10.2.9 (A lower bound for testing ℓ_1-separated multinomials)**.** *In the testing problem* (10.2.13)*,*

$$\inf_{\Psi} R_n(\Psi \mid H_0, H_1) \ge 1 - \frac{1}{\sqrt{2}}$$

whenever $\delta \le \frac{d^{1/4}}{\sqrt{n}}$.

Proof We construct a particular packing of the probability simplex $\Delta_d \in \mathbb{R}_+^d$ that guarantees that the divergence between elements of H_0 and H_1 in the test (10.2.13) is small. For simplicity, we assume d is even, as it changes nothing. For the base distribution P_0 take p.m.f. $p_0 = \frac{1}{d}\mathbf{1}$ as required by the problem (10.2.13). To construct the alternatives, let $\mathcal{V} \subset \{\pm 1\}^d$ be the collection of $2^{d/2}$ vectors of the form $v = (v', -v')$, where $v' \in \{\pm 1\}^{d/2}$, so that $\langle \mathbf{1}, v\rangle = 0$ for each $v \in \mathcal{V}$. Then for $\delta \ge 0$ to be chosen, define the p.m.f.s $p_v = \frac{1+\delta v}{d}$. Identify samples $X \in \{e_1, \ldots, e_d\}$. Then for any $x \in \{e_j\}$, we have $P_v(X = x) = \frac{1}{d}(1 + \delta\langle v, x\rangle)$, and so for any pair v, v' we have

$$\frac{P_v(X = x)P_{v'}(X = x)}{P_0(X = x)^2} = (1 + \delta\langle v, x\rangle)(1 + \delta\langle v', x\rangle).$$

From this key equality, we see that if $V, V' \stackrel{\text{iid}}{\sim} \mathsf{Uniform}(\mathcal{V})$, then for $\overline{P} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v$ we have

$$\begin{aligned} 1 + D_{\chi^2}\left(\overline{P} \| P_0\right) &= \mathbb{E}_0\left[\prod_{i=1}^n (1 + \delta\langle V, X_i\rangle)(1 + \delta\langle V', X_i\rangle)\right] \\ &= \mathbb{E}\left[\mathbb{E}_0[(1 + \delta\langle V, X\rangle)(1 + \delta\langle V', X\rangle) \mid V, V']^n\right] \\ &= \mathbb{E}\left[\left(1 + \frac{\delta^2}{d}\langle V, V'\rangle\right)^n\right], \end{aligned}$$

where the final equality follows because $\mathbb{E}_0[\langle v, X\rangle] = \frac{1}{d}\langle v, \mathbf{1}\rangle = 0$ for each $v \in \mathcal{V}$. Now we use that $1 + t \le e^t$ for all t to obtain

$$1 + D_{\chi^2}\left(\overline{P} \| P_0\right) \le \mathbb{E}\left[\exp\left(\frac{n\delta^2}{d}\langle V, V'\rangle\right)\right] = \mathbb{E}\left[\exp\left(\frac{2n\delta^2}{d}\sum_{j=1}^{d/2} U_j\right)\right]$$

for $U_j \stackrel{\text{iid}}{\sim} \mathsf{Uniform}\{\pm 1\}$. But of course these U_j are 1-sub-Gaussian, so

$$1 + D_{\chi^2}\left(\overline{P} \| P_0\right) \le \exp\left(\frac{n^2\delta^4}{d}\right).$$

Now use Pinsker's inequalities (Propositions 2.2.8 and 2.2.9), which gives $2\left\|P_0 - \overline{P}\right\|_{\text{TV}}^2 \le \frac{n^2}{\delta^4} d$. Choose $\delta^4 = \frac{d}{n^2}$. □

10.3 Geometrizing rates of convergence

JCD Comment: Outline for this section:

1. Introduce modulus of continuity (w.r.t. Hellinger), draw a picture suggesting why it should be hard or easy
2. Example with Fisher information-type quantity
3. Show that for *testing*, the rate at which we can test really is this modulus whenever we have linear functions and convex classes, because of Le Cam's result on Hellinger affinities.

JCD Comment: Write this section

10.4 Best possible lower bounds and super-efficiency

JCD Comment: Write this section. Get in super-efficiency stuff.

10.5 Bibliography

JCD Comment: We stole the mixture idea from David Pollard I believe.

Outline

I. Motivation: function values, testing certain quantities (e.g. is $\|P - Q\|_{\text{TV}} \ge \epsilon$ or not), entropy and other quantities, and allows superefficiency guarantees in an elegant way

II. Le Cam's methods

1. The general form with mixtures
2. The χ^2-type bounds, with mixtures to a point mass
3. Geometrizing rates of convergence
4. Examples: Fisher information in classical problems (especially for a one-dimensional quantity)
5. Example: testing distance to uniformity (failure from standard two-point bound)

6. More sophisticated examples:
   a. Smooth functionals (as in Birgé and Massart [30]), like differential entropy $\int h(x) \log h(x) dx$
   b. Higher-dimensional problems, which are hard

III. "Best possible" lower bounds, super-efficiency and constrained risk inequalities

1. Basic (two-point) constrained risk inequality (cf. [66])
2. Constrained risk inequality when P_1 is actually a mixture (easiest with a functional): means that any minimax bound around P_0 is quite strong
3. Potentially (?): Cai and Low [42] paper on minimax estimation for $\frac{1}{n} \|\theta\|_1$ when $y = \theta + \varepsilon$ in a Gaussian sequence model as an example and application of a constrained risk inequality. This is probably too challenging, though—can we find a case where polynomials actually allow us to do stuff?
   a. Hard because of all the polynomial approximation stuff... but maybe there is a simpler version that simply shows how approximation via polynomials allows lower bounds. Approach works for Gaussian stuff, as in Cai and Low [42] or the earlier paper "Effect of mean on variance function estimation in nonparametric regression" by Wang, Brown, Cai, Levine.
   b. Similar idea gives variation distance bounds for Poisson priors on parameters when seeking lower bounds on estimating entropy $H(X) = -\sum_x p_x \log p_x$ of discrete distributions with (unknown) support; see [174].

10.6 A useful divergence calculation

Now, let us suppose that we define the collection $\{P_v\}$ by tiltings of an underlying base distribution P_0, where each tilting is indexed by a function $g_v : \mathcal{X} \to [-1, \infty)$, and where

$$dP_v(x) = (1 + g_v(x)) dP_0(x),$$

while $\int g_v dP_0 = 0$, so that each P_v is a valid distribution. Let P_v^n be the distribution of n observations $X_i \stackrel{\text{iid}}{\sim} P_v$, and let $\overline{P^n} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v$.

Lemma 10.6.1. *Define the inner product $\langle f, g \rangle_P = \int f(x) g(x) dP(x)$ and let $V, V' \stackrel{\text{iid}}{\sim} \mathsf{Uniform}(\mathcal{V})$. Then*

$$D_{\chi^2}\left(\overline{P^n} \| P_0\right) + 1 \le \mathbb{E}[\exp(n \langle g_V, g_{V'} \rangle_{P_0})].$$

Proof The simple technical lemma 10.1.3 essentially gives us the result. We observe that

$$D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) + 1 = \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} \int \frac{dP_v^n dP_{v'}^n}{dP_0^n} = \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} \left(\int (1 + g_v(x))(1 + g_{v'}(x)) dP_0(x) \right)^n$$

because $P_v^n(x_1, \ldots, x_n) = \prod_{i=1}^n (1 + g_v(x_i)) dP_0(x_i)$, so that the integral decomposes into a product of integrals. Then expanding $(1 + g_v)(1 + g_{v'})$ and noting that each has zero mean under P_0 gives

$$D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) + 1 = \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} \left(1 + \mathbb{E}_0[g_v(X) g_{v'}(X)]\right)^n.$$

Lastly, we note that $(1+t) \le e^t$ for all t, and so

$$\frac{1}{|\mathcal{V}|^2} \sum_{v,v'} \left(1 + \mathbb{E}_0[g_v(X) g_{v'}(X)]\right)^n \le \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} \exp\left(n \mathbb{E}_0[g_v(X) g_{v'}(X)]\right),$$

which is of course equivalent to the result we desired. □

A specialization of Lemma 10.6.1 follows when we choose our functions g to correspond to a partition of $\mathcal{X}$-space. Here, we define the following.

Definition 10.1. *Let $k \in \mathbb{N}$ and the functions $\phi_j : \mathcal{X} \to [-b, b]$. Then the functions ϕ_j are an* admissible partition with variances σ_j^2 *of $\mathcal{X}$ with respect to a probability distribution P_0 if*

(i) The supports $E_j = \operatorname{supp} \phi_j$ of each of the functions are disjoint.

(ii) Each function has P_0 mean 0, i.e., $\mathbb{E}_{P_0}[\phi_j(X)] = 0$ for each j.

(iii) Function j has variance $\sigma_j^2 = \mathbb{E}_{P_0}[\phi_j^2(X)] = \int \phi_j^2(x) dP_0(x)$.

With such a partition, we can define the functions $g_v(x) = t\langle v, \phi(x)\rangle = t \sum_{j=1}^k v_j \phi_j(x)$ for $|t| \le 1/b$, and if we take $\mathcal{V} = \{-1, 1\}^k$, we obtain the following.

Lemma 10.6.2. *Let the functions $\{\phi_j\}_{j=1}^k$ be an admissible partition of $\mathcal{X}$ with variances σ_j^2. Fix $|t| \le \frac{1}{b}$, and let $dP_{tv} = (1 + t\langle v, \phi(x)\rangle) dP_0(x)$ and $\overline{P_t^n} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v^n$. Then*

$$D_{\chi^2}\left(\overline{P_t^n} \| P_0\right) \le \exp\left(\frac{n^2 t^4}{2} \sum_{j=1}^k \sigma_j^4\right) - 1,$$

and if $|t| \le \frac{1}{\sqrt{n}} \frac{1}{(\sum_{j=1}^k \sigma_j^4)^{1/4}}$, then

$$D_{\chi^2}\left(\overline{P_t^n} \| P_0\right) \le n^2 t^4 \sum_{j=1}^k \sigma_j^4.$$

Proof First, if $\phi(x) = [\phi_j(x)]_{j=1}^k$, then $\mathbb{E}_0[\phi(X)\phi(X)^T] = \operatorname{diag}(\sigma_j^2)$, that is, the diagonal matrix with σ_j^2 on its diagonal. By Lemma 10.6.1, we therefore have

$$D_{\chi^2}\left(\overline{P_t^n} \| P_0\right) + 1 \le \mathbb{E}\left[\exp\left(nt^2 \sum_{j=1}^k \sigma_j^2 V_j V_j'\right)\right] \le \mathbb{E}\left[\exp\left(\frac{n^2 t^4}{2} \sum_{j=1}^k \sigma_j^4 V_j^2\right)\right]$$

by Hoeffding's Lemma (see Example 4.1.6), as $V_j \stackrel{\text{iid}}{\sim} \mathsf{Uniform}(\{\pm 1\})$ Noting that $V_j^2 = 1$ gives the first part of the lemma. The final statement is immediate once we observe that $e^x \le 1 + (e-1)x \le 1 + 2x$ for $0 \le x \le 1$. □

10.7 Exercises

Exercise 10.1: Recall the Hellinger distance between distributions P and Q with densities p, q is $d_{\rm hel}(P,Q)^2 = \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx$. Let P be $\mathsf{N}(\mu_0, \Sigma)$ and Q be $\mathsf{N}(\mu_1, \Sigma)$. Show that

$$\frac{1}{2} d_{\rm hel}(P,Q)^2 = 1 - \exp\left(-\frac{1}{8}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1)\right).$$

Exercise 10.2: Suppose that the test Ψ has test risk for testing between $\mathcal{P}_0$ and $\mathcal{P}_1$ satisfying $R_n(\Psi \mid \mathcal{P}_0, \mathcal{P}_1) \le \frac{1}{3}$ Let $k \in \mathbb{N}$. Show how, given a sample of size kn, we can develop a test $\Psi^\star$ with

$$R_{kn}(\Psi^\star \mid \mathcal{P}_0, \mathcal{P}_1) \le 2\exp\left(-ck\right),$$

where $c > 0$ is a numerical constant. *Hint.* Split the sample into k samples of size n, and then apply Ψ to each.

Exercise 10.3 (Poissonization: lower bounds [174])**:** Prove the lower bound in Proposition 10.2.7, inequality (10.2.8), that is, that for numerical constants C, c,

$$\mathfrak{M}_{\mathsf{Poi}(2n)} - Cr^2 \exp(-cn) \le \mathfrak{M}_n.$$

Hint. Bound $\mathfrak{M}_{\mathsf{Poi}(2n)}$ with a weighted sum of $\mathfrak{M}_m$. Use the MGF calculation that for $X \sim \mathsf{Poi}(\lambda)$, $\mathbb{E}[e^{tX}] = \exp(\lambda(e^t - 1))$ to show that $N \sim \mathsf{Poi}(2n)$ is concentrated above n.

Exercise 10.4 (Poissonization: upper bounds [174])**:** Assume the minimax result that

$$\mathfrak{M}_n = \sup_\pi \inf_{T_n} \mathbb{E}\left[(T_n(X_1^n) - T(p))^2\right],$$

where the supremum is over probability distributions (priors π) on $p \in \Delta_k$, and the expectation is now over the random choice of p and the sample $X_1^n \stackrel{\rm iid}{\sim} p$ drawn conditional on p. (This is a standard infinite-dimensional saddle point result generalizing von-Neumann's minimax theorem; cf. [81, 160].) You will show the upper bound in Proposition 10.2.7, Eq. (10.2.8).

Let $\{T_m\}$ be an arbitrary sequence of estimators and define the sequence of averaged risks

$$r_m := \mathbb{E}[(T_m(X_1^m) - T(p))^2].$$

Define the modified risks $\widetilde{r}_m = \min\{r_1, \ldots, r_m\} = \min\{\widetilde{r}_{m-1}, r_m\}$, and the "corrected" estimators

$$\widetilde{T}_m(x_1^m) := \begin{cases} \widetilde{T}_{m-1}(x_1^{m-1}) & \text{if } r_m \ge \widetilde{r}_{m-1} \\ T_m(x_1^m) & \text{if } r_m < \widetilde{r}_{m-1}. \end{cases}$$

(a) Show that $\mathbb{E}[(\widetilde{T}_m(X_1^m) - T(p))^2] \le \mathbb{E}[(T_m(X_1^m) - T(p))^2]$.

(b) Show that

$$\frac{1}{2}\inf_{T_n} \mathbb{E}[(T_n(X_1^n) - T(p))^2] \le \mathbb{E}[(T_N(X_1^N) - T(p))^2]$$

for $N \sim \mathsf{Poi}(n/2)$ and $p \sim \pi$, then X_i drawn i.i.d. conditionally on p.

(c) Finalize the proof of the upper bound in inequality (10.2.8).

Exercise 10.5: Consider the hypothesis testing problem of testing whether a collection of independent Bernoulli random variables $X_1, \ldots, X_n$ is is fair (H_0, so that $\mathbb{P}(X_i = 1) = \frac{1}{2}$ for each i) or that there are unfair subcollections. That is, we wish to test

$$\begin{aligned} H_0: &\quad X_i \stackrel{\text{iid}}{\sim} \mathsf{Bernoulli}(\tfrac{1}{2}) \\ H_1: &\quad X_i \stackrel{\text{ind}}{\sim} \mathsf{Bernoulli}(\tfrac{1+\theta_i}{2}),\ \theta \in C \end{aligned}$$

for a set $C \subset [-1,1]^n$. Show that if the set C is orthosymmetric, meaning that whenever $\theta \in C$ then $S\theta \in C$ for any diagonal matrix S of signs, i.e. $\mathrm{diag}(S) \in \{\pm 1\}^n$, then no test can reliably distinguish H_0 from H_1 (in a minimax sense). *Hint.* Let $v \in \mathcal{V} := \{\pm 1\}^n$ index coordinate signs and define $\theta_v = Dv$ for some diagonal D, where $Dv \in C$. Let P_v be the product distribution with $X_i \sim \mathsf{Bernoulli}(\frac{1+D_i v_i}{2})$. What is $\frac{1}{2^n}\sum_{v \in \mathcal{V}} P_v$?

Exercise 10.6 (Testing a trend in independent Bernoullis)**:** Consider testing whether a collection of Bernoulli random variables has an "upward trend" over time, by which we mean that if $X_i \sim \mathsf{Bernoulli}(p_i)$ independently, then

$$\overline{p}_{\text{end}} := \frac{1}{n/4} \sum_{i=\frac{3n}{4}+1}^{n} p_i > \overline{p}_{\text{beg}} := \frac{1}{n/4} \sum_{i=1}^{n/4} p_i.$$

Consider the following more quantitative version of this problem: we wish to test

$$\begin{aligned} H_0: &\quad X_i \stackrel{\text{iid}}{\sim} \mathsf{Bernoulli}(\tfrac{1}{2}) \\ H_1: &\quad X_i \stackrel{\text{ind}}{\sim} \mathsf{Bernoulli}(p_i),\ \overline{p}_{\text{end}} - \overline{p}_{\text{beg}} \geq \delta. \end{aligned}$$

(a) Use Le Cam's two-point method to show that there exists a numerical constant $c > 0$ such that for $\delta \leq \frac{c}{\sqrt{n}}$, no test can reliably distinguish H_0 from H_1.

(b) Use the statistic

$$T_n := \frac{1}{n/4} \sum_{i=3n/4+1}^{n} X_i - \frac{1}{n/4} \sum_{i=1}^{n/4} X_i$$

to develop a test Ψ (use Proposition 10.2.2) that achieves test risk $R_n(\Psi \mid H_0, H_1) \leq \frac{1}{4}$ whenever $\delta \geq \frac{C}{\sqrt{n}}$, where $C < \infty$ is a constant.

Exercise 10.7: Prove the identity (10.2.10).

Exercise 10.8 (Unbiased estimators of distance for multinomials)**:** Let $X_i \stackrel{\text{iid}}{\sim} p$, $i = 1, \ldots, n$, and $Y_i \stackrel{\text{iid}}{\sim} q$, $i = 1, \ldots, m$, meaning that X_1^n and Y_1^m are multinomial samples for $p, q \in \Delta_d$. Define the empirical estimators $\widehat{p}_j = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = j\}$ and $\widehat{q}_j = \frac{1}{m}\sum_{i=1}^n \mathbf{1}\{Y_i = j\}$.

(a) Give $\mathbb{E}[\|\widehat{p}\|_2^2]$.

(b) Show that $T_n := \|\widehat{p} - \widehat{q}\|_2^2$ satisfies

$$\mathbb{E}[T_n] = \|p - q\|_2^2 + \frac{1}{n} + \frac{1}{m} - \frac{1}{n}\|p\|_2^2 - \frac{1}{n}\|q\|_2^2.$$

(c) Modify T_n into a new statistic $T_n^{\rm unb}$ so that $\mathbb{E}[T_n^{\rm unb}] = \|p - q\|_2^2$.

Exercise 10.9: Show that in the hypothesis testing problem (10.2.12), there is a numerical constant $c > 0$ such that $\delta \le c/\sqrt{n}$ implies that no test can reliably distinguish H_0 from H_1.

JCD Comment:

1. Poissonization: remark in main text.
2. Work through Liam's ℓ_1-multinomial testing
3. Lower bound for testing whether collection of coins is fair or some number are unfair.

Part III

Entropy, predictions, divergences, and information

Chapter 11

Predictions, loss functions, and entropies

In prediction problems broadly construed, we have a random variable X and a label, or target or response, Y, and we wish to fit a model or predictive function that accurately predicts the value of Y given X. There are several perspectives possible when we consider such problems, each with attendant advantages and challenges. We can roughly divide these into three approaches, though there is considerable overlap between the tools, techniques, and goals of the three:

(1) Point prediction, where we wish to find a prediction function f so that $f(X)$ most accurately predicts Y itself.

(2) Probabilistic prediction, where we output a predicted distribution P of Y, and we seek $\mathbb{P}(Y = y \mid X = x) \approx P(Y = y \mid X = x)$, where here $\mathbb{P}$ denotes the "true" probability and P the predicted one. A relaxed version of this is *calibration*, the subject of the next chapter, where we ask that $\mathbb{P}(Y = y \mid P) \approx P(Y = y)$, that is, the distribution of Y given a predicted distribution P is accurate.

(3) Predictive inference, where for a given level $\alpha \in (0, 1)$, we seek a confidence set mapping C such that $\mathbb{P}(Y \in C(X)) \approx 1 - \alpha$.

We focus mostly on the former two, though there is overlap between the approaches.

In this first chapter of the sequence, we focus on the probabilistic prediction problem. Our main goal will be to elucidate and identify loss functions for choosing probabilistic predictions that are *proper*, meaning that the true distribution of Y minimizes the loss, and *strictly proper*, meaning that the true distribution of Y uniquely minimizes the loss. As part of this, we will develop mappings between losses and entropy-type functionals; these will repose on convex analytic techniques for their cleanest statements, highlighting the links between convex analysis, prediction, and information. Moreover, we highlight how *any* proper loss (which will be defined) is in correspondence with a particular measure of entropy on the distribution P, and how these connect with an object known as the *Bregman divergence* central to convex optimization. For the deepest understanding of this chapter, it will therefore be useful to review the basic concepts of convexity (e.g., convex sets, functions, and subgradients) in Appendix B, as well as the more subtle tools on optimality and stability of solutions to convex optimization problems in Appendix C. We give an overview of the important results in Section 11.1.1.

11.1 Proper losses, scoring rules, and generalized entropies

As motivation, consider a weather forecasting problem: a meteorologist wishes to prediction the weather Y_t on days $t = 1, 2, \ldots$, where $Y_t = 1$ indicates rain and $Y_t = 0$ indicates no rain. At time t, using covariates X_t (for example, the weather the previous day, long term trends, or simulations), the forecaster predicts a probability $p_t \in [0, 1]$. We would like the forecaster's predictions to be as accurate as possible, so that $\mathbb{P}(Y_t = 1) \approx p_t$. Following the standard dicta of decision theory, we choose a loss function $\ell(p, y)$ that scores a prediction p for a given outcome y. Ideally, the forecaster should have an incentive to make predictions as accurately as possible, so the distribution minimizing the expected loss should coincide with the true distribution of Y.

This leads to proper losses. In our treatment, we will sometimes allow infinite values, so we work with the upper and lower extended real lines, recalling that $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ and $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$.

Definition 11.1. *Let $\mathcal{P}$ be a collection of distributions on $\mathcal{Y}$. A loss $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is* proper *if, whenever $Y \sim P \in \mathcal{P}$,*

$$\mathbb{E}[\ell(P, Y)] \le \mathbb{E}[\ell(Q, Y)] \quad \text{for all } Q \in \mathcal{P}.$$

The loss is strictly proper *if the preceding inequality is strict whenever $Q \neq P$.*

In much of the literature on prediction, one instead considers *proper scoring rules*, which are simply negative proper losses, that is, functions $S : \mathcal{P} \times \mathcal{Y}$ satisfying $S(P, y) = -\ell(P, y)$ for a (strictly) proper loss. We focus on losses for consistency with the convex analytic tools we develop. In addition, frequently we will work with discrete distributions, so that Y has a probability mass function (p.m.f.), in which case we will use $p \in \Delta_k := \{p \in \mathbb{R}^k_+ \mid \langle \mathbf{1}, p \rangle = 1\}$ to identify the distribution and $\ell(p, y)$ instead of $\ell(P, y)$.

Perhaps the two most famous proper losses are the log loss and the squared loss (often termed *Brier scoring*). For simplicity let us assume that $\mathcal{Y} \in \{1, 2, \ldots, k\}$, and let $\Delta_k = \{p \in \mathbb{R}^k_+ \mid \mathbf{1}^T p = 1\}$ be the probability simplex; we then identify distributions P on $\mathcal{Y}$ with vectors $p \in \Delta_k$, and abuse notation to write $\ell(p, y)$ accordingly and when it is unambiguous. The squared loss is then

$$\ell_{\rm sq}(p, y) = (p_y - 1)^2 + \sum_{i \neq y} p_i^2 = \|p - e_y\|_2^2 ,$$

where e_y is the yth standard basis vector, while the log loss (really, the negative logarithm) is

$$\ell_{\log}(p, y) = -\log p_y.$$

Both of these are strictly proper. To this propriety, let Y have p.m.f. $p \in \Delta_k$, so that $\mathbb{P}(Y = y) = p_y$. Then for the squared loss and any $q \in \Delta_k$, we have

$$\mathbb{E}[\ell_{\rm sq}(q, Y)] - \mathbb{E}[\ell_{\rm sq}(p, Y)] = \mathbb{E}[\|q - e_Y\|_2^2] - \mathbb{E}[\|p - e_Y\|_2^2] = \|q\|_2^2 - 2\langle q, p \rangle + 2\langle p, p \rangle = \|q - p\|_2^2 .$$

For the log loss, we have

$$\mathbb{E}[\ell_{\log}(q, Y)] - \mathbb{E}[\ell_{\log}(p, Y)] = -\sum_{y=1}^k p_y \log q_y + \sum_{y=1}^k p_y \log p_y = \sum_{y=1}^k p_y \log \frac{p_y}{q_y} = D_{\rm kl}\left(p \| q\right).$$

It is immediate that $q = p$ uniquely minimizes each loss.

That the gap between the expected losses at q and p reduced to a particular divergence-like measure—the squared ℓ_2-distance in the case of the squared loss and the KL-divergence in the

case of the log loss—is no accident. In fact, for proper losses, we will show that this divergence representation necessarily holds.

The key underlying our development is a particular construction, which we present in Section 11.1.2, that transforms a loss into a generalized notion of entropy. Because it is so central, we highlight it here, though before doing so, we take a brief detour through a few of the concepts in convexity we require. Figures representing these results capture most of the mathematical content, while Chapters B and C in the appendices contain proofs of the results we require.

11.1.1 A convexity primer

Recall that a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is convex if for all $x, y \in \dom f$ and $\lambda \in [0, 1]$, we have

$$f(\lambda x + (1 - \lambda)y) \le \lambda f(x) + (1 - \lambda)f(y),$$

where for $x \notin \dom f$ we define $f(x) = +\infty$. We exclusively work with proper convex functions, so that $f(x) > -\infty$ for each x. Typically, we work with closed convex f, meaning that the epigraph $\epi f = \{(x, t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \le t\} \subset \mathbb{R}^{d+1}$ is a closed set; equivalently, f is lower semi-continuous, so that $\liminf_{y \to x} f(y) \ge f(x)$. A concave function f is one for which $-f$ is convex.

Three main concepts form the basis for our development. The first is the *subgradient* (see Appendix B.3). For a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$, the *subgradient set* (also called the subdifferential) at the point x is

$$\partial f(x) := \left\{s \in \mathbb{R}^d \mid f(y) \ge f(x) + \langle s, y - x\rangle \text{ for all } y \in \mathbb{R}^d\right\}. \tag{11.1.1}$$

If f is a convex function, then at any point x in the relative interior of its domain, $\partial f(x)$ is nonempty (Theorem B.3.3). Moreover, a quick calculation shows that x minimizes $f(x)$ if and only if $0 \in \partial f(x)$, and (a more challenging calculation) that if $\partial f(x) = \{s\}$ is a singleton, then $s = \nabla f(x)$ is the usual gradient. See the left plot of Figure 11.1. We shall in some cases allow subgradients to take values in the extended reals $\overline{\mathbb{R}}^k$ and $\underline{\mathbb{R}}^k$, which will necessitate some additional care.

The second concept is that the supremum of a collection of convex functions is always convex, that is, if f_α is convex for each index $\alpha \in \mathcal{A}$, then

$$f(x) := \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$$

is convex, and f is closed in f_α is closed for each α. The closure of f is immediate because $\epi f = \bigcap \epi f_\alpha$, and convexity follows because

$$f(\lambda x + (1 - \lambda)y) \le \sup_{\alpha \in \mathcal{A}} \{\lambda f_\alpha(x) + (1 - \lambda) f_\alpha(y)\} \le \lambda \sup_{\alpha \in \mathcal{A}} f_\alpha(x) + (1 - \lambda) \sup_{\alpha \in \mathcal{A}} f_\alpha(y).$$

Conveniently, subdifferentiability of individual f_α implies the subdifferentiability of f when the supremum is attained. Indeed, let $\mathcal{A}(x) = \{\alpha \mid f_\alpha(x) = f(x)\}$. Then

$$\partial f(x) \subset \Conv \{s_\alpha \in \partial f_\alpha(x) \mid \alpha \in \mathcal{A}(x)\} \tag{11.1.2}$$

because if $s = \sum_{\alpha \in \mathcal{A}(x)} \lambda_\alpha s_\alpha$ for some $\lambda_\alpha \ge 0$ with $\sum_\alpha \lambda_\alpha = 1$, then

$$f(y) \ge \sum_{\alpha \in \mathcal{A}(x)} \lambda_\alpha f_\alpha(y) \ge \sum_{\alpha \in \mathcal{A}(x)} \lambda_\alpha \left[f_\alpha(x) + \langle s_\alpha, y - x\rangle\right] = f(x) + \langle s, y - x\rangle.$$

See the right plot of Figure 11.1.

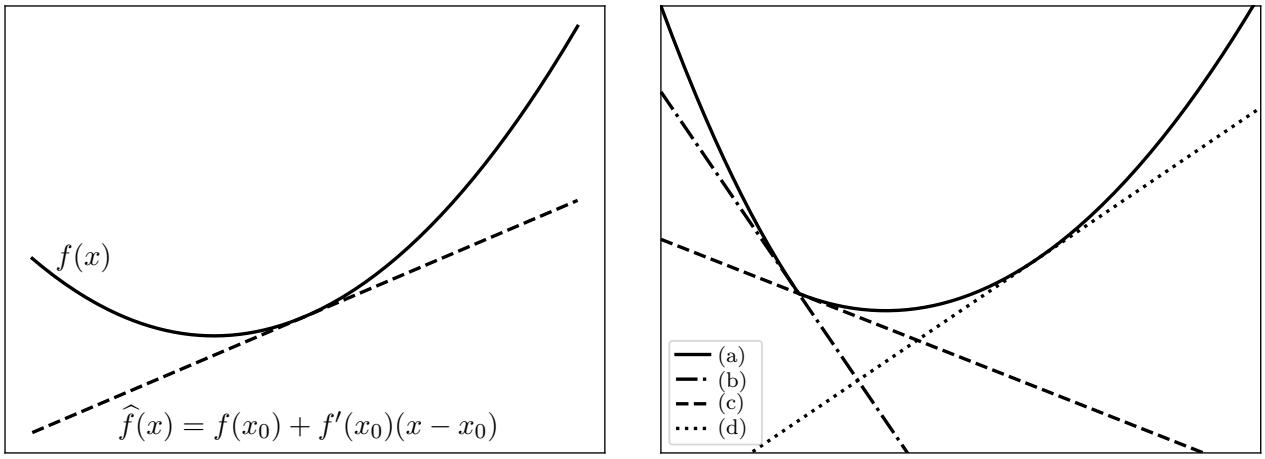

Figure 11.1. *Left:* The quadratic $f(x) = \frac{1}{2}x^2$ and the linear approximation $\widehat{f}(x) = f(x_0)+s(x-x_0)$, where $x_0 = \frac{1}{2}$ and $s = f'(x_0)$. *Right:* the piecewise quadratic $f(x) = \max\{f_0(x), f_1(x)\}$ where $f_0(x) = \frac{1}{2}x^2$ and $f_1(x) = \frac{1}{4}(x + \frac{1}{4})^2 + \frac{1}{8}$, intersecting at $x_0 = \frac{1-\sqrt{10}}{4}$. (a) The function $f(x)$. (b) The linear underestimator $\widehat{f}(x) = f(x_0) + s_0(x - x_0)$ for $s_0 = f_0'(x_0)$. (c) The linear underestimator $\widehat{f}(x) = f(x_0)+s_1(x-x_0)$ for $s_1 = f_1'(x_0)$. (d) The linear approximation $\widehat{f}(x) = f(x_1)+f'(x_1)(x-x_1)$ around the point $x_1 = \frac{1}{4}$.

Lastly, we revisit a special duality relationship that all closed convex functions f enjoy (see Appendix C.2 for a fuller treatment). The *Fenchel-Legendre conjugate* or *convex conjugate* of a function f is

$$f^*(s) := \sup_x \left\{\langle s, x\rangle - f(x)\right\}. \tag{11.1.3}$$

The function f^* is always convex, as it is the supremum of linear functions of s, and for any $x^\star(s)$ maximizing $\langle s, x\rangle - f(x)$, we have that $x^\star(s) \in \partial_s f^*(s)$ by the relationship (11.1.2); by a bit more work, we see that if $s \in \partial f(x)$, then $0 \in \partial_x\{f(x) - \langle s, x\rangle\}$ and so x maximizes $\langle s, x\rangle - f(x)$. See Figure 11.2 for a graphical representation of this process. Flipping this argument by replacing f with f^* and x with s, when $s \in \partial f(x)$ and x maximizes $\langle s, x\rangle - f(x)$ in x, then $x \in \partial f^*(s)$ and so s maximizes $\langle s, x\rangle - f^*(s)$ in s. From this development comes the *biconjugate*, that is, $f^{**}(x) = \sup_s\{\langle s, x\rangle - f^*(s)\}$, or $f^{**} = (f^*)^*$. The biconjugate f^{**}, it turns out, is the supremum of *all* linear functionals below f, because $\langle s, x\rangle - f^*(s) \le f(x)$ for all s, and if $\partial f(x)$ is non-empty, then the preceding argument guarantees that $\langle s, x\rangle - f^*(s) = f(x)$ for $s \in \partial f(x)$. Theorem C.2.1 in the appendices makes this rigorous, and shows that if f is a closed convex function, then

$$f(x) = f^{**}(x) = \sup_s \left\{\langle s, x\rangle - f^*(s)\right\}$$

for all x. In particular, by passing through the conjugate, we can recover the function f directly whenever f is closed convex.

We immediately have the *Fenchel-Young inequality* that

$$f^*(s) + f(x) \ge \langle s, x\rangle \quad \text{for all} \quad s, x,$$

and (see Proposition C.2.2) if f is a closed convex function, then equality holds if and only if

$$s \in \partial f(x) \quad \text{or} \quad x \in \partial f^*(s), \tag{11.1.4}$$

which are equivalent. Thus we obtain the identities

$$\partial f^* = (\partial f)^{-1} \quad \text{and} \quad \partial f = (\partial f^*)^{-1},$$

and we have the characterization

$$\partial f^*(s) = \operatorname*{argmin}_x \left\{-\langle s, x\rangle + f(x)\right\} = \operatorname*{argmax}_x \left\{\langle s, x\rangle - f(x)\right\}.$$

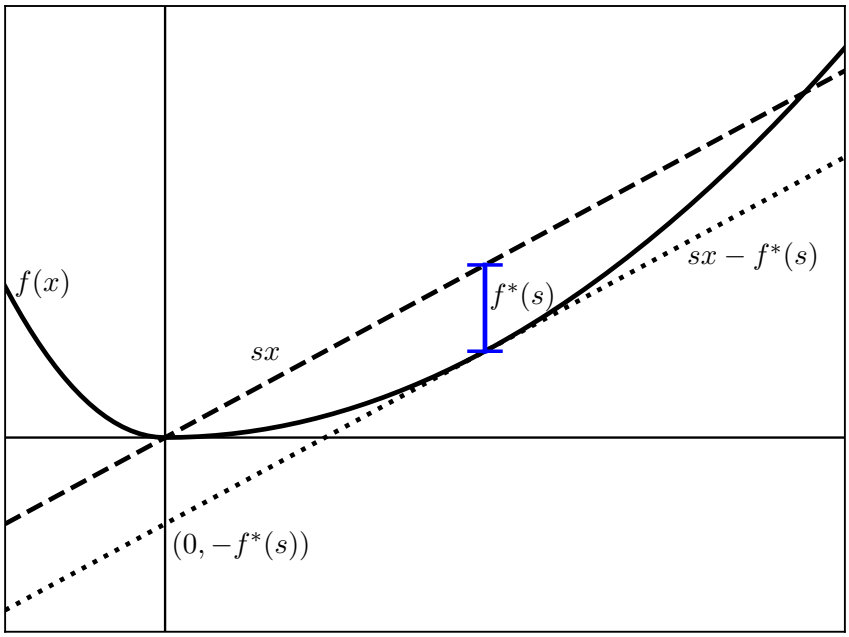

Figure 11.2. The conjugate function. The line of long dashes is $f(x) = sx$, while the dotted line is $x \mapsto sx - f^*(s)$. The blue line is the largest gap between sx and $f(x)$, which equals $f^*(s)$. Note that $x \mapsto sx - f^*(s)$ meets the graph of $f(x)$ at exactly the point of maximum difference $sx - f(x)$, where $f'(x) = s$.

11.1.2 From a proper loss to an entropy

The key construction underlying all of our proper losses is the optimal value of the expected loss. To any loss ℓ acting on a family $\mathcal{P}$ of distributions, we construct the *generalized entropy* associated with the loss ℓ by

$$H_\ell(Y) := \inf_{Q \in \mathcal{P}} \mathbb{E}[\ell(Q, Y)], \tag{11.1.5}$$

where we have paralleled the typical notation $H(Y)$ for the Shannon entropy. In many cases, it will be more convenient to write this entropy directly as a function of the distribution P of Y, in which case we write

$$H_\ell(P) = \inf_{Q \in \mathcal{P}} \mathbb{E}_P[\ell(Q, Y)], \tag{11.1.6}$$

where Y follows the distribution P; we will use whichever is more convenient. As the notation (11.1.6) makes clear, $H_\ell(P)$ is the infimum of a collection of linear functions of the form $P \mapsto \mathbb{E}_P[\ell(Q, Y)]$, one for each $Q \in \mathcal{P}$), so that necessarily $H_\ell(P)$ is concave in P. The remainder

of this chapter, and several parts of the coming chapters, highlights the ways that this particular quantity informs the properties of the loss ℓ, and more generally, how we may always view any concave function H on a family of distributions $\mathcal{P}$ as a generalized entropy function.

In Section 11.2, we show how such entropy-type functionals map back to losses themselves, so for now we content ourselves with a few examples to see why we call these entropies. Let us temporarily assume that Y has finite support $\{1, \ldots, k\}$ with $\mathcal{P} = \Delta_k = \{p \in \mathbb{R}_+^k \mid \langle \mathbf{1}, p \rangle = 1\}$ the collection of probability mass functions on elements $\{1, \ldots, k\}$.

Example 11.1.1 (Log loss)**:** Consider the log loss $\ell_{\log}(p, y) = -\log p_y$. Then

$$H_{\ell_{\log}}(p) = \inf_{q \in \Delta_k} \mathbb{E}_p[-\log q_Y] = \inf_{q \in \Delta_k} \left\{ -\sum_{y=1}^k p_y \log \frac{q_y}{p_y} - \sum_{y=1}^k p_y \log p_y \right\} = -\sum_{y=1}^k p_y \log p_y,$$

the classical Shannon entropy. $\diamond$

This highlights an operational interpretation of entropy distinct from that arising in coding: the (Shannon) entropy is the minimal expected loss of a player in a prediction game, where the player chooses a distribution Q on Y, nature draws $Y \sim P$, and upon observing $Y = y$, the player suffers loss $-\log Q(Y = y)$.

Example 11.1.2 (0-1 error)**:** If instead we take the 0-1 loss, that is, $\ell_{0\text{-}1}(p, y) = 1$ if $p_y \leq p_j$ for some $j \neq y$ and $\ell_{0\text{-}1}(p, y) = 0$ otherwise, then

$$H_{\ell_{0\text{-}1}}(p) = \inf_{q \in \Delta_k} \mathbb{E}_p[\ell(q, y)] = 1 - \max_y p_y.$$

So $H_{\ell_{0\text{-}1}}(e_y) = 0$ for any standard basis vector, that is, distribution with all mass on a single point y, and $H_{\ell_{0\text{-}1}}(p) > 0$ otherwise. Moreover, the vector $p = \mathbf{1}/k$ maximizes $H_{\ell_{0\text{-}1}}(p)$, with $H_{\ell_{0\text{-}1}}(\mathbf{1}/k) = \frac{k-1}{k}$. $\diamond$

Example 11.1.3 (Brier scoring and squared error)**:** For the squared error (Brier scoring) loss $\ell_{\rm sq}(p, y) = \|p - e_y\|_2^2$, where $e_y \in \{0, 1\}^k$ is the yth standard basis vector, let Y have p.m.f. $p \in \Delta_k$. Then

$$H_{\ell_{\rm sq}}(Y) = \mathbb{E}[\ell_{\rm sq}(p, Y)] = \|p\|_2^2 - 2\|p\|_2^2 + 1 = 1 - \|p\|_2^2 .$$

So as above, we have $H_{\ell_{\rm sq}}(Y) \geq 0$, with $H_{\ell_{\rm sq}}(Y) = 0$ if and only if Y is a point mass on one of $\{1, \ldots, k\}$, and the uniform distribution with p.m.f. $p = \frac{1}{k}\mathbf{1}$ maximizes the entropy, with $H_{\ell_{\rm sq}}(\mathsf{Uniform}([k])) = 1 - 1/k$. $\diamond$

These examples highlight how these entropy functions are types of uncertainty measures, giving rise to "maximally uncertain" distributions p, which are typically uniform on Y.

11.1.3 The information in an experiment

In classical information theory, the mutual (Shannon) information between random variables X and Y is the gap between the entropy of Y and the remaining entropy given X, that is,

$$I(X; Y) = H(Y) - H(Y \mid X).$$

In complete analogy with our development in Chapter 2, then, we can define the information between variables X and Y relative to a particular loss function ℓ. Thus, we define the *ℓ-conditional entropy*

$$H_\ell(Y \mid X = x) := \inf_{Q \in \mathcal{P}} \mathbb{E}\left[\ell(Q, Y) \mid X = x\right]$$

and, in analogy to the definitions in Section 2.1.1, the conditional entropy of Y given X is

$$H_\ell(Y \mid X) := \mathbb{E}\left[\inf_{Q \in \mathcal{P}} \mathbb{E}[\ell(Q, Y) \mid X]\right] = \int_{\mathcal{X}} H_\ell(Y \mid X = x) dP(x),$$

the average minimal expected loss when one observes X.

With this definition, we then can discuss the *information in an experiment*. This nomenclature follows classical statistical parlance, where by an experiment, we mean the observation of a variable X in a Markov chain $X \to Y$, where we think of Y as a hypothesis to be tested or a value to be predicted, and we ask how much observing X helps to actually allow this prediction. Then we define

$$I_\ell(X; Y) := H_\ell(Y) - H_\ell(Y \mid X), \tag{11.1.7}$$

which is nonnegative and is the gap between the prior entropy of Y and its posterior entropy conditional on the observation X. That is, this information measure is precisely the gap between the best achievable loss in the prediction of a distribution P for Y *a priori*, when we observe nothing, and that achievable *a posteriori*, when we observe X. In parallel to our alternative view of the entropy as the (expected) minimal loss of a player in a prediction game, then, the information between X and Y is the improvement an observation X offers a player in predicting Y when measuring error with the loss ℓ. The information (11.1.7) is typically asymmetrical in X and Y, so we are careful about the ordering (this lack of symmetric holds, essentially, unless ℓ is the log loss).

The next three examples show different information quantities, where in each we let $\mathcal{Y}$ have finite cardinality k, and thus identify $\mathcal{P}$ with the probability simplex $\Delta_k = \{p \in \mathbb{R}^k_+ \mid \langle \mathbf{1}, p \rangle = 1\}$.

Example 11.1.4 (Shannon information)**:** Taking the log loss $\ell(p, y) = -\log p_y$, we have

$$I_\ell(X; Y) = H_\ell(Y) - H_\ell(Y \mid X) = H(Y) - H(Y \mid X) = I(X; Y),$$

the classical Shannon information. $\diamond$

Example 11.1.5 (0-1 error)**:** Consider the 0-1 error $\ell_{0\text{-}1}(p, y) = 1$ if $p_y \le \max_j p_j$ and $\ell_{0\text{-}1}(p, y) = 0$ if $p_y > \max_{j \ne y} p_j$. Then letting $y^\star = \operatorname{argmax}_y \mathbb{P}(Y = y)$ and $y^\star(x) = \operatorname{argmax}_y \mathbb{P}(Y = y \mid X = x)$, we have

$$I_{\ell_{0\text{-}1}}(X; Y) = \mathbb{P}(Y = y^\star) - \mathbb{E}\left[\mathbb{P}(Y = y^\star(X) \mid X)\right] = \mathbb{P}(Y = y^\star) - \mathbb{P}(Y = y^\star(X)),$$

the gap between the prior probability of making a mistake when guessing Y and the posterior probability given X. $\diamond$

Example 11.1.6 (Squared error)**:** For the Brier score with squared error $\ell_{\text{sq}}(p, y) = \|p - e_y\|_2^2$, we have $H_{\ell_{\text{sq}}}(p) = 1 - \|p\|_2^2$, and so

$$I_{\ell_{\text{sq}}}(X; Y) = \sum_{j=1}^k \mathbb{E}\left[\mathbb{P}(Y = j \mid X)^2\right] - \sum_{j=1}^k \mathbb{P}(Y = j)^2 = \sum_{j=1}^k \operatorname{Var}(\mathbb{P}(Y = j \mid X)),$$

the summed variances of the random variables $\mathbb{P}(Y = j \mid X)$. The higher the variance of these quantities, the more information X carries about Y. $\diamond$

11.2 Characterizing proper losses and Bregman divergences

With the definition (11.1.5) of the fundamental generalized entropy, we can now proceed to a characterization of all proper losses. We do this in three settings: in the first (Section 11.2.1), we give a representation for proper losses when $\mathcal{Y}$ is finite and discrete, so we can identify it with $\mathcal{Y} = \{1, \ldots, k\}$ and distributions P on $\mathcal{Y}$ with probability mass functions $p \in \Delta_k$. We then demonstrate a full characterization of propriety (Section 11.2.2), which requires measure-theoretic tools and can be skipped. As the final approach to considering propriety, we modify the results for finite $\mathcal{Y}$ to consider cases in which Y is vector-valued and $\mathcal{Y} \subset \mathbb{R}^k$ is contained in a compact set. This case transparently generalizes the finite representations of Section 11.2.1 and will form the basis of our development going forward, as it allows us to more directly apply to tools of convexity and analysis.

11.2.1 Characterizing proper losses for Y taking finitely many vales

Here, we present the *Savage representation* of proper losses, which characterizes all proper losses using the entropies (11.1.5) or, equivalently, (11.1.6). To avoid pathological cases, we work with regular losses, which always assign a finite value to the correct predicted distribution; we assume regularity without further comment.

Definition 11.2. *Let $\mathcal{P}$ be a family of distribution on $\mathcal{Y}$. The loss $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is* regular for the family $\mathcal{P}$ *if $\mathbb{E}_P[\ell(P, Y)]$ is real valued for all $P \in \mathcal{P}$.*

We do allow losses to attain infinite values, for example, we can allow $\ell(Q, y) = +\infty$ if Q assigns probability 0 to an event y, as in the case of the logarithmic loss. The following theorem then provides the promised representation of proper losses, and additionally, highlights the centrality of the generalized entropy functionals.

Theorem 11.2.1 (Proper scoring rules: the finite case)**.** *Let $\mathcal{Y} = \{1, \ldots, k\}$ be finite and $\mathcal{P} \subset \Delta_k$ a convex collection of distributions on $\mathcal{Y}$. Then the following are true.*

(i) If the loss $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ satisfies the representation

$$\ell(p, y) = -h(p) - \langle \nabla h(p), e_y - p \rangle \tag{11.2.1}$$

for a subdifferentiable closed convex function $h : \mathcal{P} \to \mathbb{R}$, where $\nabla h(p) \in \partial h(p)$, then ℓ is proper.

(ii) Conversely, if ℓ is proper, then choosing h to be the negative generalized entropy

$$h_\ell(p) := -H_\ell(p) = \sup_q \left\{-\mathbb{E}_p[\ell(q, Y)] \mid q \in \mathcal{P}\right\}$$

satisfies equality (11.2.1) (and h is closed).

Additionally, if ℓ is real valued, then $\nabla h(p) \in \mathbb{R}^k$ in the representation (11.2.1). If $\ell(p, y)$ can take the value $+\infty$, then we allow $\nabla h(p) \in \underline{\mathbb{R}}^k$ when $p \notin \operatorname{relint} \Delta_k$. The loss is strictly proper if and only if the convex h is strictly convex.

Proof If ℓ has the given representation and $\mathbb{P}(Y = y) = p_y$, then we have

$$\mathbb{E}[\ell(q, Y)] = -h(q) - \langle \nabla h(q), p - q \rangle \geq -h(p) = \mathbb{E}[\ell(p, Y)]$$

by the first-order convexity property of convex functions (that is, the definition (11.1.1)) of a subdifferential).

Conversely, suppose that the loss is proper, and let $h(p) = h_\ell(p)$. Clearly h is convex, as it is the supremum of linear functionals of p. Moreover, propriety of ℓ guarantees that

$$h(p) \geq -\mathbb{E}[\ell(q, Y)] = h(q) + \sum_{y=1}^{k} -\ell(q, y)(p_k - q_k)$$

That is, for each $q \in \mathcal{P}$ the vector $[-\ell(q, y)]_{y=1}^k \in \partial h(q)$, so h is subdifferentiable. Choosing the vector $\nabla h(p) = [-\ell(p, y)]_{y=1}^k$, we have

$$\ell(p, y) = -h(p) + \ell(p, y) + h(p) = -h(p) - \sum_{i=1}^{k} p_i \ell(p, i) + \ell(p, y) = -h(p) - \langle \nabla h(p), e_y - p \rangle$$

as desired. Note that $\ell(p, y) < \infty$ except when $p_y = 0$, in which case our definition $\nabla h(p) = [-\ell(p, y)]_{y=1}^k$ remains sensible as $-\langle \nabla h(p), e_y - p \rangle = +\infty$.

As an alternative argument more directly using convexity, definition of $h(p) = \sup_q \{-\mathbb{E}_p[\ell(q, Y)] \mid q \in \mathcal{P}\}$ and the immediate calculation (11.1.2) of the subdifferential of the supremum shows that

$$\partial h(p) \supset \left\{ [-\ell(q, y)]_{y=1}^k \mid q \in \Delta_k \text{ satisfies } -\mathbb{E}_p[\ell(q, Y)] = h(p) \right\}.$$

But propriety guarantees that the set of such q includes p, so that $\partial h(p) \supset [-\ell(p, y)]_{y=1}^k$.

For the strict inequalities and strict propriety, trace the argument replacing inequalities with strict inequalities for $q \neq p$ and use Corollary B.3.2 or C.1.7. □

The negative generalized entropy h in Theorem 11.2.1 is essentially unique and marks an important duality between proper losses and convex functions: to each loss, we can assign a generalized entropy, and from this generalized entropy, we can reconstruct the loss. Exercise 11.2 explores this connection. We can also give a few examples that show how to recover standard losses. For each, we begin with a convex function h, then exhibit the associated proper or strictly proper scoring rule. One thing to notice in this representation is that, typically, we do *not* expect to achieve a loss function convex in p, which is a weakness of the representation (11.2.1). In Section 11.3 (and Chapter 14 in more depth), however, we will show how to convert suitable proper losses into *surrogates* that are convex in their arguments and which, after a particular transformation based on convex duality, are proper and yield the correct distributional predictions. We defer this, however, and instead provide a few examples.

Example 11.2.2 (Logarithmic losses)**:** Consider the negative entropy $h(p) = \sum_{y=1}^k p_y \log p_y$. We have $\frac{\partial}{\partial p_y} h(p) = 1 + \log p_y \in [-\infty, 1]$, and

$$\ell_{\log}(p, y) = -\sum_{j=1}^{k} p_j \log p_j + \sum_{j=1}^{k} p_y(1 + \log p_j) - (1 + \log p_y) = -\log p_y,$$

yielding the log loss. Note that for this case, we do require that the gradients $\nabla h(p)$ take values in the (downward) extended reals $\underline{\mathbb{R}}^k$. ◇

Example 11.2.3 (Brier scores and squared error)**:** When we have the squared error $\ell_{\rm sq}(p, y) = \|p - e_y\|_2^2$, we can directly check that $h(p) = \|p\|_2^2$ gives the loss. Indeed,

$$- \|p\|_2^2 - 2\langle p, e_y - p\rangle = \|p\|_2^2 - 2\langle p, e_y\rangle + 1 - 1 = \|p - e_y\|_2^2 - 1.$$

So aside from an additive constant, we have the desired result. $\diamond$

More esoteric examples exist in the literature, such as the spherical score arising from $h(p) = \|p\|_2$ (note the lack of a square).

Example 11.2.4 (Spherical scores)**:** Let $h(p) = \|p\|_2$, which is strictly convex on Δ_k. Then

$$\nabla h(p) = p/\|p\|_2$$

and $\ell(p, y) = -\|p\|_2 - \frac{1}{\|p\|_2}\langle p, e_y - p\rangle = -p_y/\|p\|_2$, which is strictly proper but does not retain convexity. $\diamond$

Bregman divergences

A key aspect of the Savage representation (11.2.1) is that associated to any proper loss is a first-order divergence (or, less evocatively, the *Bregman divergence*). Recall from Chapter 3 that for a function function $h : \mathbb{R}^k \to \overline{\mathbb{R}}$, the first-order divergence associated with h is

$$D_h(u, v) := h(u) - h(v) - \langle \nabla h(v), u - v\rangle. \tag{11.2.2}$$

In typical definitions of the divergence, one requires that h be differentiable; here, we allow non-differentiable h so long as the choice $\nabla h(v) \in \partial h(v)$ is given. In particular, we see that

$$D_h(u, v) \geq 0$$

for all u and v, and moreover, if h is strictly convex

$$D_h(u, v) > 0 \quad \text{whenever} \quad u \neq v.$$

(See, e.g., Corollaries B.3.2 and C.1.7 in the appendices.)

Familiar examples include the squared Euclidean norm $h(u) = \frac{1}{2}\|u\|_2^2$, which by inspection gives

$$D_h(u, v) = \frac{1}{2}\|u - v\|_2^2,$$

and the negative entropies $h(u) = \sum_{j=1}^k u_j \log u_j$, which implicitly encodes the constraint that $u \succ 0$. This gives

$$D_h(u, v) = \sum_{j=1}^k u_j \log u_j - \sum_{j=1}^k v_j \log v_j - \sum_{j=1}^k (1 + \log v_j)(u_j - v_j) = \sum_{j=1}^k u_j \log \frac{u_j}{v_j} + \mathbf{1}^T(u - v).$$

If $u, v \in \Delta_k$, then evidently $D_h(u, v) = D_{\rm kl}(u\|v)$ because $\mathbf{1}^T u = \mathbf{1}^T v = 1$, where we identify u and v with probability mass functions.

Continuing this identification of distributions on $\mathcal{Y}$ with elements $p \in \Delta_k$ in the probability simplex, we can reconsider the gaps between a loss evaluated at a true distribution p and an alternative q. In this case, the representation Theorem 11.2.1 provides allows us to connect proper

losses with first-order divergences immediately. Indeed, let $h : \Delta_k \to \overline{\mathbb{R}}$ be a convex function and loss ℓ be the associated proper loss, with $\ell(p, y) = -h(p) - \langle \nabla h(p), e_y - p \rangle$. Now, suppose that Y has p.m.f. p; then for any $q \in \Delta_k$, the gap

$$\begin{aligned}\mathbb{E}_p[\ell(q, Y)] - \mathbb{E}_p[\ell(p, Y)] &= h(p) - h(q) - \sum_{y=1}^{k} p_y \langle \nabla h(q), e_y - q \rangle \\ &= h(p) - h(q) - \langle \nabla h(q), p - q \rangle = D_h(p, q).\end{aligned}$$

We record this as a corollary to Theorem 11.2.1, highlighting the links between propriety, first-order divergences, and proper loss functions.

Corollary 11.2.5. *Let the conditions of Theorem 11.2.1 hold. Then ℓ is (strictly) proper if and only if there exists a (strictly) convex $h : \Delta_k \to \mathbb{R}$ for which*

$$\mathbb{E}_p[\ell(q, Y)] - \mathbb{E}_p[\ell(p, Y)] = D_h(p, q)$$

for all $p, q \in \Delta_k$.

11.2.2 General proper losses

More generally, we can consider predicting distributions P on general sets $\mathcal{Y}$. For example, recalling the meteorological motivation of predicting the weather, suppose we wish to predict a distribution of the (real-valued) amount Y of rainfall on a given day. Many predictions place a point mass at $Y = 0$, with a decaying tail for higher amounts of rainfall. Then it is natural to predict a cumulative distribution function $F : \mathbb{R} \to [0, 1]$, measuring error relative to the actual amount of rain that falls. Several losses are common in the literature; one common example is the *continuous ranked probability score*.

Example 11.2.6 (Continuous ranked probability score (CRPS))**:** The CRPS loss for a CDF F at y is

$$\ell_{\text{crps}}(F, y) = \int (F(t) - \mathbf{1}\{y \le t\})^2 dt. \tag{11.2.3}$$

This is a strictly proper scoring rule: let G be any cumulative distribution function, meaning that $\lim_{t \to -\infty} G(t) = 0$ and $\lim_{t \to \infty} G(t) = 1$, and let Y have CDF F. Then

$$\begin{aligned}\mathbb{E}[\ell_{\text{crps}}(G, Y)] - \mathbb{E}[\ell_{\text{crps}}(F, Y)] &= \int \left(G(t)^2 - F(t)^2 - 2(G(t) - F(t))\mathbb{E}[\mathbf{1}\{Y \le t\}]\right) dt \\ &= \int (G(t) - F(t))^2 \, dt\end{aligned}$$

because $\mathbb{E}[\mathbf{1}\{Y \le t\}] = F(t)$. This is the (squared) Cramér-von-Mises distance between F and G, and which is positive unless $F = G$. Unfortunately, computing the CRPS loss (11.2.3) is often challenging except for specially structured F. $\diamond$

Because the computation of the continuous ranked probability score is challenging, it can be advantageous to consider other losses on probability distributions, which can allow more flexibility in modeling. To that end, we define the *quantile loss*: for a probability distribution P on Y, let

$$\mathsf{Quant}_\alpha(Y) = \mathsf{Quant}_\alpha(P) := \inf\{t \mid P(Y \le t) \ge \alpha\}$$

to be the α-quantile of the distribution P. (When Y has cumulative distribution F, this is the inverse CDF mapping $F^{-1}(\alpha) = \inf\{t \mid F(t) \ge \alpha\}$.) Defining the quantile penalty

$$\rho_\alpha(t) = \alpha \left[t\right]_+ + (1-\alpha)\left[-t\right]_+,$$

for a collection $\mathcal{A}$ of values in $[0,1]$, the *quantile loss* is

$$\ell_{\text{quant},\mathcal{A}}(P, y) := \sum_{\alpha \in \mathcal{A}} \rho_\alpha(y - \mathsf{Quant}_\alpha(P)). \tag{11.2.4}$$

The propriety of the quantile loss is relatively straightforward; it is, however, not strictly proper.

Example 11.2.7 (Quantile loss)**:** To see that the quantile loss (11.2.4) is proper, consider the single quantile penalty ρ_α: let $g(t) = \mathbb{E}[\rho_\alpha(Y-t)] = \alpha\mathbb{E}[[Y-t]_+] + (1-\alpha)\mathbb{E}[[t-Y]_+]$, which we claim is minimized by $\mathsf{Quant}_\alpha(Y)$. Indeed, g is convex, and it has left and right derivatives

$$\partial_- g(t) := \lim_{s \uparrow t} \frac{g(s) - g(t)}{s - t} = -\alpha\mathbb{P}(Y \ge t) + (1-\alpha)\mathbb{P}(Y < t) = \mathbb{P}(Y < t) - \alpha \quad \text{and}$$
$$\partial_+ g(t) := \lim_{s \downarrow t} \frac{g(s) - g(t)}{s - t} = -\alpha\mathbb{P}(Y > t) + (1-\alpha)\mathbb{P}(Y \le t) = \mathbb{P}(Y \le t) - \alpha.$$

Indeed, for $t = \mathsf{Quant}_\alpha(Y)$, we have $\partial_- g(t) = \mathbb{P}(Y < t) - \alpha \le 0$ and $\partial_+ g(t) = \mathbb{P}(Y \le t) - \alpha \ge 0$, because $t \mapsto \mathbb{P}(Y \le t)$ is right continuous. So convexity yields

$$\mathbb{E}[\rho_\alpha(Y - \mathsf{Quant}_\alpha(Y))] \le \mathbb{E}[\rho_\alpha(Y - t)]$$

for all t. Applying this argument for each $\alpha \in \mathcal{A}$, we thus have

$$\mathbb{E}[\ell_{\text{quant},\mathcal{A}}(Q, Y)] \ge \mathbb{E}[\ell_{\text{quant},\mathcal{A}}(P, Y)]$$

for any Q whenever $Y \sim P$, and equality holds whenever Q and P have identical α quantile for each $\alpha \in \mathcal{A}$. $\diamond$

The general case of Theorem 11.2.1 allows us to address such scenarios, though it does require measure theory to properly define. Happily, the generality does not require a particularly more sophisticated proof. For a (convex) function $h : \mathcal{P} \to \overline{\mathbb{R}}$ on a family of distributions $\mathcal{P}$ on a set $\mathcal{Y}$, we say $h'(P; \cdot) : \mathcal{Y} \to \underline{\mathbb{R}}$ is a subderivative of h at $P \in \mathcal{P}$ whenever

$$\begin{aligned} h(Q) &\ge h(P) + \int_{\mathcal{Y}} h'(P, y)(dQ(y) - dP(y)) \\ &= h(P) + \mathbb{E}_Q[h'(P, Y)] - \mathbb{E}_P[h'(P, Y)] \end{aligned} \quad \text{for all } Q \in \mathcal{P}. \tag{11.2.5}$$

When $\mathcal{Y}$ is discrete and we can identify $\mathcal{P}$ with the simplex Δ_k, the inequality (11.2.5) is simply the typical subgradient inequality (11.1.1) that $h(q) \ge h(p) + \langle \nabla h(p), q - p\rangle$ for $p, q \in \Delta_k$, where $\nabla h(p) \in \partial h(p)$. We then have the following generalization of Theorem 11.2.1.

Theorem 11.2.8. *Let $\mathcal{P}$ be a convex collection of distributions on $\mathcal{Y}$. Then the following are true.*

(i) If the loss $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ satisfies the representation

$$\ell(P, y_0) = -h(P) - h'(P, y_0) + \int h'(P, y) dP(y), \quad \textit{for all } y_0 \in \mathcal{Y}, \tag{11.2.6}$$

where $h'(P, \cdot) : \mathcal{Y} \to \underline{\mathbb{R}}$ is a subderivative of h at $P \in \mathcal{P}$, then it is proper.

(ii) Conversely, if ℓ is proper, then choosing h to be the negative generalized entropy $h_\ell(P) = -H_\ell(P) = \sup\{-\mathbb{E}_P[\ell(Q,Y)] \mid Q \in \mathcal{P}\}$ satisfies equality (11.2.6).

The loss is strictly proper if and only if the convex h is strictly convex.

Proof If ℓ has the representation (11.2.6), then we have

$$-\mathbb{E}_P[\ell(P,Y)] = h(P) \geq h(Q) + \int h'(Q,y)(dP(y) - dQ(y)) = -\mathbb{E}_P[\ell(Q,Y)]$$

for any $Q \in \mathcal{P}$ by the definition (11.2.5) of a subderivative. Rewriting, we have $\mathbb{E}_P[\ell(P,Y)] \leq \mathbb{E}_P[\ell(Q,Y)]$ and ℓ is proper.

Conversely, if ℓ is proper and regular, then as in the proof of Theorem 11.2.1 we define

$$h(P) := \sup_{Q \in \mathcal{P}} -\mathbb{E}_P[\ell(Q,Y)] = -\mathbb{E}_P[\ell(P,Y)],$$

which is the supremum of linear functionals of P and hence convex. If we let $h'(P,y) = -\ell(P,y) \in \underline{\mathbb{R}}$ for $P \in \mathcal{P}$, then

$$h(P) \geq -\mathbb{E}_P[\ell(Q,Y)] = h(Q) + \mathbb{E}_Q[\ell(Q,Y)] - \mathbb{E}_P[\ell(Q,Y)] = h(Q) + \int h'(P,y)(dP(y) - dQ(y))$$

by propriety, so that evidently $h'(P,y)$ is a subderivative of h at $P \in \mathcal{P}$. That $L(P,y_0) = -h(P) - h'(P,y_0) + \int h'(P,y)dP(y)$ is then immediate.

The arguments for strict propriety/convexity are similar. □

The obvious corollary to Theorem 11.2.8 follows.

Corollary 11.2.9. *Let $\mathcal{P}$ be a convex collection of probability distributions on $\mathcal{Y}$. Then the loss $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is proper if and only if there exists a convex function $h : \mathcal{P} \to \overline{\mathbb{R}}$ with subderivatives $h'(P,\cdot) : \mathcal{Y} \to \underline{\mathbb{R}}$ such that*

$$\ell(P,y_0) = -h(P) - h'(P,y_0) + \mathbb{E}_P[h'(P,Y)] \quad \text{for all } y_0 \in \mathcal{Y}.$$

The loss ℓ is strictly proper if and only if h is strictly concave.

The subdifferentials and differentiability in this potentially infinite dimensional case can make writing the particular representation (11.2.6) challenging; for example, the representation of the quantile loss in Example 11.2.7 is quite complex. In the case of predictions involving the cumulative distribution function F, however, one can obtain the subderivative by taking directional (Gateaux) derivatives in directions $G - F$ for cumulative distributions G. In this case, for the point cumulative distribution G_y with $G_y(t) = \mathbf{1}\{y \leq t\}$, we define

$$h'(F,y) = \lim_{\epsilon \downarrow 0} \frac{h(F + \epsilon(G_y - F)) - h(F)}{\epsilon}.$$

The continuous ranked probability score (Example 11.2.6) admits this expansion.

Example 11.2.10 (CRPS (Example 11.2.6) continued)**:** The strict propriety of the CRPS loss (11.2.3) means that the generalized entropy

$$h(F) = \sup_G -\mathbb{E}[\ell(G, Y)] = -\mathbb{E}[\ell_{\rm crps}(F, Y)] = \int (F(t) - 1)F(t)dt$$

by definition. Expanding $h(F + \epsilon(G - F))$ for small ϵ as in the recipe above, we have

$$h(F + \epsilon(G - F)) = h(F) - \epsilon \int (G(t) - F(t))dt + 2\epsilon \int F(t)(G(t) - F(t))dt + O(\epsilon^2).$$

to obtain the y-based derivative $h'(F, y)$, we choose $G_y(t) = \mathbf{1}\{y \le t\}$ to obtain directional derivative

$$h'(F, y) = \lim_{\epsilon \downarrow 0} \frac{h(F + \epsilon(G_y - F)) - h(F)}{\epsilon} = \int (\mathbf{1}\{y \le t\} - F(t))dt - 2\int (F(t)(\mathbf{1}\{y \le t\} - F(t))dt.$$

By inspection, when Y has cumulative distribution function F, $\mathbb{E}[h'(F, Y)] = 0$ and so

$$\begin{aligned}
&-h(F) - h'(F, y) + \mathbb{E}[h'(F, Y)] \\
&= \int \left(-F(t)^2 + F(t) - F(t) + \mathbf{1}\{y \le t\} + 2F(t)^2 - 2F(t)\mathbf{1}\{y \le t\}\right) dt \\
&= -\int \left(F(t) - \mathbf{1}\{y \le t\}\right)^2 dt = \ell_{\rm crps}(F, y),
\end{aligned}$$

as desired. $\diamond$

11.2.3 Proper losses and vector-valued Y

The final variant of propriety we consider generalizes that when $\mathcal{Y}$ is finite and identified with $\{1, \ldots, k\}$ in Section 11.2.1. Now, we assume that Y is vector-valued, with $\mathcal{Y} \subset \mathbb{R}^k$, and assume the convex hull

$$\mathrm{Conv}(\mathcal{Y}) = \{\mathbb{E}_P[Y] \mid P \text{ is a distribution on } \mathcal{Y}\}$$

is bounded. (Typically, it will also be compact, though this will not be central to our development, and pathological cases, such as $\mathcal{Y} = \{1/n\}_{n \in \mathbb{N}}$, exist.) An example showing how to use this representation for multinomial $Y \in \{1, \ldots, k\}$ may be clarifying.

Example 11.2.11 (Multinomial Y as vectors)**:** If Y is a multinomial taking values in a discrete set of size k, we can instead identify Y with the first k standard basis vectors $e_1, \ldots, e_k$. Then $p = \mathbb{E}[Y] \in \Delta_k$ is the p.m.f. of Y, and $\mathrm{Conv}(\mathcal{Y}) = \Delta_k$. $\diamond$

Example 11.2.12 (Binary Y as a scalar)**:** When $Y \in \{0, 1\}$ is a Bernoulli random variable, we identify Y with itself, so that $p = \mathbb{E}[Y] = \mathbb{P}(Y = 1) \in [0, 1]$ and $\mathrm{Conv}(\mathcal{Y}) = [0, 1]$. $\diamond$

Example 11.2.13 (Ordinal Y as a scalar)**:** Consider a rating problem of predicting the rating Y of a movie from 1 to 5 stars. In this case, Y takes values $\{1, \ldots, 5\} \subset \mathbb{R}$, but the ordering between the elements is important; it is unnatural to treat Y as a multinomial. More generally, Y may take values in $\{y_1, \ldots, y_k\} \subset \mathbb{R}$, where $y_1 < \cdots < y_k$. As in the binary case, we identify Y with its scalar value, so that $\mathbb{E}[Y] \in [y_1, y_k]$ and $\mathrm{Conv}(\mathcal{Y}) = [y_1, y_k]$. $\diamond$

In this vector-valued Y case, instead of prediction distributions P, the goal is to predict the *mean mapping*

$$\mu(P) := \mathbb{E}_P[Y] \in \mathrm{Conv}(\mathcal{Y}),$$

so that $\mu : \mathcal{P} \to \mathbb{R}^k$ for the collection $\mathcal{P}$ of distributions on Y. Our goal is to reward predictions of the correct expectation, leading to the following definition.

Definition 11.3. *Let $C = \mathrm{cl}\,\mathrm{Conv}(\mathcal{Y})$ be a convex set. Then $\ell : C \times \mathcal{Y} \to \overline{\mathbb{R}}$ is* proper *if*

$$\mathbb{E}_P[\ell(\mu, Y)] \geq \mathbb{E}_P[\ell(\mathbb{E}_P[Y], Y)] \quad \text{for all } \mu \in C,$$

and strictly proper *if the inequality is strict whenever $\mu \neq \mathbb{E}_P[Y]$.*

Definition 11.3 generalizes Definition 11.2 in the multinomial case, where $\mathcal{Y}$ is a discrete set that we may identify with the basis vectors $\{e_1, \ldots, e_k\}$, as Example 11.2.11 makes clear.

With this definition, we can extend Theorem 11.2.1 to a more general case, where as usual we say that ℓ is regular if $\mathbb{E}_P[\ell(\mathbb{E}_P[Y], Y)] < \infty$ for all distributions P on $\mathcal{Y}$.

Theorem 11.2.14. *Let $\mathcal{Y} \subset \mathbb{R}^k$ be finite, $\mathcal{P}$ be the collection of distributions on $\mathcal{Y}$, and $C = \mathrm{Conv}(\mathcal{Y}) = \{\mathbb{E}_P[Y] \mid P \in \mathcal{P}\}$. A regular loss $\ell : C \times \mathcal{Y} \to \overline{\mathbb{R}}$ is proper if and only if there exists a closed convex $h : C \to \mathbb{R}$ such that*

$$\ell(\mu, y) = -h(\mu) - \langle \nabla h(\mu), y - \mu \rangle$$

for some subgradient $\nabla h(\mu) \in \partial h(\mu) \subset \underline{\mathbb{R}}^k$. Additionally, if $\ell : C \times \mathcal{Y} \to \mathbb{R}$, then $\partial h(\mu) \subset \mathbb{R}^k$, and if $\mu \in \mathrm{relint}\, C$, we have $\partial h(\mu) \subset \mathbb{R}^k$. The loss is strictly proper if and only if the associated h is strictly convex.

With this theorem, we have an essentially complete analogy with Theorem 11.2.1. There are subtleties in the proof because the mapping from probabilities P to $\mathbb{E}_P[Y]$ can be many-to-one, necessitating some care in the calculations, and making infinite losses somewhat challenging. A few examples centered around ordinal regression illustrate the scenarios.

Example 11.2.15 (Ordinal regression, Example 11.2.13 continued)**:** Let $Y \in \{0, 1, \ldots, k\}$ be a value to be predicted, where the ordering on Y is important, as in ratings of items. In this case, the set $C = \mathrm{Conv}(\mathcal{Y}) = [0, k]$, and any strictly convex loss with domain $[0, k]$ gives rise to a proper loss via the construction $\ell_h(\mu, y) = -h(\mu) - h'(\mu)(y - \mu)$. First, we take $h(\mu) = \frac{1}{2}\mu^2$. This gives rise to a (modified) squared error

$$\ell_h(\mu, y) = \frac{1}{2}(\mu - y)^2 - \frac{1}{2}y^2,$$

which is strictly convex and proper.

Other choices of h are possible. One natural choice is a variant of the negative binary entropy, and we define

$$h(\mu) = (k - \mu) \log(k - \mu) + \mu \log \mu,$$

which is convex in $\mu \in [0, k]$, with $h(\mu) = +\infty$ for $\mu > k$ or $\mu < 0$, while $h(0) = h(k) = k \log k$. We have $h'(\mu) = \log \frac{\mu}{k - \mu}$, and so

$$\ell_h(\mu, y) = -y \log \mu + (y - k) \log(k - \mu),$$

for $y \in \{0, \ldots, k\}$. Here, however, note the importance of allowing infinite values in the loss ℓ when $\mu \to \{0, k\}$. $\diamond$

Proof One direction is, as in the previous cases, straightforward. Let ℓ have the given representation. Then for $\mu(P) = \mathbb{E}_P[Y]$,

$$\mathbb{E}_P[\ell(\mu, Y)] = -h(\mu) - \langle \nabla h(\mu), \mu(P) - \mu \rangle \geq -h(\mu(P)) = \mathbb{E}_P[\ell(\mu(P), Y)],$$

and the inequality is strict if h is strictly convex.

The converse direction (from a proper loss to function h) is more subtle. We first give the argument in the case that the losses ℓ are finite-valued, so that $\ell(\mu, y) < \infty$ for each $\mu \in C$ and $y \in \mathcal{Y}$, deferring the proof of the general case to Section 11.5.1 as it yields little additional intuition. Let $\mathcal{Y} = \{y_1, \ldots, y_m\} \subset \mathbb{R}^k$, and assume w.l.o.g. that that the matrix $A = [y_1 \cdots y_m]$ with columns y_j has rank k (otherwise, we simply work in a subspace). We may identify $\mathcal{P}$ with the probability simplex Δ_m, and then the mean mapping $\mu(p) = \sum_{i=1}^m p_i y_i$ for $p \in \mathbb{R}^m$ is surjective. Now for $\mu \in \mathbb{R}^k$ define

$$h(\mu) := \inf_{p : \mu(p) = \mu} \sup_{\alpha} \left\{-\mathbb{E}_p[\ell(\alpha, Y)]\right\} \stackrel{(\star)}{=} \inf_{p \in \Delta_m} \left\{-\mathbb{E}_p[\ell(\mu, Y)] \mid \mu(p) = \mu\right\},$$

where the equality $(\star)$ follows because ℓ is proper. The function h is closed convex, as it is the partial infimum of the closed convex function $p \mapsto -\mathbb{E}_p[\ell(\mu, Y)] + \mathbf{I}_{\Delta_m}(p)$, where we recall $\mathbf{I}_{\Delta_m}(p) = 0$ if $p \in \Delta_m$ and $+\infty$ otherwise (see Proposition B.3.11).

We compute $\partial h(\mu)$ directly now. The infimum over p in the definition of $h(\mu)$ is attained, as Δ_m is compact and $g(p) := -\mathbb{E}_p[\ell(\mu, Y)]$ is necessarily continuous in p satisfying $\mu(p) = \mu$, because regularity of the loss guarantees $\ell(\mu, y_i) \in \mathbb{R}$ whenever $p_i > 0$ is feasible in the mean mapping constraint $\mu(p) = \mu$. Moreover, it is immediate that

$$\nabla g(p) = \begin{bmatrix} -\ell(\mu, y_1) \\ \vdots \\ -\ell(\mu, y_m) \end{bmatrix} \in \mathbb{R}^m.$$

Let $p^\star(\mu)$ be any p attaining the infimum. By Proposition B.3.27 on the subgradients of partial minimization, we thus obtain

$$\partial h(\mu) = \left\{s \in \mathbb{R}^k \mid y_i^T s = -\ell(\mu, y_i) \text{ for } i = 1, \ldots, m\right\},$$

and moreover, this set is necessarily non-empty for all $\mu \in \operatorname{relint} C = \{\mu(p) \mid p \succ 0, p \in \Delta_m\}$. Using this equality, we have

$$\begin{aligned} \ell(\mu, y) = -h(\mu) + h(\mu) + \ell(\mu, y) &= -h(\mu) + \mathbb{E}_{p^\star(\mu)}[-\ell(\mu, Y)] + \ell(\mu, y) \\ &= -h(\mu) + \sum_{i=1}^m p_i^\star(\mu) y_i^T s - y^T s \\ &= -h(\mu) + \langle s, \mathbb{E}_{p^\star(\mu)}[Y] - y \rangle = -h(\mu) - \langle s, y - \mu \rangle \end{aligned}$$

for any $s \in \partial h(\mu)$, as $\mathbb{E}_p[Y] = \mu(p) = \mu$ by construction.

Lastly, to obtain strict convexity of h, note that if $\mathbb{E}_p[Y] = \mu$, then we can use the representation

$$\mathbb{E}_p[\ell(\mu', Y)] - \mathbb{E}_p[\ell(\mu, Y)] = -h(\mu') - \langle \nabla h(\mu'), \mu - \mu' \rangle + h(\mu) = D_h(\mu, \mu')$$

which is positive whenever $\mu \neq \mu'$ if and only if h is strictly convex. □

11.3 From entropies to convex losses, arbitrary predictions, and link functions

Frequently, when we fit models, it is inconvenient to directly model or predict probabilities, that is, to minimize over probabilistic predictions. Instead, we often wish to fit some real-valued prediction and then transform it into a probabilistic prediction. This is perhaps most familiar from binary and multiclass logistic regression, where a *link function* transforms real-valued predictions into probabilistic predictions. For the binary logistic regression case with $Y \in \{-1, 1\}$, we assume that we predict a score $s \in \mathbb{R}$, where $s > 0$ indicates a prediction that Y is more likely to be 1 and $s < 0$ that it is more likely negative. The implied (modelled) probability that $Y = y$ is then

$$p(y \mid s) = \frac{1}{1 + \exp(-ys)} \text{ for } y \in \{-1, 1\}.$$

Similarly, for k-class classification problems, when using multiclass logistic regression, we predict a score vector $s \in \mathbb{R}^k$, where s_y indicates a score associated to one of the k potential class labels y; this then implies the probabilites

$$p(y \mid s) = \frac{\exp(s_y)}{\sum_{i=1}^k \exp(s_i)} = \frac{1}{1 + \sum_{i \neq y} \exp(s_i - s_y)},$$

where we clearly have $\sum_y p(y \mid s) = 1$.

In binary and logistic regression, instead of directly minimizing negative log probabilities of error over the probability simplex (though one does this implicitly), instead we use surrogate logistic losses whose arguments can range over all of $\mathbb{R}$ or $\mathbb{R}^k$. In the case of binary logistic regression with $y \in \{-1, 1\}$, this is

$$\varphi(s, y) = \log(1 + \exp(-sy)),$$

while in the multiclass case we use the multiclass logistic loss

$$\varphi(s, y) = -s_y + \log\left(\sum_{i=1}^k \exp(s_i))\right) = \log\left(1 + \sum_{i \neq y} \exp(s_i - s_y)\right).$$

Note that for each of these, we have a direct relationship between the probabilistic predictions and derivatives of φ. In the binary logistic regression case, we have

$$p(y \mid s) = 1 + \frac{\partial}{\partial s}\varphi(s, y) = 1 - \frac{1}{1 + \exp(ys)} = \frac{1}{1 + \exp(-ys)},$$

while in the multiclass case we similarly have

$$p(y \mid s) = 1 + \frac{\partial}{\partial s_y}\varphi(s, y) = \frac{\exp(s_y)}{\sum_{i=1}^k \exp(s_i)}.$$

11.3.1 Convex conjugate linkages

These dualities turn out to hold in substantially more generality, and they are the key to transforming proper losses (as applied on probabilities) into proper surrogate losses that apply directly to real-valued scores and which are convex in their arguments, allowing us to bring the tools of convex optimization to bear on actually fitting predictive models. We work in the general setting of

Section 11.2.3 of losses for vector-valued y where $\mathcal{Y} \subset \mathbb{R}^k$, so that instead of predicting probability distributions on Y itself we predict elements μ of the set $\{\mathbb{E}_P[Y]\} = \mathrm{Conv}(\mathcal{Y})$, and let ℓ be a strictly proper loss. Theorems 11.2.1 and 11.2.14 demonstrate that if the loss ℓ is proper, there exists a (negative) generalized entropy, which in the case of Theorem 11.2.1 is $h(p) = \sup_q\{-\mathbb{E}_p[\ell(q, Y)]\}$, for which

$$\ell(\mu, y) = -h(\mu) - \langle \nabla h(\mu), y - \mu\rangle.$$

Note that h is always a *closed* convex function, meaning that it is lower semicontinuous or that its epigraph $\mathrm{epi}\, h = \{(\mu, t) \mid h(\mu) \le t\}$ is closed.

Let us suppose temporarily that we have *any* such entropy. Recalling the convex conjugate (11.1.3), the negative generalized entropy h is closed convex, and so its conjugate $h^*(s) = \sup\{\langle s, \mu\rangle - h(\mu)\}$ satisfies $h^{**}(\mu) = h(\mu)$. In particular, if we define the *surrogate loss*

$$\varphi(s, y) := h^*(s) - \langle s, y\rangle,$$

which is defined for all $s \in \mathbb{R}^k$ (instead of $\mathrm{Conv}(\mathcal{Y})$), then

$$\mathbb{E}_P[\varphi(s, Y)] = h^*(s) - \langle s, \mathbb{E}_P[Y]\rangle = h^*(s) - \langle s, \mu(P)\rangle$$

for the mean mapping $\mu(P) = \mathbb{E}_P[Y]$. Moreover,

$$\inf_s \mathbb{E}_P[\varphi(s, Y)] = \inf_s \{h^*(s) - \langle s, \mu(P)\rangle\} = -h^{**}(\mu(P)) = -h(\mu(P)),$$

and so it generates the same negative entropy as the original loss ℓ, as

$$\inf_\mu \mathbb{E}_P[\ell(\mu, Y)] = \inf_\mu\{-h(\mu) - \langle \nabla h(\mu), \mu(P) - \mu\rangle\} = -h(\mu(P)).$$

This identification of (generalized) entropies will underpin much of our development of the consistency of losses in sections to come. For now, we content ourselves with addressing how to understand propriety of the surrogate loss φ and how to transform predictions $s \in \mathbb{R}^k$ into probabilistic predictions μ.

The key will be to consider what we term *convex-conjugate-linkages*, or *conjugate linkages* for short. Recall the duality relationships (11.1.4) from the Fenchel-Young inequality we present in the convexity primer in Section 11.1.1. The negative generalized entropy h is convex, and the dualities associated with its conjugate $h^*(s) = \sup_\mu\{\langle s, \mu\rangle - h(\mu)\}$ will form the basis of our transformations. We first give a somewhat heuristic presentation, as the intuition is important (but details to make things precise can be a bit tedious). Essentially, we require that h^* and h are continuously differentiable, in which case we have

$$\nabla h(\mu) = s \;\text{ if and only if }\; \nabla h^*(s) = \mu \;\text{ if and only if }\; h^*(s) + h(\mu) = \langle s, \mu\rangle$$

by the Fenchel-Young inequalities (11.1.4). That is, the gradient ∇h^* of the conjugate transforms a score vector $s \in \mathbb{R}^k$ into elements c to predict $\mathcal{Y}$: we transform s into a prediction μ via the *conjugate link* function

$$\mathsf{pred}_h(s) = \operatorname*{argmax}_\mu \{\langle s, \mu\rangle - h(\mu)\} = \nabla h^*(s) = (\nabla h)^{-1}(s), \tag{11.3.1}$$

which finds the μ that best trades having maximal "entropy" $-h(\mu)$, or uncertainty, with alignment with the scores $\langle s, \mu\rangle$.

With this, it is then natural to consider the function substituting the prediction $\mu = \mathsf{pred}_h(s)$ into $\ell(\mu, y)$, and so we consider

$$\ell(\mathsf{pred}_h(s), y).$$

Immediately, if $\mu = \mathsf{pred}_h(s) = \nabla h^*(s)$, we have $s = \nabla h(\mu)$ by construction (or the Fenchel-Young inequality (11.1.4)), and so $h(\mu) = \langle s, \mu\rangle - h^*(s)$ for this particular pair (s, μ), and $\nabla h(\mu) = \nabla h(\nabla h^*(s)) = s$ because ∇h and ∇h^* are inverses. Substituting, we obtain

$$\begin{aligned}\ell(\mathsf{pred}_h(s), y) = -h(\mathsf{pred}_h(s)) - \langle \nabla h(\mathsf{pred}_h(s)), y - \mathsf{pred}_h(s)\rangle &= -h(\mu) - \langle s, y - \mu\rangle \\ &= h^*(s) - \langle s, \mu\rangle - \langle s, y - \mu\rangle,\end{aligned}$$

that is, we have recovered the surrogate

$$\varphi(s, y) = h^*(s) - \langle s, y\rangle. \tag{11.3.2}$$

The surrogate loss (11.3.2) constructed from the negative entropy h is the key transformation of the loss ℓ into a convex loss, and (no matter the properties of ℓ) is always convex.

As we have already demonstrated, the construction (11.3.2) is more general than we have presented; certainly, h^* is always convex, and so φ is always convex in s. Moreover, if Y has expectation $\mathbb{E}[Y] = \mu$, then

$$\inf_s \mathbb{E}[\varphi(s, \mu)] = \inf_s \{h^*(s) - \langle s, \mu\rangle\} = -h(\mu)$$

by conjugate duality, so the surrogate φ always recovers the negative entropy h; without some type of differentiability conditions, however, the construction of the prediction mapping pred_h requires more care. Chapter 14 more deeply investigates these connections.

All that remains is to give more precise conditions under which the prediction (11.3.1) is always unique and exists for all possible score vectors $s \in \mathbb{R}^k$. To that end, we make the following definition.

Definition 11.4. *Let $h : \mathbb{R}^k \to \overline{\mathbb{R}}$. Then h is a* Legendre negative entropy *if it is strictly convex, continuously differentiable, and*

$$\|\nabla h(\mu)\| \to \infty \quad \text{if either} \begin{cases} \mu \to \mathrm{bd}\,\mathrm{dom}\, h & \text{or} \\ \|\mu\| \to \infty. \end{cases} \tag{11.3.3}$$

This is precisely the condition we require to make each step in the development of the surrogate (11.3.2) airtight; as a corollary to Theorem C.2.9 in the appendices, we have the following.

Corollary 11.3.1. *Let h be a Legendre negative entropy. Then the conjugate link prediction (11.3.1) is unique and exists for all $s \in \mathbb{R}^k$. In particular, the conjugate h^* is strictly convex, continuously differentiable, satisfies $\mathrm{dom}\, h^* = \mathbb{R}^k$, and $\nabla h^* = (\nabla h)^{-1}$.*

With this corollary in place, we can then give a theorem showing the equivalence of the strictly proper loss ℓ and its surrogate.

Theorem 11.3.2. *Let $\ell : C \times \mathcal{Y} \to \overline{\mathbb{R}}$ be the strictly proper loss associated with the Legendre negative entropy h. Then*

$$\ell(\mathsf{pred}_h(s), y) = \varphi(s, y) := h^*(s) - \langle s, y\rangle.$$

Moreover, the convex surrogate φ satisfies the consistency that if

$$\mathbb{E}[\varphi(s_n, Y)] \to \inf_s \mathbb{E}[\varphi(s, Y)]$$

then $\mu_n = \mathsf{pred}_h(s_n)$ satisfies

$$\mathbb{E}[\ell(\mu_n, Y)] \to \inf_\mu \mathbb{E}[\ell(\mu, Y)].$$

Proof The first equality we have already demonstrated. For the minimization claim, we note that if $\mu = \mathbb{E}[Y]$, then $\mathbb{E}[\varphi(s, Y)] = h^*(s) - \langle \mu, s\rangle$ and $\inf_s\{h^*(s) - \langle \mu, s\rangle\} = -h(\mu)$. Strict propriety of ℓ then gives $\inf_{\mu'} \mathbb{E}[\ell(\mu', Y)] = -h(\mu)$. □

Said differently, the surrogate φ is consistent with the loss ℓ and (strictly) proper, in that if s minimizes $\mathbb{E}[\varphi(s, Y)]$, then $\mathsf{pred}_h(s)$ minimizes $\mathbb{E}[\ell(\mu, Y)]$. The statement in terms of limits is necessary, however, as simple examples show, because with some link functions it is in fact impossible to achieve the extreme points of $\mathrm{Conv}(\mathcal{Y})$, as in logistic regression. We provide a few example applications (and non-applications) of Theorem 11.3.2. For the first, let us consider binary logistic regression.

Example 11.3.3 (Binary logistic regression)**:** For a label $Y \in \{0, 1\}$ and predictions $p \in [0, 1]$, take the generalized entropy

$$h(p) = p \log p + (1 - p) \log(1 - p).$$

By inspection, $\mathrm{dom}\, h = [0, 1]$, and $h'(p) = \log \frac{p}{1-p}$ satisfies $|h'(p)| \to \infty$ as $p \to \{0, 1\}$. For $s \in \mathbb{R}$, the conjugate is

$$h^*(s) = \sup_p\{sp - p \log p - (1 - p) \log(1 - p)\} = \log(1 + e^s),$$

where the supremum is achieved by $p = \mathsf{pred}_h(s) = \frac{e^s}{1+e^s}$. Then we have

$$\varphi(s, y) = \log(1 + e^s) - sy = -\log p(y \mid s),$$

where $p(y \mid s) = \frac{e^{ys}}{1+e^s}$ is the binary logistic probability of the label $y \in \{0, 1\}$.

For the induced loss $\ell(p, y) = -y \log p - (1 - y) \log(1 - p)$ (the log loss), if $\mathbb{P}(Y = 1) = 1$, then $p = 1$ minimizes $\mathbb{E}[\ell(p, Y)]$. Similarly, if $\mathbb{P}(Y = 0) = 1$, then $p = 0$ minimizes $\mathbb{E}[\ell(p, Y)]$. Neither of these is achievable by a finite $\dot{\ell}$ in $p(y \mid s) = \frac{e^{ys}}{1+e^s}$, showing how the limiting argument in Theorem 11.3.2 is necessary. ◇

The next example shows that we sometimes need to elaborate the setting of Theorem 11.3.2 to deal with constraints.

Example 11.3.4 (Multiclass logistic regression)**:** Identify the set $\mathcal{Y} = \{e_1, \ldots, e_k\}$ with the k standard basis vectors, and for $p \in \Delta_k = \{p \in \mathbb{R}^k_+ \mid \mathbf{1}^T p = 1\}$, consider the negative entropy

$$h(p) = \sum_{y=1}^k p_y \log p_y.$$

This function is strictly convex and of Legendre type for the positive orthant $\mathbb{R}^k_+$ but *not* for Δ_k. Shortly, we shall allow linear constraints on the predictions to address this shortcoming. As an alternative, take $\mathcal{Y} = \{0, e_1, \ldots, e_{k-1}\}$, so that $\mathrm{Conv}(\mathcal{Y}) = \{p \in \mathbb{R}^{k-1}_+ \mid \mathbf{1}^T p \le 1\}$, which has an interior and so more easily admits a conjugate duality relationship. In this case, the negative entropy-type function

$$h(p) = \sum_{y=1}^{k-1} p_y \log p_y + (1 - \mathbf{1}^T p) \log(1 - \mathbf{1}^T p) \tag{11.3.4}$$

is of Legendre type. A calculation for $s \in \mathbb{R}^{k-1}$ yields

$$h^*(s) = \log \left(1 + \sum_{y=1}^{k-1} e^{s_y}\right),$$

with

$$\mathsf{pred}_h(s) = \left(\frac{e^{s_1}}{1 + \sum_{j=1}^{k-1} e^{s_j}}, \ldots, \frac{e^{s_{k-1}}}{1 + \sum_{j=1}^{k-1} e^{s_j}}\right).$$

Letting p denote the entries of this vector, we can then assign a probability to class k via $p_k = 1 - \sum_{j=1}^{k-1} p_j$. $\diamond$

In Section 11.4 we revisit exponential families in the (proper) loss minimization framework we have thus far developed, which gives some additional perspective on these problems.

11.3.2 Convex conjugate linkages with affine constraints

As Example 11.3.4 shows, in some cases a "natural" formulation fails to satisfy the desiderata of our link functions. Accordingly, we make a slight modification to the Legendre type (11.3.3) negative entropy h to allow for *affine* constraints, which still allows us to develop the precise convexity dualities with proper losses we require. Continuing to work in the scenario in which $\mathcal{Y} \subset \mathbb{R}^k$, suppose now that the affine hull

$$\mathcal{A} = \mathrm{aff}(\mathcal{Y}) := \left\{\sum_{j=1}^{m} \alpha_j y_j \mid y_j \in \mathcal{Y}, \alpha^T \mathbf{1} = 1, m \in \mathbb{N}\right\}$$

is a proper subspace of $\mathbb{R}^k$. The key motivating example here is the "failure" case of Example 11.3.4 on multiclass logistic regression, where $\mathcal{Y} = \{e_1, \ldots, e_k\}$, whose affine hull is exactly those vectors $p \in \mathbb{R}^k$ satisfying $\langle p, \mathbf{1}\rangle = 1$. Naturally, in this case we wish to predict probabilities, and so given a score vector $s \in \mathbb{R}^k$ and using the negative entropy $h(p) = \sum_{y=1}^k p_y \log p_y$, we let

$$\mathsf{pred}(s) = \operatorname*{argmin}_p \left\{h(p) - \langle s, p\rangle \mid \mathbf{1}^T p = 1\right\} = \left[e^{s_y} / \sum_{j=1}^{k} e^{s_j}\right]_{y=1}^{k}.$$

Generalizing this approach to arbitrary regularizers h, we modify the prediction (11.3.1) to be

$$\mathsf{pred}_{h,\mathcal{A}}(s) = \operatorname*{argmax}_{\mu \in \mathcal{A}} \left\{\langle s, \mu\rangle - h(\mu)\right\}.$$

Then for the loss $\ell(\mu, y) = -h(\mu) - \langle \nabla h(\mu), y - \mu \rangle$ associated with the negative entropy h, we define the surrogate

$$\varphi(s, y) := \ell(\mathsf{pred}_{h,\mathcal{A}}(s), y).$$

Perhaps remarkably, this construction still yields a well-defined convex loss with the same consistency properties as those in Theorem 11.3.2. Indeed, defining

$$h_{\mathcal{A}}(\mu) = h(\mu) + \mathbf{I}_{\mathcal{A}}(\mu)$$

and the associated conjugate $h^*_{\mathcal{A}}(s) = \sup\{\langle s, \mu \rangle - h(\mu) \mid \mu \in \mathcal{A}\}$, we have the following theorem.

Theorem 11.3.5. *Let $\ell : C \times \mathcal{Y} \to \overline{\mathbb{R}}$ be the strictly proper loss associated with the Legendre negative entropy h and $\mathcal{A} = \mathrm{aff}(\mathcal{Y})$ be the affine hull of $\mathcal{Y}$. Then*

$$\varphi(s, y) := \ell(\mathsf{pred}_{h,\mathcal{A}}(s), y) = h^*_{\mathcal{A}}(s) - \langle s, y \rangle.$$

Moreover, the convex surrogate φ satisfies the consistency that if

$$\mathbb{E}[\varphi(s_n, Y)] \to \inf_s \mathbb{E}[\varphi(s, Y)]$$

then $\mu_n = \mathsf{pred}_{h,\mathcal{A}}(s_n)$ satisfies

$$\mathbb{E}[\ell(\mu_n, Y)] \to \inf_{\mu} \mathbb{E}[\ell(\mu, Y)].$$

We return to proving the theorem presently, focusing here on how it applies to Example 11.3.4.

Example 11.3.6 (Multiclass logistic regression)**:** Consider Example 11.3.4, where we identify $\mathcal{Y} = \{e_1, \ldots, e_k\} \subset \mathbb{R}^k$, which has affine hull $\mathcal{A} = \{p \in \mathbb{R}^k \mid \langle \mathbf{1}, p \rangle = 1\}$. Then taking $h(p) = \sum_{y=1}^k p_k \log p_k$, a calculation with a Lagrangian shows that

$$\mathsf{pred}_{h,\mathcal{A}}(s) = \operatorname*{argmin}_{p \in \Delta_k} \{-\langle s, p \rangle + h(p)\} = \left[e^{s_y} / \sum_{j=1}^k e^{s_j} \right].$$

In turn, this gives surrogate logistic loss

$$\varphi(s, y) = \log \left(\sum_{j=1}^k e^{s_j - s_y} \right).$$

Notably, the logistic loss is *not* strictly convex, as $\varphi(s + t\mathbf{1}, y) = \varphi(s, y)$ for $t \in \mathbb{R}$. If Y is a multinomial random variable with $\mathbb{P}(Y = e_y) = p_y$, then by another calculation, the vector with entries

$$s^\star_y = \log p_y$$

minimizes $\mathbb{E}[\varphi(s, Y)]$, which in turn gives $\mathsf{pred}_{h,\mathcal{A}}(s^\star) = p$, maintaining propriety. $\diamond$

Proof of Theorem 11.3.5

Before proving the theorem proper, we show how the key identity that $s = \nabla h(c)$ we use to develop equality (11.3.2) generalizes in the presence of the affine constraint. The function $h_{\mathcal{A}}$ is strictly convex on its domain $\operatorname{dom} h \cap \mathcal{A}$, and moreover, $\nabla h_{\mathcal{A}}^*$ exists and is continuous. The following corollary (a consequence of Corollary C.2.12 in Appendix C.2) extends Corollary 11.3.1 and allows us to address equality (11.3.2).

Corollary 11.3.7. *The conjugate $h_{\mathcal{A}}^*$ is continuously differentiable with $\operatorname{dom} h_{\mathcal{A}}^* = \mathbb{R}^k$, and if $\mu = \nabla h_{\mathcal{A}}^*(s)$, then $\mu \in \operatorname{int} \operatorname{dom} h$ and*

$$\nabla h(\mu) = s + v$$

for some vector v normal to $\mathcal{A}$, that is, a vector $v \in \mathbb{R}^k$ satisfying $\langle v, \mu_0 - \mu_1 \rangle = 0$ for all $\mu_0, \mu_1 \in \mathcal{A}$.

While the proof of the corollary requires some care to make precise, a sketch can give intuition.

Sketch of Proof Because h is strictly convex and its derivatives $\nabla h(\mu)$ explode as $\mu \to \operatorname{bd} \operatorname{dom} h$, the minimizer of $-\langle s, \mu \rangle + h(\mu)$ over $\mu \in \mathcal{A}$ exists and is unique. Let $\mathcal{A} = \{\mu \mid A\mu = b\}$ for shorthand, where $A \in \mathbb{R}^{n \times k}$ for some $n < k$. Then introducing Lagrange multiplier $w \in \mathbb{R}^n$ for the constraint $\mu \in \mathcal{A}$, the Lagrangian for finding $\mathsf{pred}_{h,\mathcal{A}}(s) = \operatorname{argmin}_\mu \{h(\mu) - \langle s, \mu \rangle \mid \mu \in \mathcal{A}\}$ is

$$\mathcal{L}(\mu, w) = h(\mu) - \langle s, \mu \rangle + w^T (A\mu - b).$$

Minimizing out μ by setting $\nabla_\mu \mathcal{L}(\mu, w) = 0$, we obtain

$$\nabla h(\mu) - s + A^T w = 0.$$

But if $\mu_0, \mu_1 \in \mathcal{A}$, then $v = A^T w$ satisfies $\langle v, \mu_0 - \mu_1 \rangle = w^T A(\mu_0 - \mu_1) = w^T (b - b) = 0$, so that v is normal to $\mathcal{A}$. □

Finally, we return to prove the theorem. Take any vector $s \in \mathbb{R}^k$. Then because $\mathsf{pred}_{h,\mathcal{A}}(s) = \nabla h_{\mathcal{A}}^*(s)$, we have

$$\varphi(s, y) = \ell(\mathsf{pred}_{h,\mathcal{A}}(s), y) = -h(\nabla h_{\mathcal{A}}^*(s)) - \langle \nabla h(\nabla h_{\mathcal{A}}^*(s)), y - \nabla h_{\mathcal{A}}^*(s) \rangle.$$

As $\nabla h_{\mathcal{A}}^*(s) \in \mathcal{A}$ and using the shorthand $\mu = \nabla h_{\mathcal{A}}^*(s) \in \mathcal{A}$, we have $\nabla h(\mu) = s + v$ for some v normal to $\mathcal{A}$. Moreover, $h(\mu) = h_{\mathcal{A}}(\mu)$, and so the Fenchel-Young inequality (11.1.4) guarantees $-h_{\mathcal{A}}(\mu) = h_{\mathcal{A}}^*(s) - \langle s, \mu \rangle$. Substituting in the expression for φ, we obtain

$$\begin{aligned} \varphi(s, y) &= h_{\mathcal{A}}^*(s) - \langle s, \mu \rangle - \langle s + v, y - \mu \rangle \\ &= h_{\mathcal{A}}^*(s) - \langle s, \mu \rangle + \langle s, \mu - y \rangle = h_{\mathcal{A}}^*(s) - \langle s, y \rangle \end{aligned}$$

where the second equality follows because $v \perp \mu - y$.

For the consistency argument, let $\mu_n = \mathsf{pred}_{h,\mathcal{A}}(s_n)$. Then $\mathbb{E}[\ell(\mu_n, Y)] = \mathbb{E}[\varphi(s_n, Y)]$ and if $\mu = \mathbb{E}[Y]$, then $\mathbb{E}[\varphi(s, Y)] = h_{\mathcal{A}}^*(s) - \langle \mu, s \rangle$ and $\inf_s \mathbb{E}[\varphi(s, Y)] = -h_{\mathcal{A}}(\mu) = -h(\mu)$. Strict propriety of ℓ gives $\inf_{\mu'} \mathbb{E}[\ell(\mu', Y)] = -h(\mu)$.

11.4 Exponential families, maximum entropy, and log loss

Realistically, making predictions using an arbitrary distribution P on an arbitrary space $\mathcal{X}$ is statistically infeasible: we could never collect enough data to accurately model complex phenomena without any assumptions on P. Accordingly, we may seek more tractable models to make predictions feasible, and we can then investigate the consequences of moving from the entire family $\mathcal{P}$ of distributions on $\mathcal{X}$ to smaller families is. A particularly important class of distributions, which allows us to study these questions in great detail, are the exponential families from Chapter 3; here, we investigate them in the framework that we have developed for proper losses.

Let $\{P_\theta\}$ be a regular exponential family indexed by θ on a space $\mathcal{X}$ with sufficient statistic $\phi : \mathcal{X} \to \mathbb{R}^d$, where for a base measure ν on $\mathcal{X}$, P_θ has density

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$$

with respect to ν, where $A(\theta) = \log \int e^{\langle \theta, \phi(x) \rangle} d\nu(x)$ is the log partition function. (Recall that regularity means that the domain

$$\Theta := \operatorname{dom} A = \{\theta \mid A(\theta) < \infty\}$$

is open, as in Definition 3.1). Consider the log loss $-\log p_\theta(x)$, which we suggestively denote with the surrogate φ as a function of θ,

$$\varphi(\theta, x) := -\log p_\theta(x) = A(\theta) - \langle \theta, \phi(x) \rangle.$$

Proposition 3.2.1 guarantees this is always convex in θ because the log partition function is convex, and it is $\mathcal{C}^\infty$ (Proposition 3.2.2). While the log loss $-\log p(x)$ is proper, the exponential family $\{P_\theta\}$ can capture only a subset of the distributions on $\mathcal{X}$.

The mean mapping $\mu(P) := \mathbb{E}_P[\phi(X)] \in \mathbb{R}^d$ will be of central importance to the development of proper losses, exponential families, and the duality relationships between maximum likelihood and entropy that we explore here. Accordingly, throughout this section we let

$$\mathcal{P} := \{\text{distributions } P \ll \nu\} = \{\text{distributions } P \text{ with a density } p \text{ w.r.t. } \nu\}$$

be the collection of distributions with densities with respect to ν (as P_θ by definition has), and we define the set of potential *mean parameters*

$$\mathcal{M} := \left\{\mu(P) = \mathbb{E}_P[\phi(X)] \in \mathbb{R}^d \mid P \ll \nu\right\} = \left\{\mu(P) \mid P \in \mathcal{P}\right\}. \tag{11.4.1}$$

Now, for any distribution $P \in \mathcal{P}$ with mean vector $\mu = \mu(P)$, the associated generalized negative entropy is

$$h(\mu) := \sup_\theta \left\{-\mathbb{E}_P[\varphi(\theta, X)]\right\} = \sup_\theta \left\{\langle \theta, \mu(P) \rangle - A(\theta)\right\} = A^*(\mu),$$

the convex conjugate of A. At this point, the centrality of the duality relationships (via gradients ∇A and ∇A^*) between Θ and $\mathcal{M}$ to fitting and modeling should come as no surprise, and so we elucidate a few of the main properties. Because $\nabla A(\theta) = \mathbb{E}_\theta[\phi(X)]$ in the exponential family, we immediately see that

$$\nabla A(\Theta) := \{\nabla A(\theta)\}_{\theta \in \Theta} \subset \mathcal{M}.$$

Recalling the duality relationship (11.1.4) that

$$\theta \in \partial A^*(\mu) \text{ if and only if } \nabla A(\theta) = \mu,$$

we can say much more.

Proposition 11.4.1. *Let $\mathcal{M}^o = \operatorname{relint} \mathcal{M}$. Then $\nabla A(\Theta) = \mathcal{M}^o$. Additionally:*

(i) If the family is minimal, then $\mathcal{M}$ has non-empty interior and h is continuously differentiable on $\mathcal{M}^o$, with $\theta = \nabla h(\mu)$ if and only if $\nabla A(\theta) = \mu$.

(ii) If the family is non-minimal, then h is continuously differentiable relative to $\operatorname{aff}(\mathcal{M})$, meaning that there exists a continuous mapping $\nabla h(\mu) \in \Theta$ such that for all $\mu \in \mathcal{M}^o$,

$$\partial h(\mu) = \left\{ \nabla h(\mu) + \operatorname{aff}(\mathcal{M})^\perp \right\}.$$

Moreover, $\Theta = \Theta + \operatorname{aff}(\mathcal{M})^\perp$.

The proof of the proposition relies on the more sophisticated duality theory we develop in Appendices B and C, so we defer it to Section 11.5.2.

We can summarize the proposition by considering minimizers and maximizers: suppose we wish to choose θ to minimize

$$\mathbb{E}_P[\varphi(\theta, X)] = \mathbb{E}_P[-\log p_\theta(X)] = A(\theta) - \langle \mu(P), \theta \rangle.$$

Then so long as the distribution P is not extremal in that $\mu(P) = \mathbb{E}_P[\phi(X)] \in \operatorname{relint} \mathcal{M}$, there exists a parameter $\theta(P)$, unique up to translation in the subspace perpendicular to $\operatorname{aff}(\mathcal{M})$, for which

$$\theta(P) \in \operatorname*{argmin}_\theta \mathbb{E}_P[\varphi(\theta, X)] = \operatorname*{argmin}_\theta \{A(\theta) - \langle \mu(P), \theta \rangle\}.$$

Moreover, this parameter satisfies the mean matching condition

$$\nabla A(\theta(P)) = \mu(P),$$

which is of course sufficient to be a minimizer of the expected log loss. As the statements in the proposition evidence, calculations become more challenging when we must perform them all in an affine subspace, though sometimes this care is unavoidable.

Example 11.4.2 (Gaussian estimation)**:** Assume we fit a distribution assuming X has a Gaussian distribution with mean μ and covariance $\Sigma \succ 0$, both to be estimated. Performing the transformation to the exponential family form with precision $K = \Sigma^{-1}$ and $\theta = \Sigma^{-1}\mu$, we have

$$p_{\theta,K}(x) = \exp\left(\langle \theta, x \rangle - \frac{1}{2} \langle xx^\top, K \rangle - A(\theta, K) \right) \quad \text{for} \quad A(\theta, K) = \frac{1}{2} \theta^\top K^{-1} \theta - \frac{1}{2} \log \det(2\pi K).$$

The log partition function has gradients

$$\nabla_\theta A(\theta, K) = K^{-1}\theta \quad \text{and} \quad \nabla_K A(\theta, K) = -\frac{1}{2} K^{-1} \theta\theta^\top K^{-1} - \frac{1}{2} K^{-1}.$$

Matching moments for a distribution P with second moment matrix $M = \mathbb{E}[XX^\top] \succ 0$ and mean $\mathbb{E}[X]$, we obtain

$$\mathbb{E}[X] = K^{-1}\theta \quad \text{and} \quad M = K^{-1}\theta\theta^\top K^{-1} + K^{-1}.$$

Setting $\theta = K\mathbb{E}[X]$ and noting that $M = \operatorname{Cov}(X) - \mathbb{E}[X]\mathbb{E}[X]^\top$, we solve $M = \mathbb{E}[X]\mathbb{E}[X]^\top + K^{-1}$ by setting $K^{-1} = \operatorname{Cov}(X)$.

When $\mathrm{Cov}(X) \not\succ 0$, the solution $K = \mathrm{Cov}(X)^{-1}$ does not exist, so we must rely instead on part (ii) of Proposition 11.4.1. With some care, one may check that we can work in the subspace spanned by the eigenvectors of $\mathrm{Cov}(X)$, that is, if $\mathrm{Cov}(X) = U\Lambda U^\top$ and $U \in \mathbb{R}^{d\times k}$, the collection of symmetric matrices K whose column space belongs to $\mathrm{span}(U)$. Then the pseudo-inverse $K = \mathrm{Cov}(X)^\dagger$ is the appropriate solution, and it recovers the covariance $\Sigma = K^\dagger = \mathrm{Cov}(X) \succeq 0$. $\diamond$

Finally, let us give a last result that shows the duality relationships between the negative generalized entropy $h(\mu)$ and log partition A, which allows us to also capture a few of the nuances of minimization of the surrogate log loss $\varphi(\theta, x) = -\log p_\theta(x)$ when we encounter distributions P for which the mean mapping $\mu(P)$ is on the boundary of $\mathcal{M}$ or even outside it.

Proposition 11.4.3. *Let $\{P_\theta\}$ be a regular exponential family with log partition $A(\theta)$ with domain Θ, and let $\mathcal{M}$ be the associated mean parameter space with relative interior $\mathcal{M}^{\mathrm{o}} = \operatorname{relint}\mathcal{M}$. Let $h(\mu) = A^*(\mu)$ be the associated negative generalized entropy. Then*

(i) $A(\theta) = h^*(\theta) = A^{**}(\theta)$ for all θ.

(ii) If $\mu \in \mathcal{M}^{\mathrm{o}}$, there exists $\theta(\mu) \in \Theta$ such that the negative entropy satisfies $h(\mu) = A^*(\mu) = \langle\theta(\mu), \mu\rangle - A(\theta(\mu)) < \infty$. If $\mu \notin \operatorname{cl}\mathcal{M}$, then $h(\mu) = +\infty$.

(iii) If $\mu \in \operatorname{bd}\mathcal{M} = \operatorname{cl}\mathcal{M} \setminus \mathcal{M}^{\mathrm{o}}$, then for any $\mu_0 \in \mathcal{M}^{\mathrm{o}}$, $h(\mu) = \lim_{t\to 0} h(t\mu_0 + (1-t)\mu)$, and there exist $\theta_t \in \Theta$ with

$$\nabla A(\theta_t) = t\mu_0 + (1-t)\mu \quad \text{and} \quad \lim_{t\to 0}\{A(\theta_t) - \langle\mu, \theta_t\rangle\} = \inf_\theta \{A(\theta) - \langle\mu, \theta\rangle\}.$$

In particular, there exist sequences of dual pairs (μ_n, θ_n) with $\mu_n \in \mathcal{M}^{\mathrm{o}}$ and $\theta_n \in \Theta$ satisfying $\mu_n = \nabla A(\theta_n)$, $\mu_n \to \mu$, $h(\mu_n) \to h(\mu)$, and $A(\theta_n) - \langle\mu, \theta_n\rangle \to \inf_\theta\{A(\theta) - \langle\mu, \theta\rangle\}$.

See Section 11.5.2 for the deferred proof.

While the statement of Proposition 11.4.3 is somewhat complex, considering minimizers of $\mathbb{E}[\varphi(\theta, X)]$ can give some understanding. If P is a distribution such that $\mu(P) \in \mathcal{M}^{\mathrm{o}}$, then there exists a parameter $\theta(P)$ minimizing $\mathbb{E}_P[\varphi(\theta, X)]$. If $\mu(P) \in \operatorname{bd}\mathcal{M}$, then either there exists a minimizer $\theta(P)$ of the loss, or there is a sequence of points θ_n such that

$$\mathbb{E}_P[\varphi(\theta_n, X)] \to \inf_\theta \mathbb{E}_P[\varphi(\theta, X)] = -h(\mu(P)), \quad \text{and} \quad \mu(P_{\theta_n}) \to \mu(P),$$

so that they asymptotically satisfy the mean identiy. Finally, if $\mu(P) \notin \operatorname{cl}\mathcal{M}$, then $\inf_\theta \mathbb{E}[\varphi(\theta, X)] = -\infty$, making the choice of exponential family model poor, as it cannot capture the mean parameters at all.

11.4.1 Maximizing entropy

As we have seen, our notion of generalized entropies as the minimal values of expected losses can recapture the classical entropy $H(P) = -\sum_x p(x)\log p(x)$ when P has a probability mass function p, as in the case of multiclass prediction. For exponential family models, this connection goes much futher, and the (negative) generalized entropy $h(\mu)$ for $\mu \in \mathcal{M}$ coincides with a more general notion of entropy known as the *Shannon entropy*. We begin with the definition:

Definition 11.5. *Let ν be a base measure on $\mathcal{X}$ and assume P has density p with respect to ν. Then the* Shannon entropy *of P is*

$$H(P) = -\int p(x) \log p(x) d\nu(x).$$

For a distribution P with probability mass function p, the base measure ν is counting measure, yielding the classical entropy $H(P) = -\sum_x p(x) \log p(x)$, while for a distribution P with density p (for Lebesgue measure ν, so that $d\nu(x) = dx$ for $x \in \mathbb{R}^d$), we recover the differential entropy $H(P) = -\int p(x) \log p(x) dx$.

Example 11.4.4: Let P be the uniform distribution on $[0, a]$. Then the differential entropy $H(P) = -\log(1/a) = \log a$. $\diamond$

Example 11.4.5: Let P be the normal distribution $\mathsf{N}(\mu, \Sigma)$ and ν be Lebesgue measure. Then

$$H(P) = \frac{1}{2} \log(\det(2\pi\Sigma)) + \frac{1}{2} \mathbb{E}[(X - \mu)^\top \Sigma^{-1} (X - \mu)] = \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log \det(\Sigma).$$

because $p(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu))$. $\diamond$

For exponential families, the log partition determines the Shannon entropy directly, highlighting that $-h$ is indeed a familiar entropy-like object.

Proposition 11.4.6. *Let $\{P_\theta\}$ be a regular exponential family with respect to the base measure ν. Then for any $\theta \in \Theta$,*

$$H(P_\theta) = -h(\mu(P_\theta)) = A(\theta) - \langle \mu(P_\theta), \theta \rangle,$$

where $h(\mu) = \sup\{\langle \mu, \theta \rangle - A(\theta)\} = A^*(\mu)$.

Proof Using $\log p_\theta(x) = \langle \theta, \phi(x) \rangle - A(\theta)$ we obtain $H(P_\theta) = -\mathbb{E}_\theta[\langle \theta, \phi(X) \rangle - A(\theta)] = A(\theta) - \langle \mu(P_\theta), \theta \rangle$, where as usual $\mu(P) = \mathbb{E}_P[\phi(X)]$. As θ and $\mu(P_\theta)$ have the duality relationship $\nabla A(\theta) = \mu(P_\theta)$, we obtain $A(\theta) - \langle \mu(P_\theta), \theta \rangle = -h(\mu(P_\theta))$ as desired. $\square$

The *maximum entropy principal*, which Jaynes [114] first elucidated in the 1950s, originates in statistical mechanics, where Jaynes showed that (in a sense) entropy in statistical mechanics and information theory were equivalent. The maximum entropy principle is this: given some constraints (prior information) about a distribution P, we consider all probability distributions satisfying said constraints. Then to encode our prior information while being as "objective" or "agnostic" as possible (essentially being as uncertain as possible), we should choose the distribution P satisfying the constraints to maximize the Shannon entropy. This principal naturally gives rise to exponential family models, and (as we revisit later) allows connections to Bayesian and minimax procedures. One caveat throughout is that the base measure ν is *essential* to all our derivations: it radically effects the distributions P we consider.

With all this said, suppose (without making any exponential family assumptions yet) we are given $\phi : \mathcal{X} \to \mathbb{R}^d$ and a mean vector $\mu \in \mathbb{R}^d$, and we wish to solve

$$\text{maximize } H(P) \quad \text{subject to} \quad \mathbb{E}_P[\phi(X)] = \mu \tag{11.4.2}$$

over all distributions $P \in \mathcal{P}$, the collection of distributions having densities with respect to the base measure ν, that is, $P \ll \nu$. Rewriting problem (11.4.2), we see that it is equivalent to

$$\begin{aligned} \text{maximize} \quad & -\int p(x) \log p(x) d\nu(x) \\ \text{subject to} \quad & \int p(x)\phi(x)d\nu(x) = \mu, \quad p(x) \geq 0 \text{ for } x \in \mathcal{X}, \quad \int p(x)d\nu(x) = 1. \end{aligned}$$

Let

$$\mathcal{P}_\mu^{\text{lin}} := \{P \ll \nu \mid \mathbb{E}_P[\phi(X)] = \mu\}$$

be distributions with densities w.r.t. ν satisfying the expectation (linear) constraint $\mathbb{E}[\phi(X)] = \mu$. We then obtain the following theorem.

Theorem 11.4.7. *For $\theta \in \mathbb{R}^d$, let P_θ have density*

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)), \quad A(\theta) = \log \int \exp(\langle \theta, \phi(x) \rangle) d\nu(x),$$

with respect to the measure ν. If $\mathbb{E}_{P_\theta}[\phi(X)] = \mu$, then P_θ maximizes $H(P)$ over $\mathcal{P}_\mu^{\text{lin}}$; moreover, the distribution P_θ is unique (though θ need not be).

Proof We first give a heuristic derivation—which is not completely rigorous—and then check to verify that our result is exact. First, we write a Lagrangian for the problem (11.4.2). Introducing Lagrange multipliers $\lambda(x) \geq 0$ for the constraint $p(x) \geq 0$, $\theta_0 \in \mathbb{R}$ for the normalization constraint that $P(\mathcal{X}) = 1$, and $\theta \in \mathbb{R}^d$ for the constraints that $\mathbb{E}_P[\phi(X)] = \mu$, we obtain the following Lagrangian:

$$\begin{aligned} \mathcal{L}(p, \theta, \theta_0, \lambda) = & \int p(x) \log p(x) d\nu(x) + \sum_{i=1}^d \theta_i \left(\mu_i - \int p(x)\phi_i(x) d\nu(x) \right) \\ & + \theta_0 \left(\int p(x) d\nu(x) - 1 \right) - \int \lambda(x) p(x) d\nu(x). \end{aligned}$$

Now, heuristically treating the density $p = [p(x)]_{x \in \mathcal{X}}$ as a finite-dimensional vector (in the case that $\mathcal{X}$ is finite, this is completely rigorous), we take derivatives and obtain

$$\frac{\partial}{\partial p(x)} \mathcal{L}(p, \theta, \theta_0, \lambda) = 1 + \log p(x) - \sum_{i=1}^d \theta_i \phi_i(x) + \theta_0 - \lambda(x) = 1 + \log p(x) - \langle \theta, \phi(x) \rangle + \theta_0 - \lambda(x).$$

To find the minimizing p for the Lagrangian (the function is convex in p), we set this equal to zero to find that

$$p(x) = \exp\left(\langle \theta, \phi(x) \rangle - 1 - \theta_0 - \lambda(x) \right).$$

Now, we note that with this setting, we always have $p(x) > 0$, so that the constraint $p(x) \geq 0$ is unnecessary and (by complementary slackness) we have $\lambda(x) = 0$. In particular, by taking $\theta_0 = -1 + A(\theta) = -1 + \log \int \exp(\langle \theta, \phi(x) \rangle) d\nu(x)$, we have that (according to our heuristic derivation) the optimal density p should have the form

$$p_\theta(x) = \exp\left(\langle \theta, \phi(x) \rangle - A(\theta) \right).$$

So we see the form of distribution we would like to have.

Consider any distribution $P \in \mathcal{P}_\mu^{\text{lin}}$, and assume that we have some θ satisfying $\mathbb{E}_{P_\theta}[\phi(X)] = \mu$. In this case, we may expand the entropy $H(P)$ as

$$\begin{aligned} H(P) &= -\int p \log p d\nu = -\int p \log \frac{p}{p_\theta} d\nu - \int p \log p_\theta d\nu \\ &= -D_{\rm kl}\left(P \| P_\theta\right) - \int p(x)[\langle \theta, \phi(x)\rangle - A(\theta)] d\nu(x) \\ &\stackrel{(\star)}{=} -D_{\rm kl}\left(P \| P_\theta\right) - \int p_\theta(x)[\langle \theta, \phi(x)\rangle - A(\theta)] d\nu(x) \\ &= -D_{\rm kl}\left(P \| P_\theta\right) + H(P_\theta), \end{aligned}$$

where in the step $(\star)$ we have used the fact that $\int p(x)\phi(x)d\nu(x) = \int p_\theta(x)\phi(x)d\nu(x) = \mu$. As $D_{\rm kl}\left(P\|P_\theta\right) > 0$ unless $P = P_\theta$, we have shown that P_θ is the unique distribution maximizing the entropy, as desired. □

We obtain the following immediate corollary, which shows the direct connection between maximum entropy and minimizing expected logarithmic loss.

Corollary 11.4.8. *Let $\{P_\theta\}$ be the exponential family with densities $p_\theta(x) = \exp(\langle \theta, \phi(x)\rangle - A(\theta))$ with respect to ν. For any $\mu \in \mathcal{M}$, if there exists θ satisfying $\mathbb{E}_{P_\theta}[\phi(X)]$, then P_θ solves*

$$\operatorname*{minimize}_p \mathbb{E}_P[-\log p(x)]$$

over all densities p satisfying $\int \phi(x)p(x)d\nu(x) = \mu$.

So if we consider minimizing the negative log loss (which is strictly proper) but wish to guarantee that the predictive distribution satisfies $\mathbb{E}_P[\phi(X)] = \mu$, then the exponential family model is the unique minimizer.

We give three examples of maximum entropy, showing how the choice of the base measure ν effects the resulting maximum entropy distribution. For all three, we assume that the space $\mathcal{X} = \mathbb{R}$ is the real line. We consider maximizing the entropy over all distributions P satisfying

$$\mathbb{E}_P[X^2] = 1.$$

Example 11.4.9: Assume that the base measure ν is counting measure on the support $\{-1, 1\}$, so that $\nu(\{-1\}) = \nu(\{1\}) = 1$. Then the maximum entropy distribution is given by $P(X = x) = \frac{1}{2}$ for $x \in \{-1, 1\}$. ◇

Example 11.4.10: Assume that the base measure ν is Lebesgue measure on $\mathcal{X} = \mathbb{R}$, so that $\nu([a, b]) = b - a$ for $b \geq a$. Then by Theorem 11.4.7, we have that the maximum entropy distribution has the form $p_\theta(x) \propto \exp(-\theta x^2)$; recognizing the normal, we see that the optimal distribution is simply $\mathsf{N}(0, 1)$. ◇

Example 11.4.11: Assume that the base measure ν is counting measure on the integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, \ldots\}$. Then Theorem 11.4.7 shows that the optimal distribution is a discrete version of the normal: we have $p_\theta(x) \propto \exp(-\theta x^2)$ for $x \in \mathbb{Z}$. That is, we choose $\theta > 0$ so that the distribution $p_\theta(x) = \exp(-\theta x^2)/\sum_{j=-\infty}^{\infty} \exp(-\theta j^2)$ has variance 1. ◇

We remark in passing that in some cases, it is interesting to instead consider *inequality* rather than equality constraints in the linear constraints defining the family $\mathcal{P}^{\rm lin}$. Exercises 11.10 and 11.11 explore these ideas.

Lastly, we consider the empirical variant of minimizing the log loss, equivalently, of maximum likelihood, where we maximize the likelihood of a given sample $X_1, \ldots, X_n$. Consider the sample-based maximum likelihood problem of solving

$$\maximize_\theta \prod_{i=1}^n p_\theta(X_i) \equiv \minimize -\frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i), \tag{11.4.3}$$

for the exponential family model $p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$. We have the following result.

Proposition 11.4.12. *Let $\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \phi(X_i)$. Then any θ solving $\mathbb{E}_{P_\theta}[\phi(X)] = \widehat{\mu}_n$ is a maximum likelihood solution, which exists if and only if $\widehat{\mu}_n \in \relint \mathcal{M}$. If the sample is drawn $X_i \stackrel{\rm iid}{\sim} P$ where $P \ll \nu$ and $\mu(P) \in \relint \mathcal{M}$, then with probability 1, $\widehat{\mu}_n \in \relint \mathcal{M}$ eventually.*

Proof Define the empirical negative log likelihood

$$\widehat{L}_n(\theta) := -\frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i) = -\langle \widehat{\mu}_n, \theta \rangle + A(\theta),$$

which is convex. Taking derivatives and using that $\Theta = \dom A$ is open, the parameter θ is a minimizer if and only if $\nabla \widehat{L}_n(\theta) = \widehat{\mu}_n - \nabla A(\theta) = 0$ if and only if $\nabla A(\theta) = \widehat{\mu}_n$. Apply Proposition 11.4.1.

For the final statement, note that $\widehat{\mu} \in \aff(\mathcal{M})$ with probability 1. Then because $\mu(P) \in \relint \mathcal{M}$ and $\widehat{\mu}_n \to \mu(P)$ with probability 1, we see that for any $\epsilon > 0$ there is some (random, but finite) N such that $n \geq N$ implies $\|\widehat{\mu}_n - \mu(P)\| \leq \epsilon$ and $\widehat{\mu}_n \in \aff(\mathcal{M})$, so that $\widehat{\mu}_n \in \relint \mathcal{M}$. □

As a consequence of the result, we have the following rough equivalences tying together the preceding material. In short, maximum entropy subject to (linear) empirical moment constraints (Theorem 11.4.7) is equivalent to maximum likelihood estimation in exponential families (Proposition 11.4.12), and these are all equivalent to minimizing the (surrogate) log loss $\mathbb{E}[\varphi(\theta, X)]$.

11.4.2 I-projections and maximum likelihood

Certainly exponential family models cannot capture all possible distributions on $\mathcal{X}$ or even distributions $P \ll \nu$ on $\mathcal{X}$. As Corollary 11.4.8 shows, exponential family models minimize the log loss. They also solve certain projection-like problems onto different families of distributions. First, suppose that we have a family Π of distributions and some fixed distribution P. Then the *I-Projection* (for information projection) of the distribution P onto the family Π is

$$P^\star := \argmin_{Q \in \Pi} D_{\rm kl}\left(Q \| P\right), \tag{11.4.4}$$

when such a distribution exists.

By making a small tweak to the exponential family models we consider, we can show that exponential family models also solve the I-projection problem. Indeed, if we assume P has density p with respect to ν and let P_θ have density

$$p_\theta(x) = p(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \quad \text{for} \quad A(\theta) := \log \int \exp(\langle \theta, \phi(x) \rangle) p(x) d\nu(x)$$

so that p is the carrier of P_θ (recall Chapter 3). The next proposition uses this to show, perhaps unsurprisingly given our derivations thus far, that I-Projection is essentially the same as maximum entropy, and the projection of a distribution P onto a family of linearly constrained distributions yields exponential family distributions.

Proposition 11.4.13. *Let $\Pi = \mathcal{P}_\mu^{\rm lin}$. If $p_\theta(x) = p(x)\exp(\langle\theta, \phi(x)\rangle - A(\theta))$ satisfies $\mathbb{E}_{P_\theta}[\phi(X)] = \mu$, then P_θ solves the I-projection problem (11.4.4). Moreover we have the Pythagorean identity*

$$D_{\rm kl}\left(Q\|P\right) = D_{\rm kl}\left(P_\theta\|P\right) + D_{\rm kl}\left(Q\|P_\theta\right)$$

for $Q \in \mathcal{P}_\mu^{\rm lin}$.

Proof We perform an expansion of the KL-divergence parallels that in the proof of Theorem 11.4.7. Indeed, for any $Q \ll \nu$, we have

$$D_{\rm kl}\left(Q\|P\right) = \int q \log \frac{q}{p} d\nu = \int q \log \frac{p_\theta}{p} d\nu + D_{\rm kl}\left(Q\|P_\theta\right) = \int q(x)\left[\langle\theta, \phi(x)\rangle - A(\theta)\right] d\nu(x) + D_{\rm kl}\left(Q\|P_\theta\right)$$

because $p_\theta(x) = p(x)\exp(\langle\theta, \phi(x)\rangle - A(\theta))$. Then because $Q \in \mathcal{P}_\mu^{\rm lin}$, we have $\int q(x)[\langle\theta, \phi(x)\rangle - A(\theta)]d\nu(x) = \int p_\theta \log \frac{p_\theta}{p} d\nu = D_{\rm kl}\left(P_\theta\|P\right)$, giving the proposition. □

In brief, the exponential family model is the projection—in the sense of the KL divergence—of a distribution P onto the collection of distributions satisfying $\mathbb{E}[\phi(X)] = \mu$.

11.5 Technical and deferred proofs

11.5.1 Finalizing the proof of Theorem 11.2.14

The issue remaining in the proof of Theorem 11.2.14 occurs when $\ell(\mu, y_i) = +\infty$ for some i. In this case, we necessarily have $p_i = 0$ for all $p \in \Delta_m$ satisfying $\mathbb{E}_p[Y] = \mu$; define the set of infinite loss indices $\mathcal{I}(\mu) := \{i \mid L(\mu, y_i) = +\infty\}$, which is evidently in the set $\{i \mid p_i = 0 \text{ whenever } Ap = 0\}$. Because of this containment, we vectors $\{y_i\}_{i\in\mathcal{I}(\mu)}$ are independent and independent of $\{y_i\}_{i\notin\mathcal{I}(\mu)}$. In particular, there exists $\Delta \in \mathbb{R}^k$ such that $y_i^T\Delta = 0$ for all $i \notin \mathcal{I}(\mu)$ but for which $y_i^T\Delta > 0$ for each $i \in \mathcal{I}(\mu)$. Working on the subspace $\{p \in \Delta_m \mid p_i = 0, i \in \mathcal{I}(\mu)\}$, we can perform precisely the same derivation except that $G(\mu) = \{s \in \mathbb{R}^k \mid y_i^T s = -\ell(\mu, y_i) \text{ for } i \notin \mathcal{I}(\mu)\}$ is non-empty. Then we have

$$h(\mu') = -\mathbb{E}_{p^\star(\mu')}[\ell(\mu', Y)] \stackrel{(i)}{\geq} -\mathbb{E}_{p^\star(\mu')}[\ell(\mu, Y)] = -\mathbb{E}_{p^\star(\mu)}[\ell(\mu, Y)] + \sum_{i=1}^m \ell(\mu, y_i)(p_i^\star(\mu) - p_i^\star(\mu')),$$

where inequality (i) follows because ℓ is proper. We then have

$$\begin{aligned}\sum_{i=1}^m \ell(\mu, y_i)(p_i^\star(\mu) - p_i^\star(\mu')) &\stackrel{(ii)}{=} \sum_{i\notin\mathcal{I}(\mu)} \ell(\mu, y_i)(p_i^\star(\mu) - p_i^\star(\mu')) - \sum_{i\in\mathcal{I}(\mu)} \ell(\mu, y_i)p_i^\star(\mu') \\ &= \sum_{i\notin\mathcal{I}(\mu)} s^T y_i(p_i^\star(\mu) - p_i^\star(\mu')) - \sum_{i\in\mathcal{I}(\mu)} \ell(\mu, y_i)p_i^\star(\mu')\end{aligned}$$

for any $s \in G(\mu)$, where equality (ii) follows because $p_i^\star(\mu) = 0$ for $i \in \mathcal{I}(\mu)$. As we allow extended reals, replace s with $s_\infty = \lim_{t\to\infty}(s + t\Delta)$, which satisfies $\langle s_\infty, y_i\rangle = \infty = \ell(\mu, y_i)$ for $i \in \mathcal{I}(\mu)$, and we finally obtain

$$h(\mu') \geq h(\mu) + \sum_{i=1}^m s_\infty^T y_i(p_i^\star(\mu) - p_i^\star(\mu')) = h(\mu) + \langle s_\infty, \mu - \mu'\rangle.$$

The equality of the loss is as before.

11.5.2 Proof of Proposition 11.4.1

We first give the proof in the case that $\{P_\theta\}$ is a minimal exponential family, meaning that $\langle u, \phi(x)\rangle$ is non-constant in x for each $u \neq 0$, addressing the non-minimal case at the end. Then A is strictly convex (Proposition 3.2.3). As part of this proof, we will show that $\mathcal{M}^{\rm o}$ is indeed open in this case. We show both inclusions $\mathcal{M}^{\rm o} \subset \nabla A(\Theta)$ and that $\nabla A(\Theta) \subset \mathcal{M}^{\rm o}$.

Showing that $\nabla A(\Theta) \subset \mathcal{M}^{\rm o}$. Fix $\theta_0 \in \Theta$, and let $\mu = \nabla A(\theta_0)$. We must show that there exists $\epsilon > 0$ such that for all $\|u\| \leq \epsilon$, the point $\mu + u \in \mathcal{M}$. Let $\theta_u = \operatorname{argmin}_\theta\{A(\theta) - \langle \mu + u, \theta\rangle\}$ whenever the minimizer exists, where evidently θ_0 does exist because $\mu = \nabla A(\theta_0)$. Note that the strict convexity of A guarantees θ_u is unique if it exists. But now, we may use the convex analyitic fact (Proposition C.1.10 in Appendix C.1.2) that $u \mapsto \theta_u$ is continuous in u in a neighborhood of 0. These minimizers necessarily satisfy $\nabla A(\theta_u) = \mu + u$, that is, $\mathbb{E}_{\theta_u}[\phi(X)] = \mu + u \in \mathcal{M}$.

Showing that $\mathcal{M}^{\rm o} \subset \nabla A(\Theta)$. Let $\mu \in \mathcal{M}^{\rm o}$, so that there exists an $\epsilon > 0$ such that $\mu + \epsilon\mathbb{B} \subset \mathcal{M}^{\rm o}$. It is enough to show that $A(\theta) - \langle \mu, \theta\rangle$ is coercive in θ, as then there necessarily exists a (unique) minimizer $\theta(\mu)$ of $A(\theta) - \langle \mu, \theta\rangle$, and this minimizer satisfies $\nabla A(\theta(\mu)) = \mu$, so that $\mu \in \nabla A(\Theta)$. For this, it is sufficient to show that for any non-zero vector v the *recession function* of the tilted version $f(\theta) := A(\theta) - \langle \mu, \theta\rangle$ of A,

$$f'_\infty(v) := \lim_{t\to\infty} \frac{A(\theta + tv) - \langle \mu, \theta + tv\rangle - (A(\theta) - \langle \mu, \theta\rangle)}{t}$$

where $\theta \in \Theta$ is otherwise arbitrary, satisfies $f'_\infty(v) > 0$ for all $v \neq 0$, which guarantees that $A(\cdot) - \langle \mu, \cdot\rangle$ has a minimizer. (See Proposition C.2.5 and Corollary C.2.6 in Appendix C.2.1).

To that end, for vectors $v \in \mathbb{R}^d$, define the essential supremum of $\phi(x)$ in the direction v by

$$\nu^\star(\phi, v) := \operatorname*{ess\,sup}_x \langle \phi(x), v\rangle = \inf_t \left\{t \in \mathbb{R} \mid \nu(\{x \in \mathcal{X} \mid \langle v, \phi(x)\rangle \geq t\}) = 0\right\}.$$

Now as $\mu \in \mathcal{M}^{\rm o}$, for any vector $v \neq 0$ we have $\langle v, \mu\rangle < \nu^\star(\phi, v)$. Let $\epsilon > 0$ satisfy $\langle v, \mu\rangle < \nu^\star(\phi, v) - \epsilon$ be otherwise arbitrary, fix $\theta \in \Theta$, and let $\mathcal{X}_\epsilon = \{x \mid \langle v, \phi(x)\rangle \geq \nu^\star(\phi, v) - \epsilon\}$, which satisfies $\nu(\mathcal{X}_\epsilon) > 0$. Then

$$\begin{aligned}
A(\theta + tv) - \langle \mu, \theta + tv\rangle &= \log \int \exp(\langle \phi(x), \theta + tv\rangle) d\nu(x) - \langle \mu, \theta + tv\rangle \\
&\geq \log \int_{\mathcal{X}_\epsilon} \exp(\langle \phi(x), \theta\rangle) e^{t(\nu^\star - \epsilon)} d\nu(x) - \langle \mu, \theta\rangle - t\langle \mu, v\rangle \\
&= t(\nu^\star(\phi, v) - \epsilon) + \log \nu(\mathcal{X}_\epsilon) - t\langle \mu, v\rangle + \log \int_{\mathcal{X}_\epsilon} e^{\langle \phi(x), \theta\rangle} d\nu(x) - \langle \mu, \theta\rangle.
\end{aligned}$$

If $\nu(\mathcal{X}_\epsilon) = +\infty$, then $A(\theta + tv) = +\infty$ and so $A'_\infty(v) > 0$ certainly. If $\nu(\mathcal{X}_\epsilon) < \infty$, then note that $\nu^\star(\phi, v) - \epsilon - \langle \mu, v\rangle > 0$, and so

$$A(\theta + tv) - \langle \mu, \theta + tv\rangle \geq t(\nu^\star(\phi, v) - \epsilon - \langle \mu, v\rangle) - \log \nu(\mathcal{X}_c) + \log \int_{\mathcal{X}_c} e^{\langle \phi(x), \theta\rangle} d\nu(x) - \langle \mu, \theta\rangle$$

and thus

$$\frac{A(\theta + tv) - \langle \mu, \theta + tv\rangle - (A(\theta) - \langle \mu, v\rangle)}{t} \geq \nu^\star(\phi, v) - \epsilon - \langle \mu, v\rangle + o(1) \tag{11.5.1}$$

as $t \to \infty$.

Extending to the non-minimal case. If the exponential family is not minimal, there exists a unit vector u and constant c such that $\langle u, \phi(x)\rangle = c$ for ν-almost all x. Let $U \in \mathbb{R}^{d\times k}$ be an orthonormal basis for all such vectors, where k is the dimension of this collection. Then there exists a vector $c \in \mathbb{R}^k$ such that $c = U^\top \phi(x)$ for ν-almost all x, and we see that $A(\theta + Uv) = A(\theta) + \langle c, v\rangle$ as $\langle \theta + Uv, \phi(x)\rangle = \langle \theta, \phi(x)\rangle + \langle c, v\rangle$ for ν-almost all x. We show both inclusions as above. Let $U_\perp \in \mathbb{R}^{d\times d-k}$ be an orthonormal basis for the orthogonal subspace to U, so that $U^\top U = I_k$ and $U_\perp^\top U_\perp = I_{d-k}$, and for any $\mu \in \mathcal{M}$, we have $\mathrm{aff}(\mathcal{M}) = \mu + \mathrm{span}(U_\perp)$.

Showing that $\nabla A(\Theta) \subset \mathcal{M}^{\mathrm{o}}$. Fix $\theta_0 \in \Theta$ and let $\mu = \nabla A(\theta_0)$. We must show that there exists $\epsilon > 0$ such that for all $u \in \mathrm{span}(U_\perp)$ satisfying $\|u\| \leq \epsilon$, the point $\mu + u \in \mathcal{M}$. To that end, note that for any vectors $v \in \mathbb{R}^{d-k}$ and $w \in \mathbb{R}^k$, we have

$$A(\theta_0 + U_\perp v + Uw) - \langle \mu + u, U_\perp v + Uw\rangle = A(\theta_0 + U_\perp v) - \langle \mu + u, U_\perp v\rangle$$

because $U^\top u = 0$ and $U^\top \mu = c$ for each $u \in \mathrm{span}(U_\perp)$ and $\mu \in \mathcal{M}$. The function $g(v) := A(\theta_0 + U_\perp v) - \langle \mu, U_\perp v\rangle$ is strictly convex as $\nabla^2 g(v) = U_\perp^\top \nabla^2 A(\theta_0 + U_\perp v) U_\perp \succ 0$, because we know that $u^\top \phi(x)$ is non-constant for all $u \in \mathrm{span}(U_\perp)$. Define $f(v) = A(\theta_0 + U_\perp v) - \langle \mu, U_\perp v\rangle$. Then applying Proposition C.1.10 as in the minimal representation case, there exists $\epsilon > 0$ such that $v_u = \mathrm{argmin}_v\{f(v) - \langle u, U_\perp v\rangle\}$ exists and is continuous in $u \in \mathrm{span}(U_\perp)$, where by inspection $v_0 = 0$. Then $\theta_u := \theta_0 + U_\perp v_u$ minimizes $A(\theta) - \langle \mu + u, \theta\rangle$, satisfying $\nabla A(\theta_u) = \mu + u$.

Showing that $\mathcal{M}^{\mathrm{o}} \subset \nabla A(\Theta)$. We again follow the logic of the minimal representation case. Let $\mu \in \mathcal{M}^{\mathrm{o}} = \mathrm{relint}\, \mathcal{M}$, and recall $\nu^\star(\phi, U_\perp v) = \mathrm{ess\,sup}_x \langle \phi(x), U_\perp v\rangle$. Then there exists $\epsilon > 0$ such that $\mu + u \in \mathcal{M}$ for each $u \in \mathrm{span}(U_\perp)$ with $\|u\| \leq \epsilon$, so that

$$\langle \mu, U_\perp v\rangle < \sup_{\|u\|_2 \leq \epsilon, u \in \mathrm{span}(U_\perp)} \langle \mu + u, U_\perp v\rangle \leq \nu^\star(\phi, U_\perp v).$$

Define $g(v) = A(\theta + U_\perp v) - \langle \mu, U_\perp v\rangle$. Then because $A(\theta + Uw + U_\perp v) - \langle \mu, U_\perp v - Uw\rangle = A(\theta + U_\perp v) - \langle \mu, U_\perp v\rangle$ for all $w \in \mathbb{R}^k, v \in \mathbb{R}^{d-k}$, it is enough to show that $g'_\infty(v) > 0$ for all $v \neq 0$. Following the same argument, *mutatis mutandis*, as that leading to inequality (11.5.1) yields that $g'_\infty(v) > 0$ for all $v \neq 0$. That is, $v \mapsto A(\theta + U_\perp v) - \langle \mu, U_\perp v\rangle$ has a minimizer $v(\mu)$ (Corollary C.2.6), which is unique by the strict convexity of $v \mapsto A(\theta + U_\perp v)$, and which necessarily satisfies $U_\perp^\top \nabla A(\theta + U_\perp v(\mu)) = U_\perp^\top \mu$. As $U^\top \nabla A(\theta) = c$ for all θ and $U^\top \mu = c$ for all $\mu \in \mathcal{M}$, this shows that there exists $\theta(\mu)$ such that $\nabla A(\theta(\mu)) = \mu$ as desired. Moreover, fixing an arbitrary θ and letting $v(\mu)$ be the unique minimizer of $A(\theta + U_\perp v) - \langle \mu, U_\perp v\rangle$, the set of all minimizers

$$\Theta^\star(\mu) = \mathrm{argmin}_\theta\{A(\theta) - \langle \mu, \theta\rangle\} = \left\{\theta + U_\perp v(\mu) + Uw \mid w \in \mathbb{R}^k\right\}.$$

This gives Proposition 11.4.1.

11.5.3 Proof of Proposition 11.4.3

For part (i), because $\Theta = \mathrm{dom}\, A \subset \mathbb{R}^d$ is open and A is $\mathcal{C}^\infty$ on its domain, A is necessarily a closed convex function and so $A^{**}(\theta) = A(\theta)$ for all $\theta \in \mathbb{R}^d$. (See Theorem C.2.1.) For part (ii), note

that if $\mu \in \mathcal{M}^o$, there exists $\theta(\mu) \in \Theta$ such that $\nabla A(\theta(\mu)) = \mu$ by Proposition 11.4.1. This $\theta(\mu)$ maximizes $\langle \theta, \mu \rangle - A(\theta)$ over all θ, and so $h(\mu) = \langle \theta(\mu), \mu \rangle - A(\theta(\mu)) < \infty$. By Corollary C.2.4 in Appendix C.2.1 and Proposition 11.4.1, $\operatorname{dom} \partial h = \mathcal{M}^o$, and as h is subdifferentiable on the relative interior of its domain, we have $\operatorname{dom} h \subset \operatorname{cl} \mathcal{M}^o = \operatorname{cl} \mathcal{M}$. As h is closed convex, any point μ outside its domain necessarily satisfies $h(\mu) = +\infty$.

Finally, for part (iii), we note that the function $g(t) = h(t\mu_0 + (1-t)\mu)$ is a one-dimensional closed convex function. One-dimensional closed convex functions are continuous on their domains (Observation B.3.6 in Appendix B.3.2), and so g is necessarily continuous. Thus $\lim_{t \downarrow 0} g(t) = g(0)$. The existence of θ_t follows from Proposition 11.4.1.

Bibliography

JCD Comment: Need to do a lot here!

Gneiting and Raftery [93]

11.6 Exercises

Exercise 11.1 (Strict propriety of the log loss)**:** Let $\Delta_k = \{p \in \mathbb{R}^k_+ \mid \mathbf{1}^T p = 1\}$ be the probability simplex. Show that if $\ell(q, y) = -\log q_y$ and $\mathbb{P}(Y = y) = p_y$, then

$$\operatorname*{argmin}_{q \in \Delta_k} \mathbb{E}[\ell(q, Y)] = p,$$

where we treat $0 \log 0$ as 0 (which is the natural limit of $t \log t$ as $t \downarrow 0$).

Exercise 11.2 (Uniqueness of generalized entropies)**:** Here we give an alternative perspective on the generalized entropies associated with losses, showing when they are unique. For a concave function $f : \Delta_k \to \mathbb{R}$, define the perspective-type transform $f_{\rm per}(p) = \langle \mathbf{1}, p \rangle f(p / \langle \mathbf{1}, p \rangle)$, where $f_{\rm per}(0) = 0$, and which gives $f_{\rm per} : \mathbb{R}^k_+ \to \underline{\mathbb{R}}$.

(a) Let $\ell : \Delta_k \to \overline{\mathbb{R}}$ be strictly proper and let Y have p.m.f. p. Show that $H(p) = \inf_{q \in \Delta_k} \mathbb{E}[\ell(q, Y)]$ is strictly concave, and that $H_{\rm per}$ is strictly concave and continuously differentiable on $\mathbb{R}^k_{++}$.

(b) Show the converse that if $H : \Delta_k \to \mathbb{R}$ is strictly concave and its perspective $H_{\rm per}$ is differentiable on $\mathbb{R}^k_{++}$, then there exists a proper scoring loss ℓ satisfying

$$H(p) = \inf_{q \in \Delta_k} \mathbb{E}_p[\ell(q, Y)]$$

and that $\ell(q, y) = \nabla_y H_{\rm per}(q)$ for all $q \in \operatorname{dom} \nabla H_{\rm per}$.

Exercise 11.3: Give the details in the computations for Example 11.3.4.

Exercise 11.4: Let $y \in \{0, 1\}$ and take the regularization function $h(p) = -\log p - \log(1-p)$.

(a) Verify that the entropy is of Legendre type (Definition 11.4).

(b) Give the associated loss ℓ and surrogate loss φ in the sense of Section 11.3.

(c) Plot the surrogate $\varphi(s, y) + \log 8$ and the logistic regression surrogate $\log(1 + e^s) - sy$ for $y \in \{0, 1\}$, each as function of s. (The shift by $\log 8$ guarantees the losses coincide at $s = 0$.)

(d) Give $\mathsf{pred}_h(s)$ for $s \in \mathbb{R}$, verifying that $\mathsf{pred}_h(s) \in [0, 1]$.

Exercise 11.5: For $h(p) = -\log p - \log(1-p)$ as in Exercise 11.4, show that h is *self-concordant*, meaning that $h'''(p) \le 2(h''(p))^{3/2}$ for all $p \in (0, 1)$. (Such functions are important in optimization; the conjugate h^* is then also guaranteed to be self-concordant.)

Exercise 11.6 (Surrogates for regression)**:** Define $h(c) = \frac{1}{4}c^4$.

(a) Give the conjugate $h^*(s)$ to h.

(b) Show directly that the surrogate loss $\varphi(s, y) = h^*(s) - sy$ satisfies that if $\hat{s} = \operatorname{argmin}_s \mathbb{E}[\varphi(s, Y)]$, then $\mathsf{pred}_h(\hat{s}) = \mathbb{E}[Y]$.

Exercise 11.7: Let P be a predicted distribution and for $\alpha \in [0, \frac{1}{2}]$, define the lower and upper quantiles $l_\alpha = \mathsf{Quant}_\alpha(P)$ and $u_\alpha = \mathsf{Quant}_{1-\alpha}(P)$. Given these quantiles, for a finite set $\mathcal{A} \subset [0, \frac{1}{2}]$, define the weighted interval loss

$$W(P, y) := \sum_{\alpha \in \mathcal{A}} \left[\alpha(u_\alpha - l_\alpha) + \operatorname{dist}(y, [l_\alpha, u_\alpha])\right],$$

which penalizes P using both the size $(u_\alpha - l_\alpha)$ of the quantile intervals and the distance of the outcome y from the predicted quantiles. Define the symmetrized set $\mathcal{A}_{\mathrm{s}} = \mathcal{A} \cup \{1 - \alpha \mid \alpha \in \mathcal{A}\}$. Show that

$$W(P, y) = \ell_{\mathrm{quant}, \mathcal{A}_{\mathrm{s}}}(P, y),$$

where ℓ_{quant} is the quantile loss (11.2.4).

Exercise 11.8: We explore a particularization of the results in Section 11.4. Let $Y \sim \mathsf{Poi}(e^\theta)$, so that Y has p.m.f. $p_\theta(y) = \exp(\theta y - e^\theta)/y!$ for $y \in \mathbb{N}$. Let $A(\theta) = e^\theta$ be the log-partition function. Define the "surrogate" loss $\varphi(\theta, y) = -\log p_\theta(y)$.

(a) Give the associated negative generalized entropy $h(\mu)$ for $\mu \in (0, \infty)$.

(b) Give the associated loss $\ell(\mu, y)$ in the proper representation of Theorem 11.2.14. Directly verify that it is strictly proper, in that $\operatorname{argmin}_\mu \mathbb{E}[\ell(\mu, Y)] = \mathbb{E}[Y]$ for any Y supported on $\mathbb{R}_+$.

Exercise 11.9: We explore a particularization of Example 11.4. Let $X \sim \mathsf{N}(0, \Sigma)$ for a covariance $\Sigma \succ 0$, and let $K = \Sigma^{-1}$ be the associated precision matrix. Then X has density $p_K(x) = \exp(-\frac{1}{2}\langle xx^T, K\rangle + \frac{1}{2}\log\det(K))$ with respect to (a scaled) Lebesgue measure, and log partition $A(K) = -\frac{1}{2}\log\det(K)$, which has domain the positive definite matrices $K \succ 0$ (and is $+\infty$ elsewhere).

(a) Give the associated negative generalized entropy $h(M)$ for symmetric matrices M. Specify the domain of h.

(b) Give the associated loss $\ell(M, x)$ in the proper representation of Theorem 11.2.14. Directly verify that it is strictly proper, in that if the second moment matrix $C := \mathbb{E}[XX^T]$ of X satisfies $C \succ 0$, then $\operatorname{argmin}_M \mathbb{E}[\ell(M, X)] = C$.

Exercise 11.10: In this extended exercise, we generalize Theorem 11.4.7 to apply to general (finite-dimensional) convex cone constraints. A set $\mathcal{C}$ is a *convex cone* if for any two points $x, y \in \mathcal{C}$, we have $\lambda x + (1-\lambda)y \in \mathcal{C}$ for all $\lambda \in [0,1]$, and $\mathcal{C}$ is closed under positive scaling: $x \in \mathcal{C}$ implies that $tx \in \mathcal{C}$ for all $t \geq 0$. The following are standard examples (the positive orthant and the semi-definite cone):

i. *The orthant.* Take $\mathcal{C} = \mathbb{R}^d_+ = \{x \in \mathbb{R}^d : x_j \geq 0, j = 1, \ldots, d\}$. Then clearly $\mathcal{C}$ is convex and closed under positive scaling.

ii. *The semidefinite cone.* Take $\mathcal{C} = \{X \in \mathbb{R}^{d \times d} : X = X^\top, X \succeq 0\}$, where a matrix $X \succeq 0$ means that $a^\top X a \geq 0$ for all vectors a. Then $\mathcal{C}$ is convex and closed under positive scaling as well.

Given a convex cone $\mathcal{C}$, we associate a cone ordering $\succeq$ with the cone and say that for two elements $x, y \in \mathcal{C}$, we have $x \succeq y$ if $x - y \succeq 0$, that is, $x - y \in \mathcal{C}$. In the orthant case, this simply means that x is component-wise larger than y. For a given inner product $\langle \cdot, \cdot \rangle$, define the dual cone

$$\mathcal{C}^* := \{y : \langle y, x \rangle \geq 0 \text{ for all } x \in \mathcal{C}\}.$$

For the standard (Euclidean) inner product, the positive orthant is thus self-dual, and similarly the semidefinite cone is also self-dual. For a vector y, we write $y \succeq_* 0$ if $y \in \mathcal{C}^*$ is in the dual cone. With this setup, consider the following linearly constrained maximum entropy problem, where the cone ordering $\preceq$ derives from a cone $\mathcal{C}$:

$$\text{maximize } H(P) \quad \text{subject to} \quad \mathbb{E}_P[\phi(X)] = \mu, \quad \mathbb{E}_P[\psi(X)] \preceq \beta, \tag{11.6.1}$$

where the base measure ν is implicit. Let $\mathcal{P}^{\text{lin}}_{\mu,\beta}$ be the collection of distributions $P \ll \nu$ satisfying the constraints in problem (11.6.1).

Prove the following theorem:

Theorem 11.6.1. *For $\theta \in \mathbb{R}^d$ and $K \in \mathcal{C}^*$, the dual cone to $\mathcal{C}$, let $P_{\theta,K}$ have density*

$$p_{\theta,K}(x) = \exp\left(\langle \theta, \phi(x) \rangle - \langle K, \psi(x) \rangle - A(\theta, K)\right), \quad A(\theta, K) = \log \int \exp(\langle \theta, \phi(x) \rangle - \langle K, \psi(x) \rangle) d\nu(x),$$

with respect to the measure ν. If

$$\mathbb{E}_{P_{\theta,K}}[\phi(X)] = \mu \quad \textit{and} \quad \mathbb{E}_{P_{\theta,K}}[\psi(X)] = \beta,$$

then $P_{\theta,K}$ maximizes $H(P)$ over $\mathcal{P}^{\text{lin}}_{\mu,\beta}$. Moreover, the distribution $P_{\theta,K}$ is unique.

Exercise 11.11 (An application of Theorem 11.6.1)**:** Let the cone $\mathcal{C}$ be the positive semidefinite cone in $\mathbb{R}^{d \times d}$, ν be the Lebesgue measure $d\nu(x) = dx$ and define $\psi(x) = \frac{1}{2} x x^\top \in \mathbb{R}^{d \times d}$. Let $\Sigma \succ 0$. Give the density solving

$$\text{maximize } -\int p(x) \log p(x) dx \text{ subject to } \mathbb{E}_P[XX^\top] \preceq \Sigma.$$

Exercise 11.12: Prove that the log determinant function is concave over the positive semidefinite matrices. That is, show that for $X, Y \in \mathbb{R}^{d \times d}$ satisfying $X \succeq 0$ and $Y \succeq 0$, we have

$$\log \det(\lambda X + (1-\lambda)Y) \geq \lambda \log \det(X) + (1-\lambda) \log \det(Y)$$

for any $\lambda \in [0, 1]$. *Hint:* think about log-partition functions.

Exercise 11.13 (Entropy and log-determinant maximization)**:** Consider the following optimization problem over symmetric positive semidefinite matrices in $\mathbb{R}^{d \times d}$:

$$\operatorname*{maximize}_{\Sigma \succeq 0} \quad \log \det(\Sigma) \quad \text{subject to } \Sigma_{ij} = \sigma_{ij}$$

where σ_{ij} are specified only for indices $i, j \in S$ (but we know that $\sigma_{ij} = \sigma_{ji}$ and $(i, i) \in S$ for all i). Let Σ^* denote the solution to this problem, assuming there is a positive definite matrix Σ satisfying $\Sigma_{ij} = \sigma_{ij}$ for $i, j \in S$. Show that for each unobserved pair $(i, j) \notin S$, the (i, j) entry $[\Sigma^{*-1}]_{ij}$ of the inverse Σ^{*-1} is 0. *Hint:* The distribution maximizing the entropy $H(X) = -\int p(x) \log p(x) dx$ subject to $\mathbb{E}[X_i X_j] = \sigma_{ij}$ has Gaussian density of the form $p(x) = \exp(\sum_{(i,j) \in S} \lambda_{ij} x_i x_j - \Lambda_0)$.

Exercise 11.14: **JCD Comment:** Finish this.
Equivalence of integrated quantile losses and continuous ranked probability score.

Chapter 12

Calibration and Proper Losses

In Chapter 11, we encountered *proper losses*, in which we assume we predict probability distributions on outcomes Y. In typical problems, we wish to predict things about Y from a given set of covariates or inputs X, and in focusing exclusively on the losses ℓ themselves, we implicitly assume that we can model $Y \mid X$ basically perfectly. Here, we move away from this focus exclusively on the loss itself to incorporate discussion of predictions, where we seek a function $f : \mathcal{X} \to \mathcal{Y}$ (or some other output space) that yields the most accurate predictions.

In this chapter, we adopt the view of Section 11.2.3, where the target $Y \subset \mathbb{R}^k$ is vector-valued, and we wish to predict its expectation $\mathbb{E}[Y \mid X]$ as accurately as possible. For binary prediction, we have $Y \in \{0, 1\}$, so that $\mathbb{E}[Y \mid X] = \mathbb{P}(Y = 1 \mid X)$; in the case of multiclass prediction problems, it is easy to represent Y as an element of the k standard basis vectors $\{e_1, \ldots, e_k\} \subset \mathbb{R}^k$, so that $p = \mathbb{E}[Y \mid X]$ is simply the p.m.f. of Y given X with entries $p_y = \mathbb{P}(Y = y \mid X)$. We focus here, therefore, on choosing functions to minimize the risk, or expected population loss,

$$L(f) := \mathbb{E}[\ell(f(X), Y)].$$

When f is chosen from a collection $\mathcal{F} \subset \{\mathcal{X} \to \mathbb{R}^k\}$ of functions, for example, to guarantee that we can generalize, we do not expect to be able to perfect minimize the population loss. Accordingly, even though the loss is proper and hence minimized by $f^\star(x) = \mathbb{E}[Y \mid X = x]$, we cannot perfectly model reality, and so it is unrealistic to expect to be able to find f satisfying $f(x) = \mathbb{E}[Y \mid X = x]$, even approximately, for all x.

We therefore depart from the goal of perfection to address a somewhat simpler criterion: that of calibration. Here, the idea is that a predictor should be accurate on average conditional on its own predictions. Consider again a weather forecasting problem, where $Y_t = 1$ indicates it rains on day t and $Y_t = 0$ indicates no rain, and we wish to predict Y_t based on observable covariates X_t at time t. While we would like a forecaster to have perfect predictions $p_t = \mathbb{E}[Y_t \mid X_t]$, we instead ask that on days where the forecaster makes a given prediction, it should rain (roughly) with that given frequency. In particular, we seek *calibration*, which is that

$$f(X) = \mathbb{E}[Y \mid f(X)]. \tag{12.0.1}$$

That is, given that the forecaster makes a prediction with value $p = f(X)$, we should have

$$\mathbb{E}[Y \mid f(X) = p] = p.$$

While in general it is challenging to achieve this perfect calibration, in this chapter we investigate several variants of the desideratum (12.0.1) that allow for more elegant statistical and information-theoretic approaches, as well as procedures to achieve calibration.

This chapter therefore proceeds as follows. The first goal is to

JCD Comment: Fix notation. Also add a transition here to make clearer why we are doing this and what we are doing.

1. First show what we want to measure.
2. Show how to measure it, specifically using partitioned methods. I think that partitioned ones should be better than non-partitioned approaches, because we can estimate the binned / partitioned calibration error
3. Show a few ways to achieve it (population and finite-sample level).

It is important to note that the literature on calibration is broad, and there are several distinct strands. We take the particular focus that most dovetails with our treatment of proper losses and scoring rules, basing our development around random variables and finite-dimensional probabilities. So, for example, if a logistic regression model (as in Example 3.4.2 or 3.4.3) for image classification assigns a probability of 80% that an image is, say, a dog, then the model is (approximately) calibrated if in the population of all images in the world to which the model assigns probabilty 80%, (approximately) 80% are dogs. The first direction of research that we essentially do not touch are the following: in the forecasting literature, one often considers predicting the distribution of a (potentially continuous) random variable Y, such as the amount of rainfall; if we predict a cumulative distribution F as in Example 11.2.6, then perfect calibration (12.0.1) becomes that

$$\mathbb{P}(Y \le u \mid F) = F(u) \quad \text{for all } u \in \mathbb{R}.$$

This is far too stringent a condition to be achievable, so that one relaxes to various forms of marginal or average calibration. See the bibliographic notes for some discussion of the approaches here.

The second strand of research on calibration that, again, we do not address, considers more adversarial and sequential settings, where instead of any probabilistic underpinnings, nature (an adversary) plays a game against the player (or predictor). Philosophically, this approach elegantly does away with the need for probabilities: there is a physical world where whether it rains tomorrow is essentially deterministic, and we use probability as a crutch to model things we cannot measure, so calibration means that of the days on which we predict rain with a chance of 50%, it rains on roughly 50% of those days. In this sequential setting, at times $t = 1, 2, \ldots, T$, the player makes a prediction p_t of the outcome, and then nature may choose the outcome Y_t. Without giving the player a bit more leeway, calibration is impossible: say that $Y \in \{0, 1\}$, and nature plays $Y_t = 1$ if $p_t \le .5$ and $Y_t = 0$ if $p_t > .5$. Then any player is miscalibrated at least by an amount .5. Astoundingly, Foster and Vohra [84] show that if the player is allowed to randomize, then the forecasted probabilities p_t can be made arbitrarily close to the empirical averages of the observed Y_t. While many of the techniques we consider and develop arise from this adversarial setting in the literature, we shall mostly address the scenarios in which Y is indeed random.

12.1 Proper losses and calibration error

When we use a proper loss to measure the error $\ell(f(x), y)$ in making the prediction $f(x)$ for the value y, it turns out we can *always* improve the losses by modifying f to be a calibrated version of itsef: calibration is always useful. To make this precise, assume we are making predictions in the convex hull of $\mathcal{Y}$, that is, that can be represented as $\mathbb{E}[Y]$ for some distribution, so $f : \mathcal{X} \to \mathcal{M} = \mathrm{Conv}(\mathcal{Y})$.

Then by Theorems 11.2.1 and 11.2.14, there exists a convex function h such that

$$\ell(\mu, y) = -h(\mu) - \langle \nabla h(\mu), y - \mu \rangle \tag{12.1.1}$$

for all $\mu \in \mathcal{M}, y \in \mathcal{Y}$. Recall the Bregman divergence (11.2.2)

$$D_h(u, v) = h(u) - h(v) - \langle \nabla h(v), u - v \rangle,$$

which is nonnegative for all convex h (and strictly positive whenever h is strictly convex and $u \neq v$), and Corollary 11.2.5. Then for any prediction function f, if we condition on the predicted value $S = f(X)$, then

$$\begin{aligned}\mathbb{E}[\ell(S, Y) \mid S] &= \mathbb{E}[\ell(\mathbb{E}[Y \mid S], Y) \mid S] + \mathbb{E}[\ell(S, Y) - \ell(\mathbb{E}[Y \mid S], Y) \mid S] \\ &= \mathbb{E}[\ell(\mathbb{E}[Y \mid S], Y) \mid S] + \mathbb{E}[D_h(\mathbb{E}[Y \mid S], S) \mid S],\end{aligned}$$

where we use the linearity $\mathbb{E}[\ell(s, Y)] = \ell(s, \mathbb{E}[Y])$ for any distribution on Y and fixed $s \in \mathbb{R}^k$ in the second equality. We record this as a theorem.

Theorem 12.1.1. *Let ℓ be a proper loss with representation (12.1.1). Then for any $f : \mathcal{X} \to \mathbb{R}^k$,*

$$\mathbb{E}[\ell(f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] + \mathbb{E}[D_h(\mathbb{E}[Y \mid f(X)], f(X))].$$

In particular, the predictor $g : \mathbb{R}^k \to \mathbb{R}^k$ defined by

$$g(s) := \mathbb{E}[Y \mid f(X) = s]$$

is calibrated and satisfies

$$\mathbb{E}[\ell(g \circ f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] \le \mathbb{E}[\ell(f(X), Y)],$$

and the inequality is strict whenever f is not calibrated and ℓ is strictly proper.

Proof The first statement we have already proved. For the second, note that

$$g(s) = \mathbb{E}[Y \mid f(X) = s]$$

by construction of g, so that $\mathbb{E}[\ell(g \circ f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)]$. The inequality and its strictness are immediate because h is strictly convex if and only if ℓ is strictly proper. □

To interpret this result, it essentially says that if we can post-process f to make it calibrated, then we can only improve its risk, or expected loss, when ℓ is a proper loss. We can give an alternative version of Theorem 12.1.1, where we instead consider the conjugate linkages in Section 11.3, which can be useful when we wish to find f via convex optimization (instead of by directly minimizing a proper loss). To that end, assume that h is a strictly convex function, differentiable on the interior of its domain, satisfying the Legendre conditions (11.3.3), and define the surrogate loss (linked via duality and the negative generalized entropy h to ℓ)

$$\varphi(s, y) = h^*(s) - \langle s, y \rangle = \ell(\mathsf{pred}_h(s), y),$$

where $\ell(\mu, y) = -h(\mu) - \langle \nabla h(\mu), y - \mu \rangle$ and

$$\mathsf{pred}_h(s) = \operatorname*{argmin}_{\mu} \{-\langle s, \mu \rangle + h(\mu)\} = \nabla h^*(s).$$

Then we have the following decomposition of the population surrogate loss, which follows similarly to Theorem 12.1.1.

Theorem 12.1.2. *Let φ be the surrogate loss defined above. Then for any $f : \mathcal{X} \to \mathbb{R}^k$, we have*

$$\mathbb{E}[\varphi(f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] + \mathbb{E}\left[D_h(\mathbb{E}[Y \mid f(X)], \mathsf{pred}_h(f(X)))\right].$$

Proof The key is to rely on the duality relationships inherent in the definition of the surrogate $\varphi(s, y) = h^*(s) - \langle s, y\rangle$. We fix x and work in exclusively in the space of the scores (predictions) $s = f(x) \in \mathbb{R}^k$, as

$$\mathbb{E}[\varphi(f(X), Y) \mid X = x] = \varphi(f(x), \mathbb{E}[Y \mid X = x])$$

by definition. Let $\mu \in \mathcal{M} = \mathrm{Conv}(\mathcal{Y})$. Then $\varphi(s, \mu) = h^*(s) - \langle s, \mu\rangle$, and

$$\inf_s \varphi(s, \mu) = -\sup_s\{\langle s, \mu\rangle - h^*(s)\} = -h(\mu)$$

because h is (closed) convex. Additionally, if $\mu^*(s) = \nabla h^*(s) = \mathsf{pred}_h(s)$, then the conjugate duality relationships (11.1.4) guarantee $h^*(s) = \langle s, \mu^*(s)\rangle - h(\mu^*(s))$ and $s = \nabla h(\mu^*(s))$. Thus

$$\begin{aligned}\varphi(s, \mu) - \inf_{s'} \varphi(s, \mu) = h^*(s) - \langle s, \mu\rangle + h(\mu) &= h(\mu) - h(\mu^*(s)) - \langle s, \mu - \mu^*(s)\rangle \\ &= h(\mu) - h(\mu^*(s)) - \langle \nabla h(\mu^*), \mu - \mu^*(s)\rangle = D_h(\mu, \mu^*(s)).\end{aligned}$$

Taking the expectation over X and using the shorthand $S = f(X)$, we thus obtain

$$\begin{aligned}\mathbb{E}[\varphi(S, Y)] &= \mathbb{E}[\varphi(S, \mathbb{E}[Y \mid S])] \\ &= \mathbb{E}\left[\inf_s \varphi(s, \mathbb{E}[Y \mid S])\right] + \mathbb{E}\left[D_h(\mathbb{E}[Y \mid s], \mathsf{pred}_h(s))\right].\end{aligned}$$

Lastly, we use that $\ell(\mu, y) = -h(\mu) - \langle \nabla h(\mu), y - \mu\rangle$ is proper, so $\inf_s \varphi(s, \mu) = -h(\mu) = \ell(\mu, \mu)$, giving the first claim of the theorem. □

As in Theorem 12.1.1, Theorem 12.1.2 shows that calibrating a predictor f can only improve the surrogate loss associated with h. Any predictor $f : \mathcal{X} \to \mathbb{R}^k$ has unnecessary error arising from the average divergence of the prediction from being calibrated,

$$\mathbb{E}\left[D_h\left(\mathbb{E}[Y \mid f(X)], \mathsf{pred}_h(f(X))\right)\right].$$

In both cases, we see that any proper (or derived proper) loss has a natural decomposition into an error term relating to the typical error in predicting Y from $\mathbb{E}[Y \mid f(X)]$, which one frequently refers to as *sharpness* of the predictor. Replacing $f(X)$ with the expectation of Y given $f(X)$ (or a particular transformation thereof) does not increase this first term, but improves the second term, which measures the typical error of a prediction from calibration.

Let us consider an example with squared error:

Example 12.1.3 (Squared error and calibration)**:** In the case that $h(p) = \frac{1}{2}\|p\|_2^2$, we have $h^* = h$ and $\nabla h = \nabla h^*$ is the identity. Then Theorems 12.1.1 and 12.1.2 reduce to the statement that

$$\mathbb{E}[\|Y - f(X)\|_2^2] = \mathbb{E}\left[\|Y - \mathbb{E}[Y \mid X]\|_2^2\right] + \mathbb{E}\left[\|\mathbb{E}[Y \mid X] - f(X)\|_2^2\right],$$

so we may also see the decompositions of the theorems as bias/variance expansions. ◇

12.2 Measuring calibration

The first step to building a practicable theory of calibration is to define and measure the calibration of a predictor f. The first step, defining a calibrated predictor, is relatively easy, but measuring how "close" a particular predictor f is to being calibrated raises several challenges, as typical and naive measures of calibration are impossible to estimate. Thus, in this section, we develop several quantities to measure calibration, providing a main theorem relating the different quantities to one another and demonstrating a simple technique to estimate one of them, returning in Section 12.5.2 to show the equivalences between the measures.

We begin with a natural candidate for calibration: the expected difference, or expected calibration error,

$$\mathsf{ece}(f) := \mathbb{E}[\|\mathbb{E}[Y \mid f(X)] - f(X)\|]. \tag{12.2.1}$$

The calibration error (12.2.1) is 0 if and only if f is perfectly calibrated, as then $\mathbb{E}[Y \mid f(X)] = f(X)$, and it is positive otherwise. Unfortunately, while the next lemma guarantees that ece is lower semi-continuous, it is not continuous.

Lemma 12.2.1. *The expected calibration error* ece *is lower semi-continuous with respect to* $L^1(P)$ *on* $\mathcal{F}$*, that is, if* $\mathbb{E}[\|f_n(X) - f(X)\|] \to 0$ *and* $f \in L^1(P)$*, then*

$$\liminf_n \mathsf{ece}(f_n) \geq \mathsf{ece}(f).$$

This result requires some delicate measure-theoretic arguments, so we defer it to the technical proofs (see Section 12.6.1). The disctontinuity of ece is relatively easy to show, however, even in very simple cases.

Example 12.2.2 (Discontinuity of the calibration error)**:** Let $Y \in \{0, 1\}$ be a Bernoulli random variable, and let $X \in \{0, 1\}$. Take $Y = X$ with probability 1. Then the predictor that always predicts $\frac{1}{2}$ is perfectly calibrated, but if for $\epsilon \in [0, \frac{1}{2}]$ we define f_ϵ by

$$f_\epsilon(0) = \frac{1}{2} - \epsilon \quad \text{and} \quad f_\epsilon(1) = \frac{1}{2} + \epsilon$$

then we see that $\mathsf{ece}(f_\epsilon) = \frac{1}{2} - \epsilon$, while $\mathsf{ece}(f_0) = 0$. Certainly $f_\epsilon \to f_0$ in any L^p distance on functions, while $\lim_{\epsilon \to 0} \mathsf{ece}(f_\epsilon) = \frac{1}{2}$. $\diamond$

12.2.1 The impossibility of measuring calibration

The discontinuity Example 12.2.2 highlights suggests that estimating calibration $\mathsf{ece}(f)$ for a fixed function f should be nontrivial, and indeed, using the tools on functional estimation and testing we develop in Chapter 10, we can show strong lower bounds for estimating the calibration error unless one makes unjustifiable assumptions about the distribution of $Y \mid f(X)$. The precise reasons differ a bit from the discontinuity of $\mathsf{ece}(f)$ in f, though the intuition is relatively straightforward: if $f(X)$ has a density, then even given a very large sample (X_1^n, Y_1^n), all the observations $f(X_i)$ will be distinct, and we have no *a priori* reason to assume that $\mathbb{E}[Y \mid f(X)]$ should be continuous in the predicted value $f(X)$.

To make this more precise, fix a function f whose calibration error we wish to evaluate, and consider a hypothesis test of $H_0 : \mathsf{ece}(f) = 0$ against alternatives that f is miscalibrated, $H_1 : \mathsf{ece}(f) \geq \gamma$ for some $\gamma > 0$. We observe predictions $f(X_i)$ and outcomes Y_i, that is, pairs

$$Z_i = (f(X_i), Y_i)$$

drawn i.i.d.; the coming lower bound holds if $\mathcal{X} = [0,1]$ and $f(x) = x$, so in many cases, observing X is of no help. Recall the (worst-case) test risk from Section 10.2, that for the testing problem between classes $H_0 : P \in \mathcal{P}_0$ and $H_1 : P \in \mathcal{P}_1$ of distributions,

$$R_n(\Psi \mid \mathcal{P}_0, \mathcal{P}_1) := \sup_{P \in \mathcal{P}_0} P(\Psi(Z_1^n) \neq 0) + \sup_{P \in \mathcal{P}_1} P(\Psi(Z_1^n) \neq 1).$$

Because we consider the function f fixed and ask only whether we can evaluate its calibration error under an (unknown) distribution P, we denote the expected calibration error of f under P via $\mathsf{ece}_P(f) = \mathbb{E}_P[\|\mathbb{E}_P[Y \mid f(X)] - f(X)\|]$. We thus consider testing perfect calibration $H_0 : \mathsf{ece}(f) = 0$ against alternatives $H_1 : \mathsf{ece}(f) \geq \gamma$ of miscalibration for $\gamma > 0$, defining

$$\mathcal{P}_\gamma = \{\text{distributions } P \text{ on } (X, Y) \mid \mathsf{ece}_P(f) \geq \gamma\}$$

as the collection of distributions for which f is $(\frac{1}{2} - \gamma)$ mis-calibrated.

Theorem 12.2.3. *Let $f : \mathcal{X} \to [0,1]$ be a predictor of $Y \in \{0,1\}$. Assume for some $0 < c < \frac{1}{2}$ that $f(\mathcal{X}) \cap [c, 1-c]$ has cardinality at least N. Then there is a distribution P_0 such that $\mathsf{ece}_{P_0}(f) = 0$ and for any $0 < \gamma \leq c$,*

$$\inf_\Psi R_n(\Psi \mid \{P_0\}, \mathcal{P}_\gamma) \geq 1 - \frac{n\gamma^2}{2\sqrt{N}} \frac{1}{c(1-c)}.$$

Before proving Theorem 12.2.3, we note the following immediate corollary; part (ii) follows from part (i), which follows by taking $N \uparrow \infty$ in the theorem.

Corollary 12.2.4. *Let the conditions of Theorem 12.2.3 hold and let $\mathcal{P}_0 = \{P \mid \mathsf{ece}_P(f) = 0\}$.*

(i) If there exists $0 < c < \frac{1}{2}$ such that $f(\mathcal{X}) \cap [c, 1-c]$ has infinite cardinality, then $\mathcal{P}_0$ is non-empty and for any $0 < \gamma \leq c$,

$$\liminf_n \inf_\Psi R(\Psi \mid \mathcal{P}_0, \mathcal{P}_\gamma) = 1.$$

(ii) If there exists a neighborhood U of $\frac{1}{2}$ such that $U \subset f(\mathcal{X})$, then $\mathcal{P}_0$ is non-empty and for any $\gamma < \frac{1}{2}$, the minimax test risk satisfies

$$\liminf_n \inf_\Psi R(\Psi \mid \mathcal{P}_0, \mathcal{P}_\gamma) = 1.$$

In brief, no test exists that is better than random guessing for testing between

$$H_0 : \mathsf{ece}(f) = 0 \quad \text{and} \quad H_1 : \mathsf{ece}(f) \geq c$$

given access to the predictions $f(X_i)$ and observed outcomes Y_i. The theorem and corollary apply to binary prediction models with $Y \in \{0,1\}$, but the results immediately extend to more complicated prediction problems where Y is vector-valued or multiclass.

Proof The proof relies on the convex hull testing lower bound from Proposition 10.2.1. Without loss of generality, we can assume that $\mathcal{X} \subset [0,1]$ and that $f(x) = x$ by transforming the input space. Let $S = f(X)$ be the (random) scores that f outputs.

We first construct the perfectly calibrated distribution P_0 and miscalibrated family $\mathcal{P}_\gamma$. Define the distribution P_0 so that S is uniform on distinct points $s_1, \ldots, s_N \in [c, 1-c]$ and $Y \mid S = s \sim \mathsf{Bernoulli}(s)$, that is, given $S = s$, $Y = 1$ with probability s and $Y = 0$ with probability $1 - s$. By

construction, $\mathsf{ece}_{P_0}(f) = 0$. To construct the particular members of the alternative family $\mathcal{P}_\gamma$, for each $j \in [N]$, define the "tilting" function

$$\phi_j(y, s) := \left(\frac{y}{s_j} - \frac{1-y}{1-s_j}\right) \mathbf{1}\left\{s = s_j\right\}.$$

Then $\mathbb{E}_0[\phi_j(Y, S)] = 0$ while

$$\mathrm{Var}_0(\phi_j(Y, S)) = \frac{1}{N}\mathbb{E}_0\left[\left(\frac{Y}{s_j} - \frac{1-Y}{s_j}\right)^2 \mid S = s_j\right] = \frac{1}{N}\left(\frac{1}{s_j} + \frac{1}{1-s_j}\right) = \frac{1}{N}\frac{1}{s_j(1-s_j)}.$$

Note that $|\phi_j(y, s)| \le \frac{1}{c}$ as $c < \frac{1}{2}$, and if we define the vector $\phi(y, s) = (\phi_1(y, s), \ldots, \phi_N(y, s))$, then $\|\phi(y, s)\|_0 \le 1$ (that is, the number of non-zero entries is at most 1). Now as $\gamma \in [0, c]$, for each $v \in \{-1, 1\}^N$ we may define the tilted distribution P_v with

$$P_v(Y = y, S = s) = (1 + \gamma\langle v, \phi(y, s)\rangle) P_0(Y = y, S = s),$$

which is a valid distribution whenever $\gamma \le c$, as $|\langle v, \phi(y, s)\rangle| \le \frac{1}{c}$. We compute the calibration error for distributions $P \in \{P_v\}$. Noting that S is still uniform on $\{s_1, \ldots, s_N\}$ under P_v, we have

$$\mathbb{E}_v[Y \mid S = s_j] = s_j + \gamma v_j \mathbb{E}[\phi_j(Y, s_j) Y \mid S = s_j] = s_j + \gamma v_j,$$

and so $\mathsf{ece}_{P_v}(f) = \frac{1}{N}\sum_{j=1}^N \gamma |v_j| = \gamma$. In particular, we have $P_v \in \mathcal{P}_\gamma$.

Lastly, we compute a bound on the testing error. For this, we recall Lemma 10.1.3. Letting $\overline{P^n} = \frac{1}{2^N}\sum_v P_v^n$, we have

$$\begin{aligned} D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) + 1 &= \frac{1}{2^{2N}} \sum_{v, v'} \mathbb{E}_0\left[(1 + \gamma\langle v, \phi(Y, S)\rangle)(1 + \gamma\langle v', \phi(Y, S)\rangle)\right]^n \\ &= \frac{1}{2^{2N}} \sum_{v, v'} \left(1 + \gamma^2 v^\top \mathrm{Cov}_0(\phi(Y, S)) v'\right)^n \end{aligned}$$

because the sampling is i.i.d. By our variance calculation for ϕ and that each ϕ_j has disjoint support, we have $\mathrm{Cov}_0(\phi(Y, S)) = \frac{1}{N}\mathrm{diag}([\frac{1}{s_j(1-s_j)}]_{j=1}^N)$, and so

$$D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) + 1 = \mathbb{E}\left[\left(1 + \frac{\gamma^2}{N}\sum_{j=1}^N \frac{V_j V_j'}{s_j(1-s_j)}\right)^n\right] \le \mathbb{E}\left[\exp\left(\frac{n\gamma^2}{N}\sum_{j=1}^N \frac{V_j V_j'}{s_j(1-s_j)}\right)\right]$$

where the expectation is over $V, V' \stackrel{\mathrm{iid}}{\sim} \mathsf{Uniform}(\{\pm 1\}^N)$. But of course $V_j V_j'$ i.i.d. random signs, and hence 1-sub-Gaussian, so that

$$D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) + 1 \le \exp\left(\frac{n^2\gamma^4}{N^2}\sum_{j=1}^N \frac{1}{s_j^2(1-s_j)^2}\right) \le \exp\left(\frac{n^2\gamma^4}{2N}\frac{1}{c^2(1-c)^2}\right)$$

because $c \le s_j \le 1 - c$. Apply Proposition 10.2.1 and Pinsker's inequality (Propositions 2.2.8 and 2.2.9) to see that

$$\inf_\Psi R(\Psi \mid \{P_0\}, \mathcal{P}_\gamma) \ge 1 - \sqrt{\frac{1}{2}\log\left(1 + D_{\chi^2}\left(\overline{P^n} \| P_0\right)\right)} \ge 1 - \sqrt{\frac{n^2\gamma^4}{4N}\frac{1}{c^2(1-c)^2}}.$$

Taking square roots gives the result. □

12.2.2 Alternative calibration measures

The fundamental impossibility results in Theorem 12.2.3 and Corollary 12.2.4, even in the binary prediction case, suggest that we should choose some more easily estimable measure for calibration. In Section 12.5 we provide formal definitions for calibration measures to be continuous (or Lipschitz continuous) and equivalent to one another. Here, we provide the alternative definitions of calibration we consider, giving a corollary that captures their relationships for multiclass classification, and then describing how to estimate one of them. Let us take the general setting of this chapter, where the label space $\mathcal{Y} \subset \mathbb{R}^k$ and P is a distribution on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{F}$ be a collection of functions mapping $\mathcal{X} \to \mathbb{R}^k$ and integrable with respect to P, that is, $\mathbb{E}[\|f(X)\|] < \infty$ for each $f \in \mathcal{F}$.

In brief, we require that a calibration measure $\mathrm{M} : \mathcal{F} \to \mathbb{R}_+$ be *sound* (in analogy with proof systems, where soundness means nothing false can be proved), meaning that

$$\mathrm{M}(f) = 0 \;\; \text{implies} \;\; \mathbb{E}[Y \mid f(X)] = f(X) \tag{12.2.2a}$$

and *complete* (continuing the analogy, that everything true can be proved), meaning that

$$\mathbb{E}[Y \mid f(X)] = f(X) \;\; \text{implies} \;\; \mathrm{M}(f) = 0. \tag{12.2.2b}$$

We begin by considering types of *distance to calibration*. Let $\mathcal{C}(P)$ denote those functions g that are perfectly calibrated for P, that is, $\mathcal{C}(P) = \{g : \mathcal{X} \to \mathbb{R}^k \mid \mathbb{E}_P[Y \mid g(X)] = g(X)\}$ (where the defining equality holds with P-probability 1 over X). The set $\mathcal{P}$ always consists at least of the constant function $g(X) = \mathbb{E}_P[Y]$ and so is non-empty (but is typically larger). Then we call the minimum $L^1(P)$ distance of a function f to the set $\mathcal{C}(P)$ the *distance to calibration*

$$d_{\mathrm{cal}}(f) := \inf_g \left\{\mathbb{E}[\|g(X) - f(X)\|] \text{ s.t. } g \in \mathcal{C}(P)\right\}. \tag{12.2.3}$$

It is not always clear how to estimate the distance $d_{\mathrm{cal}}(f)$, making using it sometimes challenging.

We also consider a complementary quantity that relies on an alternative variational characterization. Let $\mathcal{W} \subset \{\mathbb{R}^k \to \mathbb{R}^k\}$ be a symmetric collection of functions, meaning that $w \in \mathcal{W}$ implies $-w \in \mathcal{W}$. We can view any such collection as potential witnesses of miscalibration, in that

$$\mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] = \mathbb{E}[\langle w(f(X)), \mathbb{E}[Y \mid f(X)] - f(X)\rangle]$$

and so if w can "witness" the portions of space where $f(X) \not\approx \mathbb{E}[Y \mid f(X)]$, it can certify miscalibration. We then arrive at what we term the *calibration error relative to the class* $\mathcal{W}$,

$$\mathrm{CE}(f, \mathcal{W}) := \sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle]. \tag{12.2.4}$$

Depending on the class $\mathcal{W}$, this is sometimes called the *weak calibration error*, and with large enough classes, we can recover the classical expected calibration error (12.2.1).

Example 12.2.5 (Recovering expected calibration error)**:** For a norm $\|\cdot\|$, let the set $\mathcal{W}$ be the collection of all functions w with bound $\sup_s \|w(s)\|_* \le 1$. Then

$$\mathrm{CE}(f, \mathcal{W}) = \mathbb{E}\left[\sup_{\|w\|_* \le 1} \langle w, \mathbb{E}[Y \mid f(X)] - f(X)\rangle\right] = \mathbb{E}[\|\mathbb{E}[Y \mid f(X)] - f(X)\|] = \mathsf{ece}(f),$$

the expected calibration error. $\diamond$

It is more interesting to consider restricted classes; one of particular interest to us is that of bounded Lipschitz functions. Let

$$\mathcal{W}_{\|\cdot\|} := \left\{w : \mathbb{R}^k \to \mathbb{R}^k \mid \|w(s_0) - w(s_1)\|_* \le \|s_0 - s_1\| \text{ and } \|w(s)\|_* \le 1 \text{ for all } s, s_0, s_1\right\} \quad (12.2.5)$$

denote the collection of functions bounded by 1 in $\|\cdot\|_*$ and that are 1-Lipschitz with respect to $\|\cdot\|$. Then (as we see presently) we can at least estimate the calibration error relative to the class $\mathcal{W}$ in the definition (12.2.4).

The final calibration measure we consider reposes on the idea of quantizing or partitioning the output space, which relates to the idea of "binning" predictions that the literature on calibration frequently considers. Here, we consider averages of Y conditioned on predictions in larger sets. Thus, instead of evaluating the precise conditioning $\mathbb{E}[Y \mid f(X)]$ we to look instead at the expectation of Y conditional on $f(X) \in A$ for a set A, so that a predicted score is (nearly) calibrated if the diameter $\operatorname{diam}(A)$ is small, and $\mathbb{E}[Y \mid f(X) \in A] \approx s$ for some $s \in A$. Given a partition $\mathcal{A}$ of the space $\mathcal{M} = \operatorname{Conv}(\mathcal{Y})$, it is then natural to evaluate the average error for each element of A (weighting by the probability of A), and consider the calibration error (12.2.4) for indicator functions of $A \in \mathcal{A}$, where we abuse notation slightly to define

$$\operatorname{CE}(f, \mathcal{A}) := \sum_{A \in \mathcal{A}} \|\mathbb{E}[(f(X) - Y)\mathbf{1}\left\{f(X) \in A\right\}]\| = \sum_{A \in \mathcal{A}} \|\mathbb{E}[f(X) - Y \mid f(X) \in A]\| \, \mathbb{P}(f(X) \in A).$$

Indeed, taking a supremum over all such partitions gives $\sup_{\mathcal{A}} \operatorname{CE}(f \mid \mathcal{A}) = \mathbb{E}[\|\mathbb{E}[Y \mid f(X)] - f(X)\|]$, the original expected calibration error (12.2.1). Additionally, and here we elide details, if $f(X)$ is a continuous random variable with suitably nice density and $\mathcal{A}_n$ denotes any partition satisfying $\operatorname{diam}(A) \le 1/n$ for $A \in \mathcal{A}_n$, then $\lim_n \operatorname{CE}(f, \mathcal{A}_n) = \mathbb{E}[\|\mathbb{E}[Y \mid f(X)] - f(X)\|]$. Instead of considering $\operatorname{CE}(f, \mathcal{A})$ directly, we optimize over all partitions, but penalize the average size of elements of $\mathcal{A}$, giving the *partitioned calibration error*

$$\mathsf{pce}(f) := \inf_{\mathcal{A}} \left\{\operatorname{CE}(f, \mathcal{A}) + \sum_{A \in \mathcal{A}} \operatorname{diam}(A) \mathbb{P}(f(X) \in A)\right\}. \quad (12.2.6)$$

Each of these is equivalent to within polynomial scaling.

Corollary 12.2.6. *Let $\mathcal{Y} \subset \mathbb{R}^k$ have finite diameter and $\|\cdot\|$ be any norm. Then each of the calibration measures d_{cal}, $\operatorname{CE}(\cdot, \mathcal{W}_{\|\cdot\|})$, and pce in definitions (12.2.3), (12.2.4), and (12.2.6) is sound and complete (12.2.2). Additionally, let $\mathcal{Y} = \{e_1, \ldots, e_k\}$ and $\|\cdot\| = \|\cdot\|_1$ be the ℓ_1-norm. Then for any $f : \mathcal{X} \to \mathcal{M} = \operatorname{Conv}(\mathcal{Y})$,*

$$\frac{1}{2}\operatorname{CE}(f, \mathcal{W}_{\|\cdot\|}) \le d_{\text{cal}}(f) \le \operatorname{CE}(f, \mathcal{W}_{\|\cdot\|}) + 2\sqrt{k \operatorname{CE}(f, \mathcal{W}_{\|\cdot\|})}$$

and

$$d_{\text{cal}}(f) \le \mathsf{pce}(f) \le d_{\text{cal}}(f) + 2\sqrt{k \, d_{\text{cal}}(f)}.$$

Corollary 12.2.6 will come as a consequence of the deeper development we purse in Section 12.5.

Here, we take Corollary 12.2.6 as motivation to give the type of typical result that justifies calibration estimates. As any of the calibration measures is roughly equivalent (except ece), measuring any of them on a sample can provide evidence for or against calibration of a predictor f. We focus

on the simpler binary case in which $f : \mathcal{X} \to [0,1]$ and let $\mathcal{W}_{\rm Lip}$ be bounded Lipschitz functions $w : [0,1] \to [-1,1]$. Given a sample (X_1^n, Y_1^n), the empirical variant of $\mathrm{CE}(f, \mathcal{W})$ is

$$\widehat{\mathrm{CE}}_n(f) := \sup_{\|w\|_\infty \le 1} \left\{ \frac{1}{n} \sum_{i=1}^n w_i (Y_i - f(X_i)) \text{ s.t. } |w_i - w_j| \le |f(X_i) - f(X_j)| \text{ for } i, j \le n \right\}.$$

By combining uniform covering bounds for the class of Lipschitz functions with a standard concentration inequality, we then have the following convergence guarantee for $\widehat{\mathrm{CE}}_n$.

Proposition 12.2.7. *There exists a numerical constant C such that for any $\delta > 0$,*

$$\left| \widehat{\mathrm{CE}}_n(f) - \mathrm{CE}(f, \mathcal{W}_{\rm Lip}) \right| \le C \frac{\sqrt{\log \frac{n}{\delta}}}{n^{1/3}}$$

with probability at least $1 - \delta$.

Proof Fix $\epsilon > 0$ and let $\mathcal{N}(\epsilon)$ be a minimal ϵ-cover of the set $\mathcal{W}_{\rm Lip}$ in uniform norm, meaning that $\left\|w - w^{(j)}\right\|_\infty \le \epsilon$ for each $w^{(j)} \in \mathcal{N}(\epsilon)$, and let $N(\epsilon)$ be its (minimal) cardinality. Then $\log N(\epsilon) \lesssim \frac{1}{\epsilon} \log \frac{1}{\epsilon}$ (recall Proposition 8.7.3 and Eq. (8.7.4)). For shorthand, let the error vector $E \in [-1,1]^n$ have entries $E_i = Y_i - f(X_i)$, and abusing notation, for $w \in \mathcal{W}_{\rm Lip}$ let $\langle w, E \rangle_n = \frac{1}{n} \sum_{i=1}^n w(f(X_i)) E_i$. Then for any $w \in \mathcal{W}_{\rm Lip}$, there exists $i \le N(\epsilon)$ such that

$$|\langle w, E \rangle_n - \langle w^{(i)}, E \rangle_n| \le \epsilon,$$

while $\widehat{\mathrm{CE}}_n(f) = \sup_{w \in \mathcal{W}_{\rm Lip}} \langle w, E \rangle_n$. In particular, we have

$$\left| \widehat{\mathrm{CE}}_n(f) - \mathrm{CE}(f, \mathcal{W}_{\rm Lip}) \right| \le \sup_{w \in \mathcal{W}_{\rm Lip}} |\langle w, E \rangle_n - \mathbb{E}[\langle w, E \rangle_n]| \le \max_{w \in \mathcal{N}(\epsilon)} |\langle w, E \rangle_n - \mathbb{E}[\langle w, E \rangle_n]| + 2\epsilon.$$

Thus for any $t \ge 0$, we have

$$\begin{aligned} \mathbb{P}\left(\left| \widehat{\mathrm{CE}}_n(f) - \mathrm{CE}(f, \mathcal{W}_{\rm Lip}) \right| \ge t \right) &\le \mathbb{P}\left(\max_{w \in \mathcal{N}(\epsilon)} |\langle w, E \rangle_n - \mathbb{E}[\langle w, E \rangle_n]| \ge t - 2\epsilon \right) \\ &\le 2N(\epsilon) \exp\left(-\frac{n \left[t - 2\epsilon\right]_+^2}{2} \right) \end{aligned}$$

by the Azuma-Hoeffding inequality and a union bound. Take $\epsilon = n^{-1/3}$ and $t = C n^{-1/3} \sqrt{\log \frac{n}{\delta}}$ for an appropriate numerical constant C to obtain the proposition. □

Summarizing, while the expected calibration error is fundamentally inestimable, there are alternative measures that are both sound and complete, and they can admit reasonable estimators. As the class size k grows, however, it can become statistically infeasible to estimate the calibration of predictors f, so that one must consider alternative metrics. The exercises and bibliography explore these questions in more detail.

12.3 Auditing and improving calibration at the population level

Theorems 12.1.1 and 12.1.2 provide decompositions of the expected loss of a predictor

$$\mathbb{E}[\ell(f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] + \mathbb{E}\left[D_h\left(\mathbb{E}[Y \mid f(X)], f(X)\right)\right]$$

into an average loss and an expected divergence between $f(X)$ and $\mathbb{E}[Y \mid f(X)]$, where h is the negative (generalized) entropy (11.1.6) associated with the loss ℓ, so that the loss has representation $\ell(\mu, y) = -h(\mu) - \langle \nabla h(\mu), y - \mu \rangle$. This suggests an approach to improving a predictor $f : \mathcal{X} \to \mathbb{R}^k$ without compromising its average loss: make it closer to being calibrated, so that $\mathbb{E}[Y \mid f(X)] \approx f(X)$. Here, we make this idea precise by using the weak calibration (12.2.4): if there exists a witness function w certifying that $\mathbb{E}[\langle w(f(X)), Y - f(X) \rangle] \gg 0$, then we can post-process f to $f(X) + \eta w(f(X))$ for some stepsize $\eta > 0$ and *only* improve the expected loss. We first develop the idea in the context of the squared error, where the calculations are cleanest, and extend it to general proper losses based on convex conjugates (as in Section 11.3) immediately after. Combining the ideas we develop, we also provide a (population-level) algorithm to transform a function f by post-processing its outputs that guarantees the result is nearly calibrated relative to a class $\mathcal{W}$ of witnesses. This provides an algorithmic proof quantitatively relating the calibration error $\mathrm{CE}(f, \mathcal{W})$ relative to a class $\mathcal{W}$ to the improvement achievable in minimizing $\mathbb{E}[\ell(f(X), Y)]$ by post-composition $g \circ f$.

12.3.1 The post-processing gap and calibration audits for squared error

Consider a thought experiment: instead of using f to make predictions, we use a postprocessing $g \circ f$, where $g : \mathbb{R}^k \to \mathbb{R}^k$ has the (suggestively chosen) form $g(v) = v + w(v)$, where $w(v) = (g(v) - v)$. Then using the representation $\ell(\mu, y) = -h(\mu) - \langle \nabla h(\mu), y - \mu \rangle$ for the proper loss, we recall Theorem 12.1.1 and for $\mu(f(X)) := \mathbb{E}[Y \mid f(X)]$ expand

$$\begin{aligned}
\mathbb{E}[\ell(g \circ f(X), Y)] &= \mathbb{E}[-h(g \circ f(X)) - \langle \nabla h(g \circ f(X)), Y - g \circ f(X) \rangle] \\
&= \mathbb{E}[-h(\mu(f(X)))] + \mathbb{E}[h(\mu(f(X))) - \langle \nabla h(g \circ f(X)), Y - g \circ f(X) \rangle] \\
&= \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] + \mathbb{E}[D_h(\mathbb{E}[Y \mid f(X)], g \circ f(X))],
\end{aligned}$$

where the final equality uses the linearity of $y \mapsto \ell(\mu, y)$, that is,

$$\mathbb{E}[\ell(g \circ f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] + \mathbb{E}\left[D_h\left(\mathbb{E}[Y \mid f(X)], f(X) + w(f(X))\right)\right]. \tag{12.3.1}$$

We have decomposed the expected loss $\mathbb{E}[\ell(g \circ f(X), Y)]$ into a term that post-processing does not change, which measures the sharpness with which $\mathbb{E}[Y \mid f(X)]$ predicts Y, and a divergence term D_h measuring the error in calibration of $g \circ f(X) = f(X) + w(f(X))$ for $\mathbb{E}[Y \mid f(X)]$.

The expansion (12.3.1) points toward an ability to postprocess *any* prediction function $f : \mathcal{X} \to \mathbb{R}^k$ to both (i) obtain calibration relative to a class of functions $\mathcal{W}$, as in Definition (12.2.4), and (ii) improve the expected loss $\mathbb{E}[\ell(f(X), Y)]$. Moreover, this improvement is monotone, in that changes "toward" calibration guarantee smaller expected loss, an improvement over the less refined results in Theorems 12.1.1 and 12.1.2. To that end, define the *post-processing gap* for the (proper) loss ℓ and function f relative to the class $\mathcal{W}$ of functions $\mathbb{R}^k \to \mathbb{R}^k$ by

$$\mathsf{gap}(\ell, f, \mathcal{W}) := \mathbb{E}[\ell(f(X), Y)] - \inf_{w \in \mathcal{W}} \mathbb{E}[\ell(f(X) + w(f(X)), Y)]. \tag{12.3.2}$$

The gap (12.3.2) is fundamentally tied to the calibration error relative to the class $\mathcal{W}$.

We specialize here to the simpler case of the squared error, as the statements are most transparent. We focus exclusively on symmetric convex collections of functions $\mathcal{W}$, meaning that if $w \in \mathcal{W}$, then $-w \in \mathcal{W}$, and $\mathcal{W}$ is convex.

Proposition 12.3.1. *Let $\ell(\mu, y) = \frac{1}{2}\|y - \mu\|_2^2$ be the squared error (Brier score), and let $\mathcal{W}$ be a symmetric convex collection of functions, each 1-Lipschitz with respect to the ℓ_2-norm $\|\cdot\|_2$. Define $R^2(f) = \sup_{w \in \mathcal{W}} \mathbb{E}[\|w(f(X))\|_2^2]$. Then*

$$\frac{1}{2} \min \left\{ \mathrm{CE}(f, \mathcal{W}), \frac{\mathrm{CE}(f, \mathcal{W})^2}{R^2(f)} \right\} \le \mathsf{gap}(\ell, f, \mathcal{W}) \le \mathrm{CE}(f, \mathcal{W})$$

Proof Fix x and let $\mu = \mathbb{E}[Y \mid f(X) = f(x)] \in \mathrm{Conv}(\mathcal{Y})$ and $w = w(f(x))$ be a potential update to $f(x)$. Then because $\ell(\mu, y) = \frac{1}{2}\|\mu - y\|_2^2$, for any $y \in \mathcal{Y}$

$$\ell(\mu, y) + \langle \nabla \ell(\mu, y), w \rangle + \frac{1}{2}\|w\|^2 = \ell(\mu + w, y).$$

Recognizing that $\nabla \ell(\mu, y) = (\mu - y)$, for any $w \in \mathcal{W}$ we therefore have

$$\begin{aligned} -\mathbb{E}[\langle f(X) - Y, w(f(X)) \rangle] - \frac{1}{2}\mathbb{E}[\|w(f(X))\|_2^2] &\le \mathbb{E}[\ell(f(X), Y)] - \mathbb{E}[\ell(f(X) + w(f(X)), Y)] \\ &\le -\mathbb{E}[\langle f(X) - Y, w(f(X)) \rangle]. \end{aligned}$$

Taking suprema over w on each side of the preceding inequalities and using the symmetry of $\mathcal{W}$ gives

$$\begin{aligned} \sup_{w \in \mathcal{W}} \left\{ \mathbb{E}[\langle f(X) - Y, w(f(X)) \rangle] - \frac{1}{2}\mathbb{E}[\|w(f(X))\|_2^2] \right\} &\le \mathsf{gap}(\ell, f, \mathcal{W}) \\ &\le \sup_{w \in \mathcal{W}} \mathbb{E}[\langle f(X) - Y, w(f(X)) \rangle]. \end{aligned}$$

Because $\mathrm{CE}(f, \mathcal{W}) = \sup_{w \in \mathcal{W}} \mathbb{E}[\langle f(X) - Y, w(f(X)) \rangle]$, we can use the convexity of $\mathcal{W}$ and the definition $R^2(f) := \sup_{w \in \mathcal{W}} \mathbb{E}[\|w(f(X))\|_2^2]$ to see that for any $\eta \in [0, 1]$, we may replace w with $\eta \cdot w \in \mathcal{W}$, and we have

$$\sup_{\eta \in [0,1]} \left[\eta \mathrm{CE}(f, \mathcal{W}) - \frac{\eta^2}{2} R^2(f) \right] \le \mathsf{gap}(\ell, f, \mathcal{W}) \le \mathrm{CE}(f, \mathcal{W}).$$

Maximizing over η on the left side, we choose $\eta = \min\{1, \frac{\mathrm{CE}(f, \mathcal{W})}{R^2(f)}\}$ to obtain the proposition. □

As an immediate corollary, we see that if $\mathcal{W} = \mathcal{W}_{\|\cdot\|_2}$ consists of the 1-Lipschitz functions with $\|w(\cdot)\|_2 \le 1$, we have a cleaner guarantee.

Corollary 12.3.2. *Let $\mathcal{W} = \mathcal{W}_{\|\cdot\|_2}$ and the conditions of Proposition 12.3.1 hold. Then*

$$\frac{1}{2 \operatorname{diam}(\mathcal{Y})^2} \mathrm{CE}(f, \mathcal{W})^2 \le \mathsf{gap}(\ell, f, \mathcal{W}) \le \mathrm{CE}(f, \mathcal{W}).$$

Thus, the calibration error upper and lower bounds the gap between the expected loss of f and a post-processed version of f. This yields a nearly operational interpretation of the calibration error relative to the class $\mathcal{W}$: it is, to within a square, exactly the amount we could improve the expected loss of the function f by postprocessing f itself.

12.3.2 Calibration audits for losses based on conjugate linkages

Recall as in Section 11.3.1 that, by a transformation tied to the loss ℓ via its associated generalized negative entropy, we may define the surrogate

$$\varphi(s, y) := h^*(s) - \langle s, y \rangle,$$

and we may transform arbitrary scores $s \in \mathbb{R}^k$ to predictions via the conjugate link (11.3.1), that is,

$$\mathsf{pred}_h(s) = \operatorname*{argmin}_\mu \left\{ -\langle s, \mu \rangle + h(\mu) \right\} = \nabla h^*(s).$$

So long as h is appropriately smooth, these satisfy $\ell(\mathsf{pred}_h(s), y) = \varphi(s, y)$. In complete analogy with the post-processing gap (12.3.2) when we assume f makes predictions in (the affine hull of) $\mathcal{Y}$, we can define the *surrogate post-processing gap*

$$\mathsf{gap}(\varphi, f, \mathcal{W}) := \mathbb{E}[\varphi(f(X), Y)] - \inf_{w \in \mathcal{W}} \mathbb{E}[\varphi(f(X) + w(f(X)), Y)]. \tag{12.3.3}$$

In spite of the similarity with definition (12.3.2), the actual predictions of Y from f in this case come via the link $\mathsf{pred}_h(f(X))$. Thus, in this case we instead consider the calibration error relative to a class $\mathcal{W}$ but after the composition of f with $\mathsf{pred}_h = \nabla h^*$, so that

$$\mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W}) = \sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(f(X)), Y - \mathsf{pred}_h(f(X)) \rangle] = \sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(f(X)), Y - \nabla h^*(f(X)) \rangle],$$

where as always we assume that the class of witness functions satisfies $\mathcal{W} = -\mathcal{W}$. When the prediction function is continuous enough in s, we can give an analogue of Proposition 12.3.1 to the more general surrogate case. To that end, we assume that the conjugate h^* has Lipschitz continuous gradient with respect to the dual norm $\|\cdot\|_*$, meaning that

$$\|\nabla h^*(s_0) - \nabla h^*(s_1)\| \le \|s_0 - s_1\|_*$$

for all $s_0, s_1 \in \mathbb{R}^k$. This is equivalent (see Proposition C.2.7) to the negative entropy h being strongly convex with respect to the norm $\|\cdot\|$, and also immediately implies that

$$\varphi(s + w, y) \le \varphi(s, y) + \langle \nabla_s \varphi(s, y), w \rangle + \frac{\|w\|_*^2}{2}.$$

Example 12.3.3 (Multiclass logistic regression)**:** For multiclass logistic regression, where we take $h(p) = \sum_{j=1}^k p_j \log p_j$, we know that h is strongly convex with respect to the ℓ_1 norm (this is Pinsker's inequality; see inequality (2.2.11)). Thus, the conjugate $h^*(s) = \log(\sum_{j=1}^k e^{s_j})$ has Lipschitz gradient with respect to the ℓ_∞ norm, meaning that for the prediction link

$$\mathsf{pred}_h(s) = \left[\frac{e^{s_y}}{\sum_{j=1}^k e^{s_j}} \right]_{y=1}^k,$$

we have

$$\left\| \mathsf{pred}_h(s) - \mathsf{pred}_h(s') \right\|_1 \le \left\| s - s' \right\|_\infty$$

for all $s, s' \in \mathbb{R}^k$. $\diamond$

Example 12.3.4 (The squared error)**:** When we measure the error of a predictions in $\mathbb{R}^k$ by the squared ℓ_2-norm $\frac{1}{2}\|f(x) - y\|_2^2$, this corresponds to the generalized negative entropy $h(\mu) = \frac{1}{2}\|\mu\|_2^2$. In this case, the norm $\|\cdot\| = \|\cdot\|_2 = \|\cdot\|_*$, and we have the self duality $h^* = h$, so that the prediction mapping pred_h is the identity. $\diamond$

With these examples as motivation, we then have the following generalization of Proposition 12.3.1.

Proposition 12.3.5. *Let the negative generalized entropy h be strongly convex with respect to the norm $\|\cdot\|$ and consider surrogate loss $\varphi(s, y) = h^*(s) - \langle s, y\rangle$. Define $R_*^2(f) := \sup_{w\in\mathcal{W}} \mathbb{E}[\|w(f(X))\|_*^2]$. Then*

$$\frac{1}{2}\min\left\{\mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W}), \frac{\mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W})^2}{R_*^2(f)}\right\} \le \mathsf{gap}(\varphi, f, \mathcal{W}) \le \mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W}).$$

Proof Fix x and let $s = f(x)$ and $w = w(f(x))$, and notice that for any y we have

$$\varphi(s, y) + \langle \nabla\varphi(s, y), w\rangle \le \varphi(s + w, y) \le \varphi(s, y) + \langle \nabla\varphi(s, y), w\rangle + \frac{1}{2}\|w\|_*^2.$$

Recognizing that $\nabla\varphi(s, y) = \nabla h^*(s) - y$, for any $w \in \mathcal{W}$ we have

$$\begin{aligned}-\mathbb{E}[\langle \nabla\varphi(f(X), Y), w(f(X))\rangle] - \frac{1}{2}\mathbb{E}[\|w(f(X))\|_*^2] &\le \mathbb{E}[\varphi(f(X), Y)] - \mathbb{E}[\varphi(f(X) + w(f(X)), Y)] \\ &\le -\mathbb{E}[\langle \nabla\varphi(f(X), Y), w(f(X))\rangle].\end{aligned}$$

Taking suprema over w on each side and using the symmetry of $\mathcal{W}$ gives

$$\begin{aligned}\sup_{w\in\mathcal{W}}\left\{\mathbb{E}[\langle \nabla h^*(f(X)) - Y, w(f(X))\rangle] - \frac{1}{2}\mathbb{E}[\|w(f(X))\|_*^2]\right\} &\le \mathsf{gap}(\varphi, f, \mathcal{W}) \\ &\le \sup_{w\in\mathcal{W}} \mathbb{E}[\langle \nabla h^*(f(X)) - Y, w(f(X))\rangle].\end{aligned}$$

Because $\mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W}) = \sup_{w\in\mathcal{W}} \mathbb{E}[\langle \nabla h^*(f(X)) - Y, w(f(X))\rangle]$, we can use the convexity of $\mathcal{W}$ and the definition $R_*^2(f) := \sup_{w\in\mathcal{W}} \mathbb{E}[\|w(f(X))\|_*^2]$, to see that for any $\eta \in [0, 1]$, we may replace w with $\eta \cdot w \in \mathcal{W}$ and

$$\sup_{\eta\in[0,1]}\left[\eta\mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W}) - \frac{\eta^2}{2}R_*^2(f)\right] \le \mathsf{gap}(\varphi, f, \mathcal{W}) \le \mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W}).$$

Set $\eta = \min\{1, \frac{\mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W})}{R_*^2(f)}\}$. $\square$

A corollary specializing to the case of bounded witness functions allows a somewhat cleaner statement, in analogy with Corollary 12.3.2. It provides the same operational interpretation: the calibration error $\mathrm{CE}(f, \mathcal{W})$ of f relative to $\mathcal{W}$ upper and lower bounds improvement possible through postprocessing f.

Corollary 12.3.6. *Let the conditions of Proposition 12.3.5 hold, and additionally assume that the witness functions $\mathcal{W}$ satisfy $\|w(s)\|_* \le 1$ for all $s \in \mathbb{R}^k$. Then*

$$\frac{1}{2\,\mathrm{diam}(\mathrm{dom}\, h)^2}\mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W})^2 \le \mathsf{gap}(\varphi, f, \mathcal{W}) \le \mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W}).$$

We can give an alternative perspective for this section by focusing on the definitions (12.3.2) and (12.3.3) of the post-processing gap. Suppose we have a proper loss ℓ and we wish to improve the expected loss of a predictor f by post-processing f. When there is little to be gained by replacing f with an adjusted version $f(x) + w(f(x))$ for some $w \in \mathcal{W}$, then f *must be calibrated* with respect to the class $\mathcal{W}$. So, for example, for a surrogate φ, the function f (really, its associated prediction function $\mathsf{pred}_h \circ f$) is calibrated with respect to $\mathcal{W}$ if and only if $\mathbb{E}[\varphi(f(X) + w(f(X)), Y)] \le \mathbb{E}[\varphi(f(X), Y)]$ for all $w \in \mathcal{W}$.

As a particular special case to close this section, the standard multiclass logistic loss provides a clean example.

Example 12.3.7 (Multiclass logistic losses, continued)**:** Let h be the negative entropy $h(p) = \sum_{j=1}^k p_j \log p_j$ restricted to the probability simplex $\Delta_k = \{p \in \mathbb{R}_+^k \mid \langle \mathbf{1}, p \rangle = 1\}$ and the surrogate $\varphi(s, y) = \log(\sum_{j=1}^k e^{s_j}) - s_y$. Then for any class $\mathcal{W}$ consisting of functions with $\|w(s)\|_\infty \le 1$ for all $s \in \mathbb{R}^k$ and any function $f : \mathcal{X} \to \mathbb{R}^k$,

$$\frac{1}{2}\mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W})^2 \le \mathbb{E}[\varphi(f(X), Y)] - \inf_{w \in \mathcal{W}} \mathbb{E}[\varphi(f(X) + w(f(X)), Y)].$$

(Note that $\operatorname{dom} h$ has diameter 1 in the ℓ_1-norm.) $\diamond$

12.3.3 A population-level algorithm for calibration

Implicit in each of the calibration gap bounds in Propositions 12.3.1 and 12.3.5 is bound on the improvement of a predictor f relative to processing outputs with a class $\mathcal{W}$ of functions. This suggests an algorithm for updating the predictions of f to make them calibrated, after which no improvement is possible. While we work at the population level here, similar procedures can allow calibration given access to additional data.

Working in the more general setting of surrogate losses based on the generalized negative entropy h, as these include the standard squared error as a special case, the key idea is that if we find the witness w maximizing $\mathbb{E}[\langle w(f(X)), Y - \mathsf{pred}_h(f(X)) \rangle]$ we can update f with $f - \eta \cdot w \circ f$ for some stepsize η, thus improving the calibration of f relative to the class $\mathcal{W}$ of potential witnesses. In Figure 12.1, we present a prototypical algorithm for achieving this.

The following theorem bounds the convergence of the algorithm.

Theorem 12.3.8. *Assume that the surrogate loss φ is nonnegative and that the class of witnesses $\mathcal{W}$ satisfies $R_* := \sup_s \|w(s)\|_* < \infty$. Then the algorithm in Figure 12.1 guarantees that*

$$\min_{\tau < t} \mathrm{CE}(\mathsf{pred}_h \circ f_\tau, \mathcal{W}) \le \frac{\sqrt{2R_*^2 \mathbb{E}[\varphi(f_0(X), Y)]}}{\sqrt{t}},$$

and in particular terminates with $\mathrm{CE}(\mathsf{pred}_h \circ f_t, \mathcal{W}) \le \epsilon$ for some t with

$$t \le \frac{2R_*^2 \mathbb{E}[\varphi(f_0(X), Y)]}{\epsilon^2}.$$

Proof We begin by showing a one-step progress guarantee beginning from a fixed function f. For any $w : \mathbb{R}^k \to \mathbb{R}^k$ and any f, we have

$$\mathbb{E}[\varphi(f(X) + \eta w(f(X)), Y)] \le \mathbb{E}[\varphi(f(X), Y)] + \eta \mathbb{E}[\langle w(f(X)), \nabla h^*(f(X)) - Y \rangle] + \frac{\eta^2}{2} \mathbb{E}[\|w(f(X))\|_*^2].$$

Input: Population distribution P, collection of bounded witness functions $\mathcal{W}$, generalized negative entropy h strongly convex w.r.t. norm $\|\cdot\|$, initial predictor $f_0 : \mathcal{X} \to \mathbb{R}^k$, calibration tolerance $\epsilon > 0$
Initialize: set $R_*^2 := \sup_s \|w(s)\|_*^2$
Repeat: for $t = 0, 1, \ldots$

i. Find witness w_t maximizing $\mathbb{E}[\langle w(f_t(X)), Y - \mathsf{pred}_h(f_t(X))\rangle]$

ii. Set $\eta_t = \frac{\mathbb{E}[\langle w_t(f_t(X)), Y - \mathsf{pred}_h(f_t(X))\rangle]}{R_*^2}$

iii. Update $f_{t+1} = f_t - \eta_t \cdot w_t \circ f_t$

iv. Terminate if
$$\mathrm{CE}(\mathsf{pred}_h \circ f_t, \mathcal{W}) \le \epsilon.$$

Figure 12.1: Improving calibration relative to the class $\mathcal{W}$

Let w maximize $\mathbb{E}[\langle w(f(X)), \nabla h^*(f(X)) - Y\rangle]$, so that

$$\mathbb{E}[\varphi(f(X) - \eta w(f(X)), Y)] \le \mathbb{E}[\varphi(f(X), Y)] - \eta \mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W}) + \frac{\eta^2}{2} R_*^2.$$

Choose $\eta_f = \frac{\mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W})}{R_*^2}$ to obtain

$$\mathbb{E}[\varphi(f(X) - \eta_f w(f(X)), Y)] \le \mathbb{E}[\varphi(f(X), Y)] - \frac{1}{2} \frac{\mathrm{CE}(\mathsf{pred}_h \circ f, \mathcal{W})^2}{R_*^2}. \tag{12.3.4}$$

Now we apply the obvious inductive argument. Let f_t be a function in the iteration of Algorithm 12.1. Then inequality (12.3.4) guarantees that if $\delta_t^2 := \frac{1}{2} \frac{\mathrm{CE}(\mathsf{pred}_h \circ f_t, \mathcal{W})^2}{R_*^2}$, then

$$\mathbb{E}[\varphi(f_{t+1}(X), Y)] \le \mathbb{E}[\varphi(f_t(X), Y)] - \delta_t^2.$$

In particular,

$$0 \le \mathbb{E}[\varphi(f_t(X), Y)] \le \mathbb{E}[\varphi(f_0(X), Y)] - \sum_{\tau=0}^{t-1} \delta_\tau^2.$$

In particular,

$$t \min_{\tau < t} \delta_\tau^2 \le \mathbb{E}[\varphi(f_0(X), Y)],$$

so that $\min_{\tau<t} \delta_\tau \le \sqrt{\mathbb{E}[\varphi(f_0(X), Y)]/t}$. Replacing δ_τ with its definition gives the theorem. □

12.4 Calibeating: improving squared error by calibration

Sections 12.1 and 12.3 show that at least at the population level, taking a predictor f and modifying (or postprocessing) it to guarantee its calibration can only improve the losses it suffers, whether

those are squared error or general proper losses. That is, by calibrating we can beat (and hence, calibeat) a given predictor. These arguments have exclusively been at the population level, leaving it unclear whether this approach might actually work given a finite sample. While employing these ideas for general losses and general decision settings, where we only guarantee $\mathcal{Y} \subset \mathbb{R}^k$, is challenging because of dimensionality issues, here we show how to improve calibration in finite samples while simultaneously losing little in squared error for binary predictions with $Y \in \{0, 1\}$. That is, we have *calibeating*: from any potential predictor f, we can construct a predictor g with both small calibration error and with (asymptotically) no larger squared error than f, realizing Theorem 12.1.1 but in finite samples.

Let $f : \mathcal{X} \to [0, 1]$ be any predictor of $Y \in \{0, 1\}$, and consider the squared error loss $\ell(s, y) = (s - y)^2$ with population loss $L(f) = \mathbb{E}[(Y - f(X))^2]$. The idea to improve calibration of f without losing much in accuracy (squared error) is fairly straightforward: we discretize f by binning its predictions so that the number of X_i for which $f(X_i)$ is in a bin is equal; such binning ideas are central to the theory of calibration. Then we choose the postprocessed function g by averaging observed Y values over those bins. This transforms the (population level) idea present in Theorem 12.1.1, which says to choose the post-processing conditional expectation $g(x) = \mathbb{E}[Y \mid f(X) = f(x)]$, into one implementable in finite samples, which approximately sets

$$g(x) \approx \mathbb{E}[Y \mid l(x) \le f(X) \le u(x)],$$

where l and u are lower and upper bounds over which to average the predictions of f.

To make the ideas concrete, assume we have a sample $(X_i, Y_i)_{i=1}^{2n}$ of size $2n$ drawn i.i.d. according to P (where we choose $2n$ for notational convenience), which we divide into samples $\{(X_i, Y_i)\}_{i=1}^{n}$ and $\{(X_i, Y_i)\}_{i=n+1}^{2n}$, letting $P_n^{(1)}$ denote the empirical distribution on the first sample and $P_n^{(2)}$ that on the second. We use the first to choose the binning (quantization) of f and the second to actually choose values for the binned function. Fix a number of bins $b \in \mathbb{N}$ to be chosen, for convenience assuming that b divides n. Let the indices $i_1, \ldots, i_n$ sort $f(X_i)$, so that

$$f(X_{i_1}) < f(X_{i_2}) < \cdots < f(X_{i_n}),$$

and construct index partitions I_j, $j = 1, \ldots, b$, by $I_j := \{i_{b(j-1)+1}, \ldots, i_{bj}\}$. Here, we have assumed (essentially) without loss of generality that the predictions $f(X_i)$ are distinct with probability 1.[1] Given this partitioning of indices $I_1, \ldots, I_b$, for $j = 1, \ldots, b$ define the lower and upper bin boundaries

$$\widehat{l}_j = \max_{i \in I_{j-1}} f(X_i) \quad \text{and} \quad \widehat{u}_j = \max_{i \in I_j} f(X_i),$$

except that $\widehat{l}_1 = 0$ and $\widehat{u}_b = 1$, and define the bins

$$B_1 = \left[\widehat{l}_1, \widehat{u}_1\right), \quad B_2 = \left[\widehat{l}_2, \widehat{u}_2\right), \ldots, B_b = \left[\widehat{l}_b, \widehat{u}_b\right]$$

to partition $[0, 1]$. These partition $[0, 1]$ evenly in the empirical probabilities of $f(X_i)$, $i = 1, \ldots, n$, *not* evenly in the widths $\widehat{u}_j - \widehat{l}_j$.

To construct the recalibrated and binned version g of f, for each $x \in \mathcal{X}$, define the bin mapping

$$\text{bin}(x) := \text{the bin } j \text{ such that } f(x) \in B_j,$$

[1] If this distinctness fails, we can add random dithering by letting $U_i \overset{\text{iid}}{\sim} \mathsf{Uniform}[-\frac{1}{2}, \frac{1}{2}]$ and replacing the observations X_i with pairs (X_i, U_i) and $f(X_i)$ with $f_{\text{ext}}(X_i, U_i) := f(X_i) + \epsilon U_i$ for some $\epsilon > 0$. Then $L(f_{\text{ext}}) = \mathbb{E}[(Y - f(X) - \epsilon U)^2] = \mathbb{E}[(Y - f(X))^2] + \frac{\epsilon^2}{12}$ and $\ell(f_{\text{ext}}(x, u), y) \le \ell(f(x), y) + 2\epsilon$ for all x, u, y, so that we lose little.

which implicitly depends on the first sample (X_1^n, Y_1^n). The partitioning of $[0,1]$ into the bins B_j also induces a partition on $\mathcal{X} = \bigcup_{j=1}^b f^{-1}(B_j)$, where elements x, x' belong to the same partition set if $\mathrm{bin}(x) = \mathrm{bin}(x')$. Once we have this mapping from x to the associated prediction bin, we can use the second sample (its empirical distribution) to define the binned function g by the average of the second sample distribution $P_n^{(2)}$ over those examples falling into each bin. Formally, we define g to be the the piecewise constant function

$$g(x) := \mathbb{E}_{P_n^{(2)}}[Y \mid \mathrm{bin}(X) = \mathrm{bin}(x)], \tag{12.4.1}$$

or equivalently, for each $x \in B_j$, we have

$$\begin{aligned} g(x) &:= \mathbb{E}_{P_n^{(2)}}[Y \mid f(X) \in B_j] \\ &= \frac{1}{\sum_{i=n+1}^{2n} \mathbf{1}\left\{\mathrm{bin}(X_i) = j\right\}} \sum_{i=n+1}^{2n} \mathbf{1}\left\{\mathrm{bin}(X_i) = j\right\} Y_i \end{aligned}$$

where we assign $g(x)$ an arbitrary value if no X_i satisfies $f(X_i) \in B_j$ for the index $j = \mathrm{bin}(x)$.

Informally, this function g partitions X space into regions of roughly equal (small) probability $1/b$, and for which $f(x)$ belongs to a given interval on each region. Then recalibrating f on that region changes the prediction error $(Y - f(X))^2$ little, but improves the calibration. Formally, we can show the following theorem.

Theorem 12.4.1. *Let g be the binned and recalibrated estimator (12.4.1). Assume that the number of bins b and sample size n satisfy $\frac{n}{\log n} \geq b$. Then there exists a numerical constant $c > 0$ such that for all $\delta \in (0,1)$, with probability at least $1 - 2\exp(-c\frac{n}{b}) - \delta$,*

$$L(g) \leq L(f) + \frac{3}{b} + \frac{2b \log \frac{2b}{\delta}}{n} - \mathbb{E}\left[\left(\mathbb{E}[Y \mid \mathrm{bin}(X)] - \mathbb{E}[f(X) \mid \mathrm{bin}(X)]\right)^2\right]$$

and g has expected calibration error (12.2.1) at most

$$\mathsf{ece}(g) \leq \sqrt{\frac{2b \log \frac{2b}{\delta}}{n}}.$$

JCD Comment: Put in some figures here.

The proof of Theorem 12.4.1 is long, so we defer it to Section 12.4.1. To interpret the theorem, consider the terms in it. Roughly, we see that if we choose the number of bins to be $\sqrt{n \log \frac{1}{\delta}}$, then the calibeating predictor g guarantees

$$L(g) \leq L(f) + O(1)\sqrt{\frac{\log \frac{n}{\delta}}{n}} - \mathbb{E}\left[\left(\mathbb{E}[Y \mid \mathrm{bin}(X)] - \mathbb{E}[f(X) \mid \mathrm{bin}(X)]\right)^2\right],$$

while the expected calibration error is of order $n^{-1/4}$, ignoring the logarithmic factors. So we improve the loss $L(f)$ by a factor involving the calibration error of f (relative to the random binning)—the less calibrated f is, the more improvement we can provide—and with a penalty tending to 0 at rate $\sqrt{\log n / n}$.

12.4.1 Proof of Theorem 12.4.1

Throughout the proof, we use the shorthands that $P(B_j) = P(f(X) \in B_j)$ and $P_n(B_j) = P_n(f(X) \in B_j) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{f(X_i) \in B_j\}$ to mean the (empirical) probability that $f(X) \in B_j$, and $P_n^{(1)}$ and $P_n^{(2)}$ denote empirical probabilities with respect to the samples (X_1^n, Y_1^n) and $(X_{n+1}^{2n}, Y_{n+1}^{2n})$, respectively. The key to the argument is to show three things:

1. With high probability, each bin B_j has the approximately correct probability $\frac{1}{2b} \le P(B_j) \le \frac{7}{4b}$.
2. With similarly high probability, the empirical probabilities on the second sample $P_n^{(2)}$ satisfy $\frac{1}{4b} \le P_n^{(2)}(B_j) \le \frac{2}{b}$.
3. Conditional on $P_n^{(2)}(B_j)$ being large enough, the expectations $\mathbb{E}_{P_n^{(2)}}[Y \mid f(X) \in B_j]$ are accurate, so that $g(x) \approx \mathbb{E}[Y \mid f(X) \in B_j]$ for x satisfying $f(x) \in B_j$.

Once we have each of these three, we can show that $L(g)$ is essentially no larger than $L(f)$, up to diminishing error terms in n, and that g itself is well-calibrated. We proceed through each step in turn, stating the results as lemmas whose proofs we provide at the end of this section.

Lemma 12.4.2. *Let $\frac{n}{\log n} \ge b$. For a numerical constant $c > 0$, we have*

$$\mathbb{P}\left(\frac{1}{2b} \le P(B_j) \le \frac{7}{4b} \text{ for all } j = 1, \ldots, b\right) \ge 1 - 2\exp\left(-c\frac{n}{b}\right).$$

With Lemma 12.4.2 in hand, the second step of the proof of Theorem 12.4.1 is relatively straightforward. In the lemma, conditioning on $P_n^{(1)}$ indicates conditioning on the first sample (X_1^n, Y_1^n).

Lemma 12.4.3. *Let $\frac{n}{\log n} \ge b$. Assume the first sample $P_n^{(1)}$ is such that $\frac{1}{2b} \le P(B_j) \le \frac{7}{4b}$ for each selected bin B_j, $j = 1, \ldots, b$. Then there exists a numerical constant $c > 0$ such that*

$$\mathbb{P}\left(\frac{1}{4b} \le P_n^{(2)}(B_j) \le \frac{2}{b} \mid P_n^{(1)}\right) \ge 1 - 2\exp\left(-c\frac{n}{b}\right).$$

Lemma 12.4.4. *Let the conditions of Lemma 12.4.3 hold. Then there exists a numerical constant $c > 0$ such that for any $\delta \in (0, 1)$*

$$\mathbb{P}\left(\max_{j \le b} \sup_{x : f(x) \in B_j} |g(x) - \mathbb{E}[Y \mid f(X) \in B_j]| \ge \sqrt{\frac{2b}{n}\log\frac{2b}{\delta}} \mid P_n^{(1)}\right) \le 2\exp\left(-c\frac{n}{b}\right) + \delta.$$

With the three lemmas in place, we can now expand the squared error to obtain the calibeating theorem. Recalling the population squared error $L(g) = \mathbb{E}[(Y - g(X))^2]$, let us suppose that the consequences of Lemmas 12.4.2–12.4.4 hold, so that $|g(x) - \mathbb{E}[Y \mid f(X) \in B_j]|^2 \le \frac{2b}{n}\log\frac{2b}{\delta}$ and $P(B_j) \le \frac{7}{4b}$ for each j. By the lemmas, these hold with probability $1 - 2\exp(-c\frac{n}{d}) - \delta$. Define the average function values and conditional expectations

$$\overline{f}_j := \mathbb{E}[f(X) \mid f(X) \in B_j] \quad \text{and} \quad \overline{E}_j := \mathbb{E}[Y \mid f(X) \in B_j].$$

Then we have

$$L(g) = \mathbb{E}[(Y - g(X))^2] = \sum_{j=1}^b P(B_j)\mathbb{E}[(Y - \overline{E}_j + \overline{E}_j - g(X))^2 \mid f(X) \in B_j].$$

Considering the expectation term, note that $g(X)$ is constant for $f(X) \in B_j$ by construction of the binning, and so for any $x \in f^{-1}(B_j)$, we have

$$\begin{aligned}
&\mathbb{E}[(Y - \overline{E}_j + \overline{E}_j - g(X))^2 \mid f(X) \in B_j] \\
&= \mathbb{E}[(Y - \mathbb{E}[Y \mid f(X) \in B_j])^2 \mid f(X) \in B_j] + (g(x) - \mathbb{E}[Y \mid f(X) \in B_j])^2 \\
&\le \mathbb{E}[(Y - \mathbb{E}[Y \mid f(X) \in B_j])^2 \mid f(X) \in B_j] + \frac{2b}{n} \log \frac{2b}{\delta}.
\end{aligned}$$

Now, using that $\mathbb{E}[Y \mid f(X) \in B_j] = \overline{E}_j$, we see that

$$\mathbb{E}[(Y - \overline{E}_j)^2 \mid f(X) \in B_j] = \mathbb{E}[(Y - \overline{f}_j)^2 \mid f(X) \in B_j] - (\overline{E}_j - \overline{f}_j)^2$$

by adding and subtracting $\overline{f}_j$ and expanding the square. Summarizing, we have shown so far that

$$L(g) \le \sum_{j=1}^b P(B_j) \mathbb{E}[(Y - \overline{f}_j)^2 \mid f(X) \in B_j] + \frac{2b}{n} \log \frac{2b}{\delta} - \sum_{j=1}^b P(B_j)(\overline{E}_j - \overline{f}_j)^2. \tag{12.4.2}$$

We can directly relate the first term in the expansion (12.4.2) to the expected error $\mathbb{E}[(Y - f(X))^2]$. Indeed, by expanding out the square, we have

$$\begin{aligned}
&\mathbb{E}[(Y - \overline{f}_j)^2 \mid f(X) \in B_j] \\
&= \mathbb{E}[(Y - f(X) + f(X) - \overline{f}_j)^2 \mid f(X) \in B_j] \\
&= \mathbb{E}[(Y - f(X))^2 \mid f(X) \in B_j] + 2\mathbb{E}[(Y - f(X))(f(X) - \overline{f}_j) \mid f(X) \in B_j] + \mathrm{Var}(f(X) \mid f(X) \in B_j) \\
&\le \mathbb{E}[(Y - f(X))^2 \mid f(X) \in B_j] + 2\sqrt{\mathrm{Var}(f(X) \mid f(X) \in B_j)} + \mathrm{Var}(f(X) \mid f(X) \in B_j),
\end{aligned}$$

where the inequality is Cauchy-Schwarz, as $|Y - f(X)| \le 1$. Finally, we recognize that $B_j \subset [\widehat{l}_j, \widehat{u}_j]$, so $\mathrm{Var}(f(X) \mid f(X) \in B_j) \le \frac{1}{4}(\widehat{u}_j - \widehat{l}_j)^2$, and thus

$$\mathbb{E}[(Y - \overline{f}_j)^2 \mid f(X) \in B_j] \le \mathbb{E}[(Y - f(X))^2 \mid f(X) \in B_j] + \frac{5}{4}(\widehat{u}_j - \widehat{l}_j).$$

Substituting in the bound (12.4.2) and recognizing that $\sum_{j=1}^b P(B_j)\mathbb{E}[(Y - f(X))^2 \mid f(X) \in B_j] = \mathbb{E}[(Y - f(X))^2] = L(f)$, we have

$$L(g) \le L(f) + \frac{5}{4} \sum_{j=1}^b P(B_j)(\widehat{u}_j - \widehat{l}_j) + \frac{2b}{n} \log \frac{2b}{\delta} - \sum_{j=1}^b P(B_j)(\overline{E}_j - \overline{f}_j)^2.$$

But of course, $P(B_j) \le \frac{7}{4b}$ by the assumed conclusions of Lemma 12.4.2, and so $\sum_{j=1}^b P(B_j)(\widehat{u}_j - \widehat{l}_j) \le \frac{7}{4b}$ as $\sum_{j=1}^b (\widehat{u}_j - \widehat{l}_j) = 1$. This gives the final inequality

$$L(g) \le L(f) + \frac{35}{16b} + \frac{2b}{n} \log \frac{2b}{\delta} - \sum_{j=1}^b P(B_j)(\overline{E}_j - \overline{f}_j)^2,$$

proving the first claim of the theorem. The bound on calibration error is immediate because $|g(x) - \mathbb{E}[Y \mid f(X) \in B_j]|^2 \le \frac{2b}{n} \log \frac{2b}{\delta}$ for each $x \in f^{-1}(B_j)$ with the prescribed probability, by Lemma 12.4.4.

Proof of Lemma 12.4.2 We follow the notational shorthand $P_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{f(X_i) \in A\}$. Fix a pair $0 \le l < u \le 1$ and define the interval $A = [l, u]$. Then Bernstein's inequality (4.1.8) shows that

$$\mathbb{P}\left(\left|\frac{1}{n}P_n(A) - P(A)\right| \ge v\right) \le 2\exp\left(-\frac{nv^2}{2P(A) + \frac{2}{3}v}\right)$$

for all $v \ge 0$. Partition $[0,1]$ into intervals $A_1, \ldots, A_{4b}$, $A_j = [l_j, u_j]$, each of probability $P(A_j) = \frac{1}{4b}$. Now, fix an index $j^\star \in [b]$ and consider the (empirically constructed) bin $B_{j^\star} = [\widehat{l}_{j^\star}, \widehat{u}_{j^\star})$. Then there exist some $j, k \in \mathbb{N}$ such that

$$A_j \cup \cdots \cup A_{j+k} \supset B_{j^\star} \supset A_{j+1} \cup \cdots \cup A_{j+k-1}.$$

We provide upper and lower bounds on k as a function of the error in $P_n(A_j)$. Suppose that for some $t > 0$, we have

$$\frac{1-t}{4b} \le P_n(A_j) \le \frac{1+t}{4b} \quad \text{for } j = 1, \ldots, 4b. \tag{12.4.3}$$

Then

$$\frac{1+t}{4b}(k+1) \ge P_n(A_j \cup \cdots \cup A_{j+k}) \ge P_n(B_{j^\star}) = \frac{1}{b},$$

and similarly

$$\frac{1-t}{4b}(k-1) \le P_n(A_{j+1} \cup \cdots \cup A_{j+k}) \le P_n(B_{j^\star}) = \frac{1}{b},$$

implying the bounds

$$\frac{4}{1+t} - 1 \le k \le \frac{4}{t-1} + 1.$$

In particular, if $t < \frac{1}{3}$ then $3 \le k \le 6$, and so when the bounds (12.4.3) hold with $t = \frac{1}{3}$ we obtain

$$\frac{1}{2b} \le \frac{k-1}{4b} = P(A_{j+1} \cup \cdots \cup A_{j+k-1}) \le P(B_{j^\star}) \le P(A_j \cup \cdots \cup A_{j+k}) = \frac{k+1}{4b} \le \frac{7}{4b}.$$

Apply Bernstein's inequality for using $t = \frac{1}{3}$, or $v = \frac{1}{12b}$, with variance bound $\sigma^2 \le P(A_j) \le \frac{1}{4b}$ to obtain that for each $j = 1, \ldots, 4b$, we have

$$\mathbb{P}\left(|P_n(A_j) - P(A_j)| \ge \frac{1}{12b}\right) \le 2\exp\left(-\frac{n/(12b)^2}{2/(4b) + \frac{2}{3}\frac{1}{12b}}\right) = 2\exp\left(-\frac{n}{80b}\right).$$

Apply a union bound to obtain the lemma once we recognize that $n/b - \log b \gtrsim n/b$ whenever $n/\log n \ge b$. □

Proof of Lemma 12.4.3 Assume that $P(B_j) \le \frac{7}{4b}$. Then applying Bernstein's inequality (4.1.8), and using that $\mathbf{1}\{f(X) \in B_j\}$ is a Bernoulli random variable with mean (and hence variance) at most $\frac{7}{4b}$, we have

$$\mathbb{P}\left(P_n^{(2)}(B_j) \ge \frac{2}{b}\right) \le \exp\left(-\frac{n/(4b)^2}{\frac{7}{4b} + \frac{2}{3}\frac{1}{4b}}\right) = \exp\left(-\frac{1}{28 + 8/3}\frac{n}{b}\right) \le \exp\left(-\frac{1}{31}\frac{n}{b}\right).$$

Similarly, we have $\mathbb{P}(P_n^{(2)}(B_j) \le \frac{1}{4b}) \le \exp(-\frac{1}{31}\frac{n}{b})$ as $P(B_j) \ge \frac{1}{2b}$. Applying a union bound over $j = 1, \ldots, b$, then noting that $n/b - \log b \gtrsim n/b$ whenever $n/\log n \ge b$, we again obtain □

Proof of Lemma 12.4.4 Recall that $g(x) = \mathbb{E}_{P_n^{(2)}}[Y \mid \mathrm{bin}(X) = \mathrm{bin}(x)]$, and note that g is constant on $x \in f^{-1}(B_j)$. Fix a bin j, and let $I(j) = \{i \in \{n+1, \ldots, 2n\} \mid f(X_{n+i}) \in B_j\}$ denote the indices in the second sample for which $f(X_{n+i})$ falls in bin B_j. Then conditional on $i \in I(j)$, we have $Y_i \sim P(Y \in \cdot \mid f(X) \in B_j)$, so that

$$\mathbb{P}\left(\left|\frac{1}{|I(j)|}\sum_{i \in I(j)} Y_i - \mathbb{E}[Y \mid f(X) \in B_j]\right| \ge t \mid I(j)\right) \le 2\exp\left(-2\,\mathrm{card}(I(j))t^2\right)$$

by Hoeffding's inequality. Then (conditioning on the bins $\{B_j\}$ chosen using $P_n^{(1)}$, which by assumption satisfy $P(B_j) \in [\frac{1}{2b}, \frac{7}{4b}]$, we have for any fixed $x \in f^{-1}(B_j)$ that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x \in f^{-1}(B_j)} |g(x) - \mathbb{E}[Y \mid f(X) \in B_j]| \ge t \mid P_n^{(1)}\right) \\
&= \sum_{I \subset [n]} \mathbb{P}\left(|g(x) - \mathbb{E}[Y \mid f(X) \in B_j]| \ge t, I(j) = I \mid P_n^{(1)}\right) \\
&\le \mathbb{P}\left(\mathrm{card}(I(j)) < \frac{n}{4b} \mid P_n^{(1)}\right) + \sum_{I \subset [n], \mathrm{card}(I) \ge n/4b} \mathbb{P}\left(|g(x) - \mathbb{E}[Y \mid f(X) \in B_j]| \ge t, I(j) = I \mid P_n^{(1)}\right) \\
&\le \mathbb{P}\left(P_n^{(2)}(B_j) < \frac{1}{4b}\right) + 2\exp\left(-\frac{nt^2}{2b}\right),
\end{aligned}$$

where the final line applies Hoeffding's inequality. Taking $t^2 = \frac{2b \log \frac{2b}{\delta}}{n}$ and applying Lemma 12.4.3 and a union bound gives Lemma 12.4.4. □

12.5 Continuous and equivalent calibration measures

We finally return to constructing a calculus and tools with which to measure calibration, addressing the issues of discontinuity of ece that Example 12.2.2 highlights, and building to a combination of results that imply Corollary 12.2.6. In the end, we will see that for appropriate classes $\mathcal{F}$ of predictors, several potential measures $\mathrm{M} : \mathcal{F} \to \mathbb{R}_+$ are roughly equivalent sound and complete calibration measures, all enjoying similar continuity properties. We begin with two definitions.

Definition 12.1. *A function* $\mathrm{M} : \mathcal{F} \to \mathbb{R}_+$ *is a* continuous calibration measure *for the distribution* P *on* $\mathcal{X} \times \mathcal{Y}$ *if*

(i) it is sound and complete (12.2.2)*, that is,* $\mathrm{M}(f) = 0$ *if and only if* f *is calibrated for* P*, and*

(ii) it is continuous with respect to the $L^1(P)$ *metric on* $\mathcal{F}$*, that is, for any* f*, if* f_n *is a sequence of functions with* $\mathbb{E}[\|f(X) - f_n(X)\|] \to 0$*, then*

$$\mathrm{M}(f) - \mathrm{M}(f_n) \to 0.$$

A stronger definition replaces continuity with a Lipschitz requirement.

Definition 12.2. *A function* $\mathrm{M} : \mathcal{F} \to \mathbb{R}_+$ *is a* Lipschitz calibration measure *for the distribution* P *on* $\mathcal{X} \times \mathcal{Y}$ *if it is sound and complete (Definition 12.1, part (i)), and instead of part (ii) satisfies*

(iii) it is Lipschitz continuous with respect to the $L^1(P)$ *metric on* $\mathcal{F}$*, that is, for some* $C < \infty$

$$|\mathrm{M}(f_0) - \mathrm{M}(f_1)| \le C \cdot \mathbb{E}_P[\|f_0(X) - f_1(X)\|]$$

for all $f_0, f_1 \in \mathcal{F}$.

If conditions (i) and (ii) (respectively (iii)) hold for all P in a collection of distributions $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$, we will say that M is a continuous (respectively, Lipschitz) calibration measure for $\mathcal{P}$.

The desiderata (ii) and (iii) are matters of taste; the central idea is that some type of continuity is essential for efficient modeling, estimation, and analysis. We leave the norm $\|\cdot\|$ implicit in the definition, and we typically omit the distribution P from the calibration metric as it is clear from context. The two parts of Definition 12.2 admit many possible calibration measures. We consider two types of measures, which are (almost) dual to one another, as examples. Both use a variational representation, where in one we essentially look for the "closest" function that is calibrated, while in the other, we investigate the ease with which we can (quantitatively) certify that a predictor f is uncalibrated.

A key concept will be the equivalence of calibration measures, where we target a quantitative equivalence. To define this, let $0 < \alpha, \beta < \infty$. Then we say that two candidate calibration measures M_0 and M_1 on $\mathcal{F} \subset \mathcal{X} \to \mathbb{R}^k$ are (α, β)-*equivalent* if there exist constants c_0, c_1 (which may depend on $\mathcal{Y}$) such that

$$\mathrm{M}_0(f) \le c_0 \left[\mathrm{M}_1(f) + \mathrm{M}_1(f)^\alpha\right] \quad \text{and} \quad \mathrm{M}_1(f) \le c_1 \left[\mathrm{M}_0(f) + \mathrm{M}_0(f)^\beta\right]. \tag{12.5.1}$$

Then in a strong sense, $\mathrm{M}_0(f) \to 0$ if and only if $\mathrm{M}_1(f) \to 0$.

12.5.1 Calibration measures

We revisit the potential calibration measures in Section 12.2.2 here to recapitulate definitions, providing initial results on their soundness and completeness. We focus on the distance to calibration (12.2.3) and relative calibration errors (12.2.4), as the partitioned calibration error (12.2.6) we use more as a proof device.

Distances to calibration. Recall the *distance to calibration* (12.2.3), which for $\mathcal{C}(P) = \{g : \mathcal{X} \to \mathbb{R}^k \mid \mathbb{E}_P[Y \mid g(X)] = g(X)\}$ (where the defining equality holds with P-probability 1 over X) has definition $d_{\mathrm{cal}}(f) := \inf_g \{\mathbb{E}[\|g(X) - f(X)\|] \text{ s.t. } g \in \mathcal{C}(P)\}$. The measure (12.2.3) is, after appropriate normalization, the *largest* Lipschitz measure of calibration: if M is any Lipschitz calibration measure (with constant $C = 1$ in Definition 12.2 part (iii)), then taking a perfectly calibrated g with $\mathsf{ece}(g) = 0$, we necessarily have $\mathrm{M}(g) = 0$. Then for any f we have $\mathrm{M}(f) = \mathrm{M}(f) - \mathrm{M}(g) \le \mathbb{E}[\|f(X) - g(X)\|]$, and taking an infimum over such g guarantees

$$\mathrm{M}(f) \le d_{\mathrm{cal}}(f).$$

The second related quantity, which sometimes admits cleaner properties for analysis, is the *penalized calibration distance*, which we define as

$$p_{\mathrm{cal}}(f) := \inf_g \left\{\mathbb{E}[\|f(X) - g(X)\|] + \mathbb{E}\left[\|\mathbb{E}[Y \mid g(X)] - g(X)\|\right]\right\}. \tag{12.5.2}$$

These quantities are strongly related, and in the sequel (see Corollary 12.5.8), we show that

$$p_{\rm cal}(f) \le d_{\rm cal}(f) \le p_{\rm cal}(f) + C_{\mathcal{Y}} \sqrt{p_{\rm cal}(f)},$$

where $C_{\mathcal{Y}}$ is a constant depending only on the set $\mathcal{Y}$ whenever $\mathcal{Y}$ has finite diameter.

To build intuition for the definition (12.5.2), consider the two quantities. The first measures the usual L^1 distance between the function f and a putative alternative g. The second is the expected calibration error of g. By restricting the infimum in definition (12.2.3) to functions g with $\mathsf{ece}(g) = 0$, we simply have the L^1 distance to the nearest calibrated function; as is, the additional term in (12.5.2) allows trading between the distance to a calibrated function and the actual calibration error. We also have the following proposition.

Proposition 12.5.1. *The functions $d_{\rm cal}$ and $p_{\rm cal}$ are Lipschitz calibration measures.*

Proof If f is calibrated, then $p_{\rm cal}(f) = d_{\rm cal}(f) = 0$ immediately. Conversely, if $p_{\rm cal}(f) = 0$, there exists a sequence of functions g_n satisfying $\mathbb{E}[\|f(X) - g_n(X)\|] \to 0$, as each term in the definition (12.5.2) is nonnegative. Additionally, we must have that $\mathsf{ece}(g_n) = \mathbb{E}[\|\mathbb{E}[Y \mid g_n(X)] - g_n(X)\|] \to 0$. Applying Lemma 12.2.1 we have $0 \ge \liminf_n \mathsf{ece}(g_n) \ge \mathsf{ece}(f)$. If $d_{\rm cal}(f) = 0$, then there exists a sequence of functions g_n with $\mathsf{ece}(g_n) = 0$ and $\mathbb{E}[\|f(X) - g_n(X)\|] \to 0$. Again, the lower semicontinuity of ece from Lemma 12.2.1 gives $0 = \liminf_n \mathsf{ece}(g_n) \ge \mathsf{ece}(f)$.

To see that $p_{\rm cal}$ is Lipschitz in f, let $f_0, f_1 : \mathcal{X} \to \mathbb{R}^k$, and let g_0, g_1 be within $\epsilon > 0$ of achieving the infima in definition (12.5.2) for f_0 and f_1, respectively. Then

$$\begin{aligned}
p_{\rm cal}(f_0) - p_{\rm cal}(f_1) &\le \inf_g \left\{\mathbb{E}[\|f_0(X) - g(X)\|] + \mathbb{E}[\|\mathbb{E}[Y \mid g(X)] - g(X)\|]\right\} \\
&\qquad - \mathbb{E}[\|f_1(X) - g_1(X)\|] + \mathbb{E}[\|\mathbb{E}[Y \mid g_1(X)] - g_1(X)\|] + \epsilon \\
&\le \mathbb{E}[\|f_0(X) - g_1(X)\|] - \mathbb{E}[\|f_1(X) - g_1(X)\|] + \epsilon \\
&\le \mathbb{E}[\|f_0(X) - f_1(X)\|] + \epsilon.
\end{aligned}$$

Take $\epsilon \downarrow 0$. The lower inequality is similar, as is the proof for $d_{\rm cal}$. □

Weak calibration. The calibration error (12.2.4) relative to a class $\mathcal{W}$,

$$\mathrm{CE}(f, \mathcal{W}) := \sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(f(X), Y - f(X)\rangle]$$

admits similar properties, as it also satisfies our desiderata for a calibration measure. In particular, if we take $\mathcal{W}$ to be the class $\mathcal{W}_{\|\cdot\|}$ of bounded Lipschitz witness functions (12.2.5), we have the next two propositions.

Proposition 12.5.2. *Let $\mathcal{F}$ consist of functions with $\mathbb{E}[\|f(X)\|] < \infty$ and assume $\mathbb{E}[\|Y\|] < \infty$. Then $\mathrm{CE}(\cdot, \mathcal{W}_{\|\cdot\|})$ is a continuous calibration measure over $\mathcal{F}$.*

Because continuity is such a weak requirement, the proof of this result relies on measure theoretic results, so we defer it to Section 12.6.2.

When we assume the collection $\mathcal{F}$ consists of bounded functions and $\mathcal{Y}$ itself is bounded, we can give a stronger guarantee for the weak calibration, and we no longer need to rely on careful arguments considering the order of various limits.

Proposition 12.5.3. *Assume that* $\operatorname{diam}(\mathcal{Y})$ *is finite and that* $\mathcal{F}$ *is a collection of bounded functions* $\mathcal{X} \to \mathbb{R}^k$. *Then* $\operatorname{CE}(\cdot, \mathcal{W}_{\|\cdot\|})$ *is a Lipschitz calibration measure over* $\mathcal{F}$.

Proof Let $\mathcal{W} = \mathcal{W}_{\|\cdot\|}$ for shorthand. That $\operatorname{CE}(f, \mathcal{W}) = 0$ when f is calibrated is immediate, as by definition of conditional expectation we have

$$\mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] = \mathbb{E}[\langle w(f(X)), \mathbb{E}[Y \mid f(X)] - f(X)\rangle] = 0.$$

To obtain the converse that $\operatorname{CE}(f, \mathcal{W}) = 0$ implies f is calibrated, we require an intermediate lemma, which leverages the density of Lipschitz functions in L^p spaces. As was the case for the lower semi-continuity lemma 12.2.1 central to the proof of the converse in Proposition 12.5.1, this lemma requires measure-theoretic approximation arguments, so we defer its proof to Section 12.6.3.

Lemma 12.5.4. *Let* $S \in \mathbb{R}^k$ *be a random variable and* $\mathbb{E}[\|g(S)\|] < \infty$. *If* $\mathbb{E}[\langle w(S), g(S)\rangle] = 0$ *for all bounded and* 1*-Lipschitz functions* w, *then* $g(S) = 0$ *with probability 1.*

The converse is now trivial: let $S = f(X)$, and note that $\operatorname{CE}(f, \mathcal{W}) = \sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(S), \mathbb{E}[Y \mid S] - S\rangle]$, and take $g(S) = \mathbb{E}[Y \mid S] - S$ in Lemma 12.5.4.

To see that CE is Lipschitz, let $w_0 \in \mathcal{W}$ be such that $\operatorname{CE}(f_0, \mathcal{W}) \ge \mathbb{E}[\langle w_0(f_0(X)), Y - f_0(X)\rangle] - \epsilon$, and let $C < \infty$ satisfy $C \ge \sup_{y \in \mathcal{Y}, x \in \mathcal{X}, f \in \mathcal{F}} \|y - f(x)\|$. Then

$$\begin{aligned}
\operatorname{CE}(f_0, \mathcal{W}) - \operatorname{CE}(f_1, \mathcal{W}) &\le \mathbb{E}[\langle w_0(f_0(X)), Y - f_0(X)\rangle] - \mathbb{E}[\langle w_0(f_1(X)), Y - f_1(X)\rangle] + \epsilon \\
&\le \mathbb{E}[\langle w_0(f_0(X)) - w_0(f_1(X)), Y - f_0(X)\rangle] + \mathbb{E}[\langle w_0(f_1(X)), f_1(X) - f_0(X)\rangle] + \epsilon \\
&\le C\mathbb{E}[\|w_0(f_0(X)) - w_0(f_1(X))\|_*] + \mathbb{E}[\|f_1(X) - f_0(X)\|] + \epsilon \\
&\le (1 + C)\mathbb{E}[\|f_1(X) - f_0(X)\|] + \epsilon.
\end{aligned}$$

Repeating the same argument, *mutatis mutandis*, for the lower bound gives the Lipschitz continuity as desired. □

The family of weak calibration measures $\operatorname{CE}(f, \mathcal{W})$ as we vary the collection of potential witness functions $\mathcal{W}$ yields a variety of behaviors. Different choices of $\mathcal{W}$ can give different continuous calibration measures, where we may modify Definition 12.1 part (ii) to other notions of continuity, such as Lipschitzness with respect to $L^2(P)$ norms. We explore a few of these in the exercises at the end of the chapter.

12.5.2 Equivalent calibration measures

That all three measures $d_{\text{cal}}(f)$, $p_{\text{cal}}(f)$, $\operatorname{CE}(f, \mathcal{W}_{\|\cdot\|})$ are Lipschitz calibration measures when the label space $\mathcal{Y}$ is bounded suggests deeper relationships between these and other notions of calibration, such as the equivalence (12.5.1). We elucidate this here, showing that each of the measures $d_{\text{cal}}, p_{\text{cal}}$, and CE are equivalent. Indeed, the main consequence of the results in this chapter is that this equivalence holds for multiclass classification.

Theorem 12.5.5. *Let* $\mathcal{Y} = \{e_1, \ldots, e_k\}$ *and* $\mathcal{W}_{\|\cdot\|}$ *be the collection* (12.2.5) *of bounded Lipschitz functions for a norm* $\|\cdot\|$ *on* $\mathbb{R}^k$. *Then* $d_{\text{cal}}, p_{\text{cal}}$, *and* $\operatorname{CE}(\cdot, \mathcal{W}_{\|\cdot\|})$ *are each* $(\frac{1}{2}, \frac{1}{2})$*-equivalent. Moreover, this equivalence is sharp, in that they are not* (α, β)*-equivalent for any* $\alpha, \beta > \frac{1}{2}$.

The theorem follows as a compilation of the other results in this section. Along the way to demonstrating this theorem, we introduce a few alternative measures of calibration we use as stepping stones toward our final results. While many of our derivations will apply for general sets $\mathcal{Y}$, in some cases we will restrict to multiclass classification problems, so that $\mathcal{Y} = \{e_1, \ldots, e_k\} \subset \mathbb{R}^k$ are the k standard basis vectors. We present two main results: the first, Theorem 12.5.6, shows an equivalence (up to a square root) between the penalized calibration distance (12.5.2) and the partitioned calibration error (12.2.6). As a corollary of this result, we obtain the equivalence of the distance to calibration (12.2.3) and penalized distance to calibration (12.5.2). The second main result, Theorem 12.5.9, gives a similar equivalence between the penalized distance (12.5.2) and the calibration error relative to Lipschitz functions (12.2.4). Throughout, to make the calculations cleaner and more transparent, we restrict our functions to make predictions in $\mathcal{M} = conv(\mathcal{Y})$.

Partition-based calibration measures and lifting to random variables

It is easier to work directly in the space of predictions $f(X) \in \mathbb{R}^k$ rather than addressing the underlying space $\mathcal{X}$. To that end, let $S = f(X)$ be the random vector (use the mnemonic that S is for "scores") induced by $f(X)$ and taking values in $\mathrm{Conv}(\mathcal{Y})$, which has a joint distribution (S, Y) with the label Y. Then, for example, the expected calibration error of f is simply

$$\mathsf{ece}(f) = \mathbb{E}[\|\mathbb{E}[Y \mid S] - S\|].$$

Once we work exclusively in the space of random scores $S = f(X)$, we may define alternative distances to calibration in analogy with the (penalized) distances to calibration, which will allow us to more easily relate distances to the partitioned error (12.2.6). Thus, we define

$$d_{\mathrm{cal,low}}(f) := \inf_V \{\mathbb{E}[\|S - V\|] \;\; \text{s.t.} \;\; \mathbb{E}[Y \mid V] = V\} \tag{12.5.3a}$$

and

$$p_{\mathrm{cal,low}}(f) := \inf_V \{\mathbb{E}[\|S - V\|] + \mathbb{E}[\|\mathbb{E}[Y \mid V] - V\|]\}, \tag{12.5.3b}$$

where the infimum are over all random variables V taking values in $\mathrm{Conv}(\mathcal{Y})$, which can have arbitrary distribution with (S, Y) (but do not modify the joint (S, Y)), and in case (12.5.3a) are calibrated. This formulation is convenient in that we can represent it as a convex optimization problem, allowing us to bring the tools of duality to bear on it, though we defer this temporarily. By considering $V = g(X)$ for functions $g : \mathcal{X} \to \mathrm{Conv}(\mathcal{Y})$, we immediately see that $p_{\mathrm{cal}}(f) \geq p_{\mathrm{cal,low}}(f)$. We can also consider upper distances

$$d_{\mathrm{cal,up}}(f) := \inf_g \{\mathbb{E}[\|S - g(S)\|] \;\; \text{s.t.} \;\; \mathbb{E}[Y \mid g(S)] = g(S)\}$$

and

$$p_{\mathrm{cal,up}}(f) := \inf_{g:\mathbb{R}^k \to \mathrm{Conv}(\mathcal{Y})} \{\mathbb{E}[\|S - g(S)\|] + \mathbb{E}[\|\mathbb{E}[Y \mid g(S)] - g(S)\|]\},$$

which restrict the definitions (12.2.3) and (12.5.2) to compositions. We therefore have the inequalities

$$d_{\mathrm{cal,low}}(f) \leq d_{\mathrm{cal}}(f) \leq d_{\mathrm{cal,up}}(f) \;\; \text{and} \;\; p_{\mathrm{cal,low}}(f) \leq p_{\mathrm{cal}}(f) \leq p_{\mathrm{cal,up}}(f). \tag{12.5.4}$$

The partitioned calibration error (12.2.6) allows us to provide a bound relating the calibration error and the lower and upper calibration errors. To state the theorem, we make a normalization with $\|\cdot\|$, assuming without loss of generality that $\|\cdot\|_\infty \leq \|\cdot\|$.

Theorem 12.5.6. *Let $\mathcal{Y} \subset \mathbb{R}^k$ have finite diameter $\operatorname{diam}(\mathcal{Y})$ in the norm $\|\cdot\|$. Let $S = f(X) \in \mathbb{R}^k$. Then for all $\varepsilon > 0$,*

$$\begin{aligned} p_{\mathrm{cal,up}}(f) \le d_{\mathrm{cal,up}}(f) \le \mathsf{pce}(S) &\le \left(1 + \frac{2k \operatorname{diam}(\mathcal{Y})}{\varepsilon}\right) p_{\mathrm{cal,low}}(f) + \|\mathbf{1}_k\|_* \varepsilon \\ &\le \left(1 + \frac{2k \operatorname{diam}(\mathcal{Y})}{\varepsilon}\right) d_{\mathrm{cal,low}}(f) + \|\mathbf{1}_k\|_* \varepsilon. \end{aligned}$$

While the first inequality in Theorem 12.5.6 is relatively straightforward to prove, the second requires substantially more care, so we defer the proof of the theorem to Section 12.6.4.

We record a few corollaries, one consequence of which is to show that the partitioned calibration error (12.2.6) is at least a calibration measure in the sense of Definition 12.2.(i). Theorem 12.5.6 also shows that the penalized calibration distance $p_{\mathrm{cal}}(f)$ is equivalent, up to taking a square root, to the upper and lower calibration "distances". In each corollary, we let $C_k = \|\mathbf{1}_k\|_*$ for shorthand.

Corollary 12.5.7. *Let the conditions of Theorem 12.5.6 hold. Then*

$$p_{\mathrm{cal,low}}(f) \le p_{\mathrm{cal}}(f) \le p_{\mathrm{cal,low}}(f) + 2\sqrt{C_k k \operatorname{diam}(\mathcal{Y})}\sqrt{p_{\mathrm{cal,low}}(f)}$$

and

$$d_{\mathrm{cal,low}}(f) \le d_{\mathrm{cal}}(f) \le d_{\mathrm{cal,low}}(f) + 2\sqrt{C_k k \operatorname{diam}(\mathcal{Y})}\sqrt{d_{\mathrm{cal,low}}(f)}.$$

Proof The first lower bound is immediate (recall the naive inequalitites (12.5.4)). Now set $\varepsilon = \sqrt{2k \operatorname{diam}(\mathcal{Y}) p_{\mathrm{cal,low}}(f)/C_k}$ in Theorem 12.5.6, and recognize that $p_{\mathrm{cal,low}}(f) \le p_{\mathrm{cal,up}}(f)$. □

We also obtain an approximate equivalence between the calibration distance d_{cal} and penalized calibration distance p_{cal} from definitions (12.2.3) and (12.5.2).

Corollary 12.5.8. *Let the conditions of Theorem 12.5.6 hold. Then*

$$p_{\mathrm{cal}}(f) \le d_{\mathrm{cal}}(f) \le p_{\mathrm{cal}}(f) + 2\sqrt{c_k k \operatorname{diam}(\mathcal{Y})}\sqrt{p_{\mathrm{cal}}(f)}.$$

Proof The first inequality is immediate by definition. For the second, note (see Lemma 12.6.4 in the proof of Theorem 12.5.6 in Section 12.6.4) that $p_{\mathrm{cal,low}}(f) \le \mathsf{pce}(S)$ for $S = f(X)$. Then apply Theorem 12.5.6 with $\varepsilon = \sqrt{2k \operatorname{diam}(\mathcal{Y}) p_{\mathrm{cal,low}}(f)/c_k}$ as in Corollary 12.5.7, and recognize that $p_{\mathrm{cal,low}} \le p_{\mathrm{cal}}$. □

Let us instantiate the theorem and its corollaries in a few special cases. If we make binary predictions with $\mathcal{Y} = \{0, 1\}$, then $C_k = k = \operatorname{diam}(\mathcal{Y}) = 1$, and Theorem 12.5.6 implies that

$$p_{\mathrm{cal,low}}(f) \le p_{\mathrm{cal}}(f) \le p_{\mathrm{cal,low}}(f) + 2\sqrt{p_{\mathrm{cal,low}}(f)}.$$

For k-class multiclass classification, where we identify $\mathcal{Y} = \{e_1, \ldots, e_k\}$ with the k standard basis vectors, we have the bounds

$$p_{\mathrm{cal,low}}(f) \le p_{\mathrm{cal}}(f) \le p_{\mathrm{cal,low}}(f) + 2\sqrt{k p_{\mathrm{cal,low}}(f)},$$

so long as we measure calibration errors with respect to the ℓ_1-norm, that is, $\|y - f(x)\|_1$, because $\operatorname{diam}(\mathcal{Y}) \le 1$ and $C_k = \|\mathbf{1}\|_\infty = 1$.

JCD Comment: Remark on sharpness here.

The equivalence between calibration error and the calibration distance

We can rewrite the calibration error $\mathrm{CE}(S, \mathcal{A})$ relative to partitions in the definition (12.2.6) as the supremum over a collection $\mathcal{W}_{\mathcal{A}}$ of functions of the form $w(s) = v\mathbf{1}\{s \in A\}$, where $\|v\|_* \le 1$, so that $\mathrm{CE}(S, \mathcal{W}_{\mathcal{A}}) = \sup_{w \in \mathcal{W}_{\mathcal{A}}} \mathbb{E}[\langle w(S), Y - S\rangle] = \sum_{A \in \mathcal{A}} \mathbb{E}[\|\mathbb{E}[Y \mid S] - S\|]$. Relaxing this supremum, and removing the infimum over partitions, we might expect a similar relationship to Theorem 12.5.6 to hold. Via a duality argument that the definition (12.5.3) of the lower calibration error as an infimum over joint distributions makes possible, we can directly relate the measures.

Theorem 12.5.9. *Let $\mathcal{Y} \subset \mathbb{R}^k$ have finite diameter in the norm $\|\cdot\|$ and $\mathcal{W}_{\|\cdot\|}$ be the collection (12.2.5) of bounded Lipschitz functions. Then*

$$\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) \le (1 + \mathrm{diam}(\mathcal{Y})) \cdot p_{\mathrm{cal,low}}(f).$$

Conversely, let $\mathcal{Y} = \{e_1, \ldots, e_k\}$ and define $C_k := \|\mathbf{1}_k\|_* \max\{1, \mathrm{diam}(\mathcal{Y})\}$. Then

$$d_{\mathrm{cal,low}}(f) \le C_k \cdot \mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}).$$

This proof, while nontrivial, is more elementary than the others in this chapter, so we present it here. Before giving it, however, we give a few corollaries that give a fuller picture of the relationships between the different calibration measures we have developed. These show how, for the case of k-class multiclass classification where we identify $\mathcal{Y} = \{e_1, \ldots, e_k\}$ with the standard basis vectors, the distance to calibration (12.2.3) and penalized calibration (12.5.2) provide essentially equivalent measures of calibration error, and that these in turn are equivalent to the calibration error with respect to the collection of bounded Lipschitz functions.

We first give a corollary for the penalized calibration (12.5.2).

Corollary 12.5.10. *Let $\mathcal{Y} = \{e_1, \ldots, e_k\}$ and $\|\cdot\| = \|\cdot\|_1$ be the ℓ_1-norm. Then for any $f : \mathcal{X} \to \mathrm{Conv}(\mathcal{Y})$, we have*

$$\frac{1}{2}\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) \le p_{\mathrm{cal}}(f) \le \mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) + 2\sqrt{k\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|})}.$$

Proof Let $\mathcal{W} = \mathcal{W}_{\|\cdot\|}$ for shorthand. Theorem 12.5.9 gives $\mathrm{CE}(f, \mathcal{W}) \le 2p_{\mathrm{cal,low}}(f)$, and $p_{\mathrm{cal,low}}(f) \le p_{\mathrm{cal}}(f)$, giving the lower bound. For the upper bound, Corollary 12.5.7 gives $p_{\mathrm{cal}}(f) \le p_{\mathrm{cal,low}}(f) + 2\sqrt{k}\sqrt{p_{\mathrm{cal,low}}(f)}$, then using that $p_{\mathrm{cal,low}}(f) \le d_{\mathrm{cal,low}}(f)$ and the second part of Theorem 12.5.9 gives the corollary. □

The same argument implies the following analogue for the distance to calibration (12.2.3).

Corollary 12.5.11. *Let $\mathcal{Y} = \{e_1, \ldots, e_k\}$ and $\|\cdot\| = \|\cdot\|_1$ be the ℓ_1-norm. Then for any $f : \mathcal{X} \to \mathrm{Conv}(\mathcal{Y})$, we have*

$$\frac{1}{2}\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) \le d_{\mathrm{cal}}(f) \le \mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) + 2\sqrt{k\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|})}.$$

Proof of Theorem 12.5.9

The proof of the upper bound is fairly straightforward. For any $w \in \mathcal{W}_{\|\cdot\|}$, we have

$$\begin{aligned}
\mathbb{E}[\langle w(S), Y - S\rangle] &= \mathbb{E}[\langle w(S), V - S\rangle] + \mathbb{E}[\langle w(S) - w(V), Y - V\rangle] + \mathbb{E}[\langle w(V), Y - V\rangle] \\
&\le \mathbb{E}[\|V - S\|] + \mathrm{diam}(\mathcal{Y})\mathbb{E}[\|V - S\|] + \mathbb{E}[\langle w(V), \mathbb{E}[Y \mid V] - V\rangle] \\
&\le (1 + \mathrm{diam}(\mathcal{Y}))\mathbb{E}[\|V - S\|] + \mathbb{E}[\|\mathbb{E}[Y \mid V] - V\|].
\end{aligned}$$

To prove the converse requires more; we present most of the argument for an arbitrary discrete space $\mathcal{Y}$ and specialize to the multiclass setting only at the end. The starting point is to reduce the problem to a discrete problem over probability mass functions rather than general distributions, as then it is much easier to apply the standard tools of convex duality. Consider the value

$$d_{\mathrm{cal,low}}(S) = \inf_V \left\{\mathbb{E}[\|S - V\|] \;\; \text{s.t.} \;\; \mathbb{E}[Y \mid V] = V\right\}.$$

Let $b \in \mathbb{N}$ and $\mathcal{S}_b$ be a (minimal) $1/b$ covering $\{s_1, \ldots, s_N\}$ of $\mathrm{Conv}(\mathcal{Y})$, and define S_b to be the projection of S to the nearest s_i. Then $\|S - V\| = \|S_b - V\| \pm \frac{1}{b}$, and

$$d_{\mathrm{cal,low}}(S) = \inf_V \left\{\mathbb{E}[\|S_b - V\|] \;\; \text{s.t.} \;\; \mathbb{E}[Y \mid V] = V\right\} \pm \frac{1}{b}.$$

Now, if we replace the infimum over arbitrary joint distributions of (S_b, Y, V) leaving the marginal (S_b, Y) unchanged (with V calibrated) with an infimum over only discrete distributions on V, we have

$$d_{\mathrm{cal,low}}(S) \le \inf_{V \text{ finitely supported}} \left\{\mathbb{E}[\|S_b - V\|] \;\; \text{s.t.} \;\; \mathbb{E}[Y \mid V] = V\right\} + \frac{1}{b}. \tag{12.5.5}$$

Notably, the infimum is non-empty, as we can always choose $V = Y$.

With the problem (12.5.5) in hand, we can write a finite dimensional optimization problem whose optimal value is the discretized infimum on the right side. Without loss of generality assuming that S is finitely supported, we let $p_{sy} = \mathbb{P}(S = s, Y = y)$ be the probability mass function of (S, Y). Then introducing the joint distribution Q with p.m.f. $q_{syv} = Q(S = s, Y = y, V = v)$, the infimum (12.5.5) has the constraint that $\sum_v q_{syv} = p_{sy}$. Then $\mathbb{E}[\|S - V\|] = \sum_{s,y,v} q_{syv} \|s - v\|$ and the calibration constraint $\mathbb{E}[Y \mid V] = V$ is equivalent to the equality constraint that $\sum_{s,y} q_{syv}(y - v) = 0$ for each v. This yields the convex optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{s,y,v} q_{syv} \|s - v\| \\ \text{subject to} & \sum_v q_{syv} = p_{sy}, \quad q \succeq 0, \quad \sum_{y,q} q_{syv}(y - v) = 0 \text{ for all } v \end{array} \tag{12.5.6}$$

in the variable q. We take the dual of this problem. Taking Lagrange multipliers λ_{sy} for each equality constraint that $\sum_v q_{syv} = p_{sy}$, $\theta_{syv} \ge 0$ for the nonnegativity constraints on q, and $\beta_v \in \mathbb{R}^k$ for each equality constraint that $0 = \sum_{s,y} q_{syv}(y - v)$, we have Lagrangian

$$\begin{aligned} &\mathcal{L}(q, z, \lambda, \theta, \beta) \\ &= \sum_{s,y,v} q_{syv} \|s - v\| + \sum_{s,y,v} q_{syv} \beta_v^T (y - v) - \sum_{s,y} \lambda_{sy} \left(\sum_v q_{syv} - p_{sy} \right) - \langle \theta, q \rangle. \end{aligned}$$

Taking an infimum over q, we see that unless

$$\|s - v\| + \beta_v^T (y - v) - \lambda_{sy} - \theta_{syv} = 0$$

for each triple (s, y, v), we have $\inf_q \mathcal{L}(q, \lambda, \theta, \beta) = -\infty$. The equality in the preceding display is equivalent to $\|s - v\| + \beta_v^T (y - v) \ge \lambda_{sy}$, so that eliminating $\theta \succeq 0$ variables, we have the dual

$$\begin{array}{ll} \text{maximize} & \sum_{s,y} \lambda_{sy} p_{sy} \\ \text{subject to} & \lambda_{sy} \le \|s - v\| + \beta_v^T (y - v), \quad \text{all } s, y, v \end{array}$$

to problem (12.5.6). Equivalently, recognizing that at the optimum we must saturate the constraints on λ via $\lambda_{sy} = \min_v\{\|s - v\| + \beta_v^T(y - v)\}$, we have

$$\text{maximize} \quad \sum_{s,y} p_{sy} \min_v \left\{\|s - v\| + \beta_v^T(y - v)\right\} \tag{12.5.7}$$

in the variables β_v, and strong duality obtains.

The dual problem (12.5.7) is the key to the final step in the proof. To make the functional notation clearer, let us fix any collection of vectors β_v and define $\lambda_y(s) = \min_v\{\|s - v\| + \beta_v^T(y - v)\}$ for each $y \in \mathcal{Y}$. If we can exhibit a C-Lipschitz function $s \mapsto w(s)$ that satisfies

$$\langle w(s), y - s\rangle \ge \lambda_y(s) \tag{12.5.8}$$

for each $y \in \mathcal{Y}$ and $\|w(s)\|_* \le C$, we will evidently have shown that

$$\sup_{w \in \mathcal{W}_{\|\cdot\|}} \mathbb{E}[\langle w(S), Y - S\rangle] \ge \frac{1}{C} d_{\text{cal,low}}(S),$$

by the dual formulation (12.5.7).

The functions λ_y are each 1-Lipschitz with respect to $\|\cdot\|$, as

$$\begin{aligned}\lambda_y(s) - \lambda_y(s') &\ge \min_v \left\{\|s - v\| + \beta_v^T(y - v) - \left\|s' - v\right\| - \beta_v^T(y - v)\right\} \\ &= \min_v \left\{\|s - v\| - \left\|s' - v\right\|\right\} \ge -\left\|s - s'\right\|,\end{aligned}$$

and similarly

$$\lambda_y(s) - \lambda_y(s') \le \max_v \left\{\|s - v\| + \beta_v^T(y - v) - \left\|s' - v\right\| - \beta_v^T(y - v)\right\} \le \left\|s - s'\right\|$$

by the triangle inequality. Here, we specialize to the particular multiclass classification case in which the set $\mathcal{Y} = \{e_1, \ldots, e_k\}$ consists of extreme points of the probability simplex, so that $s \in \text{Conv}(\mathcal{Y})$ means that $\langle \mathbf{1}, s\rangle = 1$ and $s \succeq 0$. Abusing notation slightly, let $\lambda_i = \lambda_{e_i}$ for $i = 1, \ldots, k$. Then define the function

$$w(s) := \begin{bmatrix} \lambda_1(s) \\ \vdots \\ \lambda_k(s) \end{bmatrix}.$$

By inspection, we have

$$\left\|w(s) - w(s')\right\|_* \le \left\|\left\|s - s'\right\| \mathbf{1}\right\|_* = \|\mathbf{1}\|_* \left\|s - s'\right\|.$$

Additionally, because $\lambda_i(s) \le \|s - e_i\|$ (take $v = e_i$ in the definition of λ_i), we have $\|w(s)\|_* \le \|\mathbf{1}\|_* \operatorname{diam}(\mathcal{Y})$. Finally, we have

$$\begin{aligned}\langle w(s), e_i - s\rangle &= (1 - s_i)\lambda_i(s) - \sum_{j \ne i} s_j \lambda_j(s) \\ &\ge (1 - s_i)\lambda_i(s) - \sum_{j \ne i} s_j \langle \beta_s, e_j - s\rangle\end{aligned}$$

because $\lambda_j(s) \le \langle \beta_s, e_j - s \rangle$ by taking $v = s$ in the definition of λ_j. Adding and subtracting $s_i \langle \beta_s, e_i - s \rangle$, we obtain

$$\begin{aligned}\langle w(s), e_i - s \rangle &\ge (1 - s_i)\lambda_i(s) - \sum_{j=1}^k s_j \langle \beta_s, e_j - s \rangle + s_i \langle \beta_s, e_i - s \rangle \\ &= (1 - s_i)\lambda_i(s) + s_i \langle \beta_s, e_i - s \rangle \ge \lambda_i(s),\end{aligned}$$

because $s \succeq 0$ and $\langle \beta_s, e_i - s \rangle \ge \lambda_i(s)$. This is the desired inequality (12.5.8).

12.6 Deferred technical proofs

Several of the proofs in this chapter rely on standard results from analysis and measure theory; we give these as base lemmas, as any book on graduate level real analysis (implicitly) contains them (see, e.g., Tao [164, Chapters 1.3 and 1.13] or Royden [154]).

Lemma 12.6.1 (Egorov's theorem)**.** *Let $f_n \to f$ in $L^p(P)$ for some $p \ge 1$. Then for each $\epsilon > 0$, there exists a set A of measure at least $P(A) \ge 1 - \epsilon$ such that $f_n \to f$ uniformly on A.*

Lemma 12.6.2 (Monotone convergence)**.** *Let $f_n : \mathcal{X} \to \mathbb{R}_+$ be a monotone increasing sequence of functions and $f(x) = \lim_n f_n(x)$ (which may be infinite). Then $\int f(x) d\mu(x) = \lim_n \int f_n(x) d\mu(x)$ for any measure μ.*

Lemma 12.6.3 (Density of Lipschitz functions)**.** *Let $\mathcal{C}_c^{\rm Lip}$ be the collection of compactly supported Lipschitz functions on $\mathbb{R}^k$ and P a probability distribution on $\mathbb{R}^k$. Then $\mathcal{C}_c^{\rm Lip}$ is dense in $L^p(P)$, that is, for each $\epsilon > 0$ and f with $\mathbb{E}_P[|f(X)|^p] < \infty$, there exists $g \in \mathcal{C}_c^{\rm Lip}$ with $\mathbb{E}_P[|g(X) - f(X)|^p]^{1/p} \le \epsilon$.*

12.6.1 Proof of Lemma 12.2.1

Let $\mathcal{W}_k$ be the collection of k-Lipschitz functions w with $\|w(s)\|_* \le 1$ for all s, and let $\mathcal{W}$ denote the collection of measurable functions with $\|w(s)\|_* \le 1$ for all s. Recall the defininition $\mathrm{CE}(g, \mathcal{W}_k) = \sup_{w \in \mathcal{W}_k} \mathbb{E}[\langle w(g(X)), Y - g(X) \rangle]$. Then if $f_n \to f$ in $L^1(P)$, by Egorov's theorem (Lemma 12.6.1), for each $\epsilon > 0$ there exists a set A with $P(A) \ge 1 - \epsilon$ and $f_n \to f$ uniformly on A. Then

$$\begin{aligned}&\mathbb{E}[\langle w(f_n(X)), Y - f_n(X) \rangle] \\ &= \mathbb{E}[\langle w(f_n(X)), Y - f_n(X) \rangle \mathbf{1}\left\{X \in A\right\}] + \mathbb{E}[\langle w(f_n(X)), Y - f_n(X) \rangle \mathbf{1}\left\{X \in A^c\right\}] \\ &\ge \mathbb{E}[\langle w(f_n(X)), Y - f_n(X) \rangle \mathbf{1}\left\{X \in A\right\}] - \mathbb{E}[\|Y - f_n(X)\| \mathbf{1}\left\{X \in A^c\right\}] \qquad (12.6.1)\end{aligned}$$

because $\|w(s)\|_* \le 1$. As $| \|y - f_n(x)\| \mathbf{1}\{x \in A^c\} - \|y - f(x)\| \mathbf{1}\{x \in A^c\}| \le \|f(x) - f_n(x)\|$ by the triangle inequality, the last term in inequality (12.6.1) converges to $\mathbb{E}[\|Y - f(X)\| \mathbf{1}\left\{X \in A^c\right\}]$ as $n \to \infty$. Focusing on the first term in (12.6.1), for any $\epsilon_1 > 0$ the uniform convergence of f_n to f on A guarantees that for large enough n, we have

$$\begin{aligned}&\mathbb{E}[\langle w(f_n(X)), Y - f_n(X) \rangle] \\ &= \mathbb{E}[\langle w(f(X)), Y - f_n(X) \rangle \mathbf{1}\left\{X \in A\right\}] + \mathbb{E}[\langle w(f_n(X)) - w(f(X)), Y - f_n(X) \rangle \mathbf{1}\left\{X \in A\right\}] \\ &\ge \mathbb{E}[\langle w(f(X)), Y - f_n(X) \rangle \mathbf{1}\left\{X \in A\right\}] - k \sup_{x \in A} \|f(x) - f_n(x)\|_* \mathbb{E}[\|Y - f_n(X)\|] \\ &\ge \mathbb{E}[\langle w(f(X)), Y - f_n(X) \rangle \mathbf{1}\left\{X \in A\right\}] - \epsilon_1\end{aligned}$$

Adding and subtracting $f(X)$ in the final expectation, we have

$$\begin{aligned}
&\mathbb{E}[\langle w(f(X)), Y - f_n(X)\rangle \mathbf{1}\left\{X \in A\right\}] \\
&= \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle \mathbf{1}\left\{X \in A\right\}] + \mathbb{E}[\langle w(f(X)), f(X) - f_n(X)\rangle \mathbf{1}\left\{X \in A\right\}] \\
&\geq \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] - \mathbb{E}[\|Y - f(X)\| \mathbf{1}\left\{X \in A^c\right\}] - \mathbb{E}[\|f(X) - f_n(X)\|] \\
&\to \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] - \mathbb{E}[\|Y - f(X)\| \mathbf{1}\left\{X \in A^c\right\}].
\end{aligned}$$

Substituting these bounds into inequality (12.6.1), we have for any $\epsilon > 0$ that there exists a set A_ϵ with $P(A_\epsilon) \geq 1 - \epsilon$ and for which

$$\begin{aligned}
&\liminf_n \mathbb{E}[\langle w(f_n(X)), Y - f_n(X)\rangle] \\
&\geq \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] - 2\mathbb{E}[\|Y - f(X)\| \mathbf{1}\left\{X \in A_\epsilon^c\right\}].
\end{aligned}$$

For each $m \in \mathbb{N}$, let $B_m = \bigcup_{n \leq m} A_{1/n}$. Certainly $P(B_m) \geq 1 - 1/m$, and $f_n \to f$ uniformly on B_m (as the guarantees on $A_{1/n}$ from Egorov's theorem apply); the same argument thus gives

$$\begin{aligned}
&\liminf_n \mathbb{E}[\langle w(f_n(X)), Y - f_n(X)\rangle] \\
&\geq \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] - 2\mathbb{E}[\|Y - f(X)\| \mathbf{1}\left\{X \in B_m^c\right\}].
\end{aligned}$$

Because B_m is an increasing sequence of sets with $P(B_m) \geq 1 - 1/m$, the limit $B_\infty = \bigcup_m B_m$ satisfies $P(B_\infty) = 1$. For any $x \in B_\infty$, we see that $x \in B_m$ for some finite m; trivially, for $x \in B_\infty$ we thus have $\|y - f(x)\| \mathbf{1}\left\{x \notin B_m\right\} \to \|y - f(x)\| \mathbf{1}\left\{x \notin B_\infty\right\} = 0$ as $m \to \infty$. Said differently, except on a null set, we have $\|y - f(x)\| \mathbf{1}\left\{x \notin B_m\right\} \to 0$ for P-almost all (x, y), and this is certainly dominated by $\|y - f(x)\|$. Lebesgue's dominated convergence theorem then implies $\mathbb{E}[\|Y - f(X)\| \mathbf{1}\left\{X \notin B_m\right\}] \to 0$ as $m \to \infty$. Summarizing, we have shown that for any $w \in \mathcal{W}_k$, we have

$$\liminf_n \mathbb{E}[\langle w(f_n(X)), Y - f_n(X)\rangle] \geq \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle].$$

By taking a supremum over $w \in \mathcal{W}_k$ in the last display and recognizing that $\epsilon > 0$ was arbitrary, we have shown that

$$\liminf_n \mathrm{CE}(f_n, \mathcal{W}_k) \geq \mathrm{CE}(f, \mathcal{W}_k)$$

for all $k < \infty$. By Lemma 12.6.3, for any integrable f and for each $\epsilon > 0$ there exists k such that

$$\sup_{w \in \mathcal{W}_k} \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] \geq \sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] - \epsilon.$$

and for this k we have

$$\liminf_n \mathrm{CE}(f_n, \mathcal{W}_k) \geq \mathrm{CE}(f, \mathcal{W}_k) \geq \mathrm{CE}(f, \mathcal{W}) - \epsilon.$$

Noting that $\mathrm{CE}(f_n, \mathcal{W}) \geq \mathrm{CE}(f_n, \mathcal{W}_k)$ for any k and taking $\epsilon \to 0$ gives the lemma.

12.6.2 Proof of Proposition 12.5.2

The proof that $\mathrm{CE}(\cdot, \mathcal{W}_{\|\cdot\|})$ identifies calibration (Definition 12.1, part (i)) is identical to the argument for Proposition 12.5.3, so we omit it.

Let $\mathcal{W} = \mathcal{W}_{\|\cdot\|}$ for shorthand, and consider a sequence of functions $f_n \to f$. Then

$$\mathrm{CE}(f, \mathcal{W}) - \mathrm{CE}(f_n, \mathcal{W}) \le \sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle - \langle w(f_n(X)), Y - f_n(X)\rangle]$$

and

$$\mathrm{CE}(f_n, \mathcal{W}) - \mathrm{CE}(f, \mathcal{W}) \le \sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(f_n(X)), Y - f_n(X)\rangle - \langle w(f(X)), Y - f(X)\rangle].$$

We focus on bounding the first display, as showing that the second tends to zero requires, *mutatis mutandis*, an identical argument.

Fix any $w \in \mathcal{W}$. Then

$$\begin{aligned}
&\mathbb{E}[\langle w(f(X)), Y - f(X)\rangle - \langle w(f_n(X)), Y - f_n(X)\rangle] \\
&= \mathbb{E}[\langle w(f(X)) - w(f_n(X)), Y - f(X)\rangle] + \mathbb{E}[\langle w(f_n(X)), f_n(X) - f(X)\rangle] \\
&\le \mathbb{E}[\min\{2, \|f(X) - f_n(X)\|\} \|Y - f(X)\|] + \mathbb{E}[\|f_n(X) - f(X)\|],
\end{aligned}$$

where the inequality follows because $\|w(s) - w(s')\|_* \le 2$ and $\|w(s) - w(s')\|_* \le \|s - s'\|$ for any s, s' by construction. The second expectation certainly tends to zero as $n \to \infty$, so we consider the first. Define $g_n(x, y) = \min\{2, \|f(x) - f_n(x)\|\} \|y - f(x)\|$. Then $g_n(x, y) \le g(x, y) = \|y - f(x)\|$, which has finite expectation by assumption. Moreover, Egorov's theorem (Lemma 12.6.1) guarantees that for each k, there is a set A_k with $P(A_k) \ge 1 - 1/k$ and for which $g_n \to 0$ uniformly on A_k (because $\mathbb{E}[\|f(X) - f_n(X)\|] \to 0$). Define $A_\infty = \bigcup_k A_k$, so that $P(A_\infty) = 1$, and $g_n(x, y) \to 0$ pointwise on A_∞. Then the dominated convergence theorem guarntees that

$$\mathbb{E}[g_n(X, Y)] = \mathbb{E}[g_n(X, Y)\mathbf{1}\{(X, Y) \in A_\infty\}] + \underbrace{\mathbb{E}[g_n(X, Y)\mathbf{1}\{(X, Y) \notin A_\infty\}]}_{=0} \to 0.$$

Notably, this convergence is independent of w, and so we obtain

$$\limsup_n \{\mathrm{CE}(f, \mathcal{W}) - \mathrm{CE}(f_n, \mathcal{W})\} \le 0.$$

A similar argument gives the converse bound.

12.6.3 Proof of Lemma 12.5.4

Define $f(s) = g(s)/\max\{1, \|g(s)\|\}$, so that $\mathbb{E}[\|g(s)\|_2] = \mathbb{E}[\langle f(s), g(s)\rangle]$. Using Lemma 12.6.3, we see that for each $n \in \mathbb{N}$ there exists a $C = C_n$-Lipschitz function (where $C < \infty$) w_n with $\mathbb{E}[\|w_n(S) - f(S)\|] \le \frac{1}{n}$, and w.l.o.g. we may assume $\|w_n(s)\|_2 \le 1$ (by projection if necessary, which is Lipschitzian). Then

$$\mathbb{E}[\|g(S)\|_2] = \mathbb{E}[\langle f(S), g(S)\rangle] = \mathbb{E}[\langle f(S) - w_n(S), g(S)\rangle] + \underbrace{\mathbb{E}[\langle w_n(S), g(S)\rangle]}_{=0}.$$

Note that $w_n \to f$ in $L^1(P)$. Then for any $\epsilon > 0$, an application of Egorov's theorem (Lemma 12.6.1) and that $\mathbb{E}[\|g(S)\|] < \infty$ gives that we can find sets A_ϵ with $P(A_\epsilon) \ge 1 - \epsilon$ and for which $w_n \to f$ uniformly on A_ϵ. Then

$$\begin{aligned}
\mathbb{E}[\|g(S)\|_2] &= \mathbb{E}[\langle f(S) - w_n(S), g(S)\rangle \mathbf{1}\{S \in A_\epsilon\}] + \mathbb{E}[\|g(S)\|_2 \mathbf{1}\{S \notin A_\epsilon\}] \\
&\le \mathbb{E}\Big[\sup_{s \in A_\epsilon} \|f(s) - w_n(s)\|_2 \|g(S)\|_2 \mathbf{1}\{S \in A_\epsilon\}\Big] + \mathbb{E}[\|g(S)\|_2 \mathbf{1}\{S \notin A_\epsilon\}] \\
&\to \mathbb{E}[\|g(S)\|_2 \mathbf{1}\{S \notin A_\epsilon\}].
\end{aligned}$$

as $n \uparrow \infty$. We now employ the same device we use in the proof of Lemma 12.2.1. For $m \in \mathbb{N}$, let $B_m = \bigcup_{n \le m} A_{1/n}$. Then $w_n \to f$ uniformly on B_m, and so $\mathbb{E}[\|g(S)\|_2] \le \mathbb{E}[\|g(S)\|_2 \mathbf{1}\{S \notin B_m\}]$, that is, $\mathbb{E}[\|g(S)\|_2 \mathbf{1}\{S \in B_m\}] = 0$. Monotone convergence implies $0 = \lim_{m\to\infty} \mathbb{E}[\|g(S)\|_2 \mathbf{1}\{S \in B_m\}] = \mathbb{E}[\|g(S)\|_2 \mathbf{1}\{S \in B_\infty\}]$ where $B_\infty = \bigcup_m B_n$. As $P(B_\infty) = 1$ by continuity of measure, we have $\mathbb{E}[\|g(S)\|_2] = 0$, giving the lemma.

12.6.4 Proof of Theorem 12.5.6

The following lemma gives the lower bound in the theorem and is fairly straightforward.

Lemma 12.6.4. *For $S = f(X)$, we have*

$$p_{\text{cal,up}}(f) \le d_{\text{cal,up}}(f) \le \mathsf{pce}(S). \tag{12.6.2}$$

Proof Fix any partition $\mathcal{A}$, and define $\mathsf{q}_{\mathcal{A}}(s)$ to be the (unique) set A such that $s \in A$ (so we quantize s). Then set $g(s) = \mathbb{E}[Y \mid S \in \mathsf{q}_{\mathcal{A}}(s)]$ to be the expectation of Y conditional on S being in the same partition element as s. Then $g(S) = \mathbb{E}[Y \mid g(S)]$ with probability 1, so that g is perfectly calibrated, and

$$\begin{aligned}
p_{\text{cal,up}}(f) \le d_{\text{cal,up}}(f) &\le \mathbb{E}[\|S - g(S)\|] \\
&= \sum_{A \in \mathcal{A}} \mathbb{E}[\|S - \mathbb{E}[Y \mid S \in A]\| \mathbf{1}\{S \in A\}] \\
&\le \sum_{A \in \mathcal{A}} \mathbb{E}\left[(\|S - \mathbb{E}[S \mid S \in A]\| + \|\mathbb{E}[S - Y \mid S \in A]\|) \mathbf{1}\{S \in A\}\right] \\
&\le \sum_{A \in \mathcal{A}} \operatorname{diam}(A) \mathbb{P}(S \in A) + \sum_{A \in \mathcal{A}} \|\mathbb{E}[(S - Y)\mathbf{1}\{S \in A\}]\|.
\end{aligned}$$

Taking an infimum gives the claim (12.6.2). □

To prove the claimed upper bound requires more work. For pedagogical reasons, let us attempt to prove a similar upper bound relating $\mathsf{pce}(S)$ to $p_{\text{cal,low}}(f)$. We might begin with a partition $\mathcal{A}$ with maximal diameter $\operatorname{diam}(A) \le \epsilon$ for $A \in \mathcal{A}$, and for random variables (S, V, Y), begin with the first term in the partition error, whence

$$\begin{aligned}
\operatorname{CE}(S, \mathcal{A}) &\le \sum_{A \in \mathcal{A}} \|\mathbb{E}[(S - V)\mathbf{1}\{S \in A\}]\| + \|\mathbb{E}[(V - Y)\mathbf{1}\{S \in A\}]\| \\
&\le \mathbb{E}[\|S - V\|] + \sum_{A \in \mathcal{A}} \|\mathbb{E}[(V - Y)\mathbf{1}\{V \in A\}]\| + \sum_{A \in \mathcal{A}} \|\mathbb{E}[(V - Y)(\mathbf{1}\{S \in A\} - \mathbf{1}\{V \in A\})]\| \\
&\le \mathbb{E}[\|S - V\|] + \mathbb{E}[\|\mathbb{E}[Y \mid V] - V\|] + \sum_{A \in \mathcal{A}} \|\mathbb{E}[(V - Y)(\mathbf{1}\{S \in A\} - \mathbf{1}\{V \in A\})]\|
\end{aligned}$$

by Jensen's inequality applied to conditional expectations, once we recognize $\mathbb{E}[(Y - V)\mathbf{1}\{V \in A\}] = \mathbb{E}[(\mathbb{E}[Y \mid V] - V)\mathbf{1}\{V \in A\}]$. For the final term, a straightforward computation yields

$$\begin{aligned}
\sum_{A \in \mathcal{A}} \|\mathbb{E}[(V - Y)(\mathbf{1}\{S \in A\} - \mathbf{1}\{V \in A\})]\| &\le \operatorname{diam}(\mathcal{Y}) \sum_{A \in \mathcal{A}} [\mathbb{P}(S \in A, V \notin A) + \mathbb{P}(S \notin A, V \notin A)] \\
&= 2 \operatorname{diam}(\mathcal{Y}) \mathbb{P}(S \text{ and } V \text{ belong to different } A \in \mathcal{A}).
\end{aligned}$$

If S and V had continuous distributions, we would expect the probability that they fail to belong to the same partition elements to scale as $\mathbb{E}[\|S - V\|]$. This may fail, but to rectify the issue, we can randomize.

Consequently, let us consider the *randomized partition error*, which we index with $\varepsilon > 0$ and for $U \sim \mathsf{Uniform}[-1, 1]^k$ define as

$$\mathsf{rpce}_\varepsilon(S) := \inf_{\mathcal{A}} \left\{ \sum_{A \in \mathcal{A}} \|\mathbb{E}[(S - Y)\mathbf{1}\left\{S + \varepsilon U \in A\right\}]\| + \sum_{A \in \mathcal{A}} \operatorname{diam}(A) \mathbb{P}(S \in A) \right\}. \tag{12.6.3}$$

(The choice of uniform $[-1, 1]^k$ is only made for convenience in the calculations to follow.) Letting $c_k = \|\mathbf{1}_k\|_*$, we see immediately that

$$\mathsf{pce}(S) \le \mathsf{rpce}_\varepsilon(S) + c_k \varepsilon$$

for all $\varepsilon \ge 0$. We can say more.

Lemma 12.6.5. *Let $\varepsilon > 0$. Then for any random variable V,*

$$\mathsf{rpce}_\varepsilon(S) \le \mathbb{E}[\|S - V\|] + \mathbb{E}[\|\mathbb{E}[Y \mid V] - V\|] + \frac{2k}{\varepsilon} \mathbb{E}[\|Y - S\| \, \|V - S\|_\infty].$$

Note that by combining Lemma 12.6.5 with the display above and recognizing that $\|Y - S\| \le \operatorname{diam}(\mathcal{Y})$ with probability 1, we have the theorem.

Proof We replicate the calculation bounding $\mathrm{CE}(S, \mathcal{A})$ above, but while allowing the randomization. Let $\mathcal{A}$ be a partition of $\mathbb{R}^k$ into hypercubes of width ε, that is, $[-\varepsilon, \varepsilon]^k + \varepsilon z$, where $z \in 2\mathbb{Z}^k$ ranges over integer vectors with even entries. Then $\operatorname{diam}(A) \le c_k \varepsilon$, and

$$\begin{aligned}
&\|\mathbb{E}[(S - Y)\mathbf{1}\left\{S + \varepsilon U \in A\right\}]\| \\
&\le \|\mathbb{E}[(S - V)\mathbf{1}\left\{S + \varepsilon U \in A\right\}]\| + \|\mathbb{E}[(V - Y)\mathbf{1}\left\{V + \varepsilon U \in A\right\}]\| \\
&\quad + \|\mathbb{E}[(V - Y)(\mathbf{1}\left\{S + \varepsilon U \in A\right\} - \mathbf{1}\left\{S + \varepsilon U \in A\right\})]\| \\
&\le \|\mathbb{E}[(S - V)\mathbf{1}\left\{S + \varepsilon U \in A\right\}]\| + \|\mathbb{E}[(V - Y)\mathbf{1}\left\{V + \varepsilon U \in A\right\}]\| \\
&\quad + \mathbb{E}\left[\|V - Y\| \cdot (\mathbb{P}(V + \varepsilon U \in A, S + \varepsilon U \notin A \mid V, S) + \mathbb{P}(S + \varepsilon U \in A, V + \varepsilon U \notin A \mid V, S, Y))\right]
\end{aligned}$$

Summing over sets A and using the triangle inequality and that $S + \varepsilon U \in A$ for some A, we find

$$\begin{aligned}
\sum_{A \in \mathcal{A}} \|\mathbb{E}[(S - Y)\mathbf{1}\left\{S + \varepsilon U \in A\right\}]\| &\le \mathbb{E}[\|S - V\|] + \mathbb{E}[\|\mathbb{E}[Y \mid V] - V\|] \\
&\quad + 2\mathbb{E}\left[\|V - Y\| \sum_{A \in \mathcal{A}} \mathbb{P}(V + \varepsilon U \in A, S + \varepsilon U \notin A \mid V, S, Y)\right].
\end{aligned} \tag{12.6.4}$$

We now may bound the probability in inequality (12.6.4). Recall that $A = [-\varepsilon, \varepsilon]^k + \varepsilon z$ for some $z \in 2\mathbb{Z}^k$, and fix $v, s \in \mathbb{R}^k$. Let $B = [-1, 1]^k$ be the ℓ_∞ ball. Then

$$\begin{aligned}
\mathbb{P}(v + \varepsilon U \in B, s + \varepsilon U \notin B) &= \mathbb{P}(U \notin \varepsilon^{-1}(B - s) \mid U \in \varepsilon^{-1}(B - v)) \mathbb{P}(v + \varepsilon U \in B) \\
&\le \frac{k \|s - v\|_\infty}{\varepsilon} \mathbb{P}(v + \varepsilon U \in B),
\end{aligned} \tag{12.6.5}$$

where inequality (12.6.5) follows because if $s, v \in \mathbb{R}^k$ are the centers of two ℓ_∞ balls B_s and B_v of radius 1, and if $\delta = \|s - v\|_\infty$, then the volume of $B_v \setminus B_s$ is at most $k\delta^k / \delta^{k-1} = k\delta$. (See

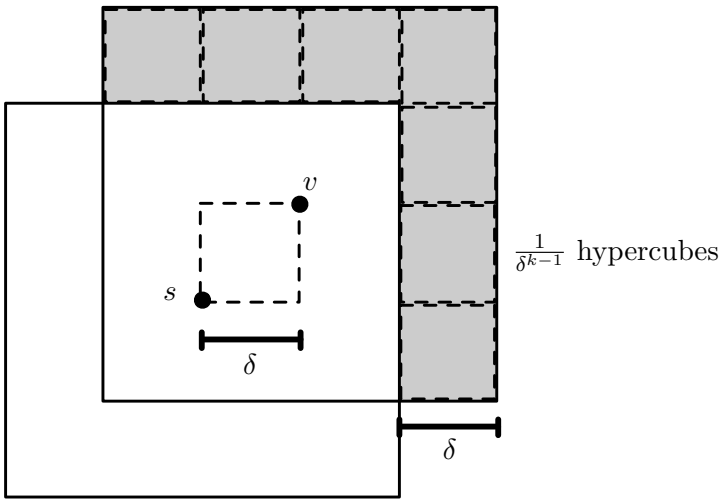

Figure 12.2. The volume argument in inequality (12.6.5). In k dimensions, the hypercube of side-length δ can be replicated $1/\delta^{k-1}$ times on each exposed base of the cube centered at v, where $\delta = \|s - v\|_\infty$. There are at most k such faces, giving volume at most $k\delta^k/\delta^{k-1} = k\delta$ to the gray region.

Figure 12.2. The k-dimensional surface area of one side of a hypercube of radius δ is $2k\delta^{k-1}$, and we can put at most $1/\delta^{k-1}$ boxes in each facial part of the grey region.)

Substituting inequality (12.6.5) into the bound (12.6.4) and conditioning and deconditioning on V, S, we find that

$$\begin{aligned}
&\sum_{A \in \mathcal{A}} \|\mathbb{E}[(S - Y)\mathbf{1}\{S + \varepsilon U \in A\}]\| \\
&\le \mathbb{E}[\|S - V\|] + \mathbb{E}[\|\mathbb{E}[Y \mid V] - V\|] + \frac{2k}{\varepsilon}\mathbb{E}\left[\|V - Y\| \sum_{A \in \mathcal{A}} \|V - S\|_\infty \mathbf{1}\{V + \varepsilon U \in A\}\right] \\
&= \mathbb{E}[\|S - V\|] + \mathbb{E}[\|\mathbb{E}[Y \mid V] - V\|] + \frac{2k}{\varepsilon}\mathbb{E}[\|Y - V\| \, \|V - S\|_\infty].
\end{aligned}$$

Taking an infimum over partitions $\mathcal{A}$ gives the lemma. □

12.7 Bibliography

Draft: Calibration remains an active research area. The initial references for online calibration are Foster and Vohra [84], Dawid and Vovk [59]. The idea of calibeating is most present in Foster and Hart [85]. Our proof of calibeating is based on Kumar et al. [123]. Blasiok et al. [31] demonstrate the equivalence of the different metrics for measuring calibration, focusing on the case of binary prediction; the extension to vector-valued Y appears to be new. The ideas of the postprocessing gap and also descend from Blasiok et al. [32], and the connections with general proper losses also appear to be new. Propositions 12.5.1, 12.5.2, and 12.5.3 are new in that they are the first to demonstrate that the measures are valid calibration measures (Definition 12.1, part (i)).

JCD Comment: A few more things to add either in the bibliography or the introduction to the section:

1. We only really do calibration for binary/multiclass things. One would also really like to predict full distributions P_t on general outcomes Y, which is harder (nearly impossible) to do in any conditional sense.
2. It's much easier to do predictive inference (cover) because don't need accuracy
3. Maybe comment on variants for top entry (from multiclass to binary) classification and why that is important. Maybe in the middle, maybe here.

12.8 Exercises

JCD Comment: Add a uniform convexity version of Proposition 12.3.5 as an exercise.

JCD Comment: Can we add an exercise about achieving weak calibration for different classes of functions?

JCD Comment: A few potential exercises:

(i) Deal with any class $\mathcal{W}$ for which $\mathbb{E}[\langle w, f\rangle] = 0$ for all $w \in \mathcal{W}$ means $f = 0$, then still get a continuous calibration measure

JCD Comment: Exercise: do Aaditya's top-class calibration approach.

JCD Comment: Do we need more commentary on calibeating? Maybe an exercise on empirics? Project ideas: calibeating with witnesses in higher dimensions, doing calibeating in higher dimensions, optimality results / lower bounds.

JCD Comment: Do Example 3.2 of Kumar et al. [123] as exercise

JCD Comment: Coding and empirical exercises on calibration?

JCD Comment: Remark on impossibility of inference of ece? Exercises on its impossibility too, perhaps, and one-sided estimation of it. And maybe some minimax lower bounds on the Lipschitz one as well I think.

JCD Comment: Exercise potential: let $\mathcal{W}$ be a collection from an RKHS

Chapter 13

Surrogate Risk Consistency: the Classification Case

I. The setting: supervised prediction problem

(a) Have data coming in pairs (X, Y) and a loss $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ (can have more general losses)

(b) Often, it is hard to minimize L (for example, if L is non-convex), so we use a surrogate φ

(c) We would like to compare the risks of functions $f : \mathcal{X} \to \mathbb{R}$:

$$R_\varphi(f) := \mathbb{E}[\varphi(f(X), Y)] \quad \text{and} \quad R(f) := \mathbb{E}[L(f(X), Y)]$$

In particular, when does minimizing the surrogate give minimization of the true risk?

(d) Our goal: when we define the Bayes risks R^*_φ and R^*

Definition 13.1 (Fisher consistency)**.** *We say the loss φ is* Fisher consistent *if for any sequence of functions f_n*

$$R_\varphi(f_n) \to R^*_\varphi \quad \textit{implies} \quad R(f_n) \to R^*$$

II. Classification case

(a) We focus on the binary classification case so that $Y \in \{-1, 1\}$

1. Margin-based losses: predict sign correctly, so for $s \in \mathbb{R}$,

$$L(s, y) = \mathbf{1}\{sy \le 0\} \quad \text{and} \quad \varphi(s, y) = \phi(ys).$$

2. Consider conditional version of risks. Let $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$ be conditional probability, then

$$\begin{aligned} R(f) &= \mathbb{E}[\mathbf{1}\{f(X)Y \le 0\}] = \mathbb{P}(\text{sign}(f(X)) \ne Y) \\ &= \mathbb{E}\left[\eta(X)\mathbf{1}\{f(X) \le 0\} + (1 - \eta(X))\mathbf{1}\{f(X) \ge 0\}\right] = \mathbb{E}[\ell(f(X), \eta(X))] \end{aligned}$$

and

$$\begin{aligned} R_\phi(f) &= \mathbb{E}[\phi(Yf(X))] \\ &= \mathbb{E}\left[\eta(X)\phi(f(X)) + (1 - \eta(X))\phi(-f(X))\right] = \mathbb{E}[\ell_\phi(f(X), \eta(X))] \end{aligned}$$

where we have defined the conditional risks

$$\ell(s, \eta) = \eta\mathbf{1}\{s \le 0\} + (1 - \eta)\mathbf{1}\{s \ge 0\} \quad \text{and} \quad \ell_\phi(s, \eta) = \eta\phi(s) + (1 - \eta)\phi(-s).$$

3. Note the minimizer of ℓ: we have $s^*(\eta) = \sign(\eta - 1/2)$, and $f^*(X) = \sign(\eta(X) - 1/2)$ minimizes risk $R(f)$ over all f
4. Minimizing f can be achieved pointwise, and we have

$$R^* = \mathbb{E}[\inf_s \ell(s, \eta(X))] \quad \text{and} \quad R^*_\phi = \mathbb{E}[\inf_s \ell_\phi(s, \eta(X))].$$

(b) **Example 13.0.1** (Exponential loss)**:** Consider the exponential loss, used in AdaBoost (among other settings), which sets $\phi(s) = e^{-s}$. In this case, we have

$$\operatorname*{argmin}_s \ell_\phi(s, \eta) = \frac{1}{2} \log \frac{\eta}{1 - \eta} \quad \text{because} \quad \frac{\partial}{\partial s} \ell_\phi(s, \eta) = -\eta e^{-s} + (1 - \eta) e^s.$$

Thus $f^*_\phi(x) = \frac{1}{2} \log \frac{\eta(x)}{1 - \eta(x)}$, and this is Fisher consistent. $\diamond$

(c) Classification calibration

1. Consider pointwise versions of risk (all that is necessary, turns out)
2. Define the infimal conditional ϕ-risks as

$$\ell^*_\phi(\eta) := \inf_s \ell_\phi(s, \eta) \quad \text{and} \quad \ell^{\text{wrong}}_\phi(\eta) := \inf_{s(\eta - 1/2) \le 0} \ell_\phi(s, \eta).$$

3. Intuition: if we always have $\ell^*_\phi(\eta) < \ell^{\text{wrong}}_\phi(\eta)$ for all η, we should do fine
4. Define the sub-optimality function $H : [0, 1] \to \mathbb{R}$

$$H(\delta) := \ell^{\text{wrong}}_\phi \left(\frac{1 + \delta}{2} \right) - \ell^*_\phi \left(\frac{1 + \delta}{2} \right).$$

Definition 13.2. *The margin-based loss ϕ is* classification calibrated *if $H(\delta) > 0$ for all $\delta > 0$. Equivalently, for any $\eta \neq \frac{1}{2}$, we have $\ell^*_\phi(\eta) < \ell^{\text{wrong}}_\phi(\eta)$.*

5. **Example** (Example 13.0.1 continued)**:** For the exponential loss, we have

$$\ell^{\text{wrong}}_\phi(\eta) = \inf_{s(2\eta - 1) \le 0} \left\{ \eta e^{-s} + (1 - \eta) e^s \right\} = e^0 = 1$$

while the unconstrained minimal conditional risk is

$$\ell^*_\phi(\eta) = \eta \sqrt{\frac{1 - \eta}{\eta}} + (1 - \eta) \sqrt{\frac{\eta}{1 - \eta}} = 2 \sqrt{\eta(1 - \eta)},$$

so that $H(\delta) = 1 - \sqrt{1 - \delta^2} \ge \frac{1}{2} \delta^2$. $\diamond$

Example 13.0.2 (Hinge loss)**:** We can also consider the hinge loss, which is defined as $\phi(s) = [1 - s]_+$. We first compute the minimizers of the conditional risk; we have

$$\ell_\phi(s, \eta) = \eta \left[1 - s \right]_+ + (1 - \eta) \left[1 + s \right]_+,$$

whose unique minimizer (for $\eta \notin \{0, \frac{1}{2}, 1\}$) is $s(\eta) = \sign(2\eta - 1)$. We thus have

$$\ell^*_\phi(\eta) = 2 \min\{\eta, 1 - \eta\} \quad \text{and} \quad \ell^{\text{wrong}}_\phi(\eta) = \eta + (1 - \eta) = 1.$$

We obtain $H(\delta) = 1 - \min\{1 + \delta, 1 - \delta\} = \delta$. $\diamond$

Comparing to the sub-optimality function for exp-loss, is tighter.

6. Pictures: use exponential loss, with η and without.

(d) Our goal: using classification calibration, find some function ψ such that $\psi(R_\phi(f) - R_\phi^*) \le R(f) - R^*$, where $\psi(\delta) > 0$ for all $\delta > 0$. Can we get a convex version of H, them maybe use Jensen's inequality to get the results? Turns out we will be able to do this.

III. Some necessary asides on convex analysis

(a) Epigraphs and closures

1. For a function f, the epigraph $\operatorname{epi} f$ is the set of points (x, t) such that $f(x) \le t$
2. A function f is said to be *closed* if its epigraph is closed, which for convex f occurs if and only if f is lower semicontinuous (meaning $\liminf_{x \to x_0} f(x) \ge f(x_0)$)
3. Note: a one-dimensional closed convex function is continuous

   Lemma 13.0.3. *Let $f : \mathbb{R} \to \mathbb{R}$ be convex. Then f is continuous on the interior of its domain.*

   (Proof in notes; just give a picture)

   Lemma 13.0.4. *Let $f : \mathbb{R} \to \mathbb{R}$ be closed convex. Then f is continuous on its domain.*
4. The *closure* of a function f is the function $\operatorname{cl} f$ whose epigraph is the closed convex hull of $\operatorname{epi} f$ (picture)

(b) Conjugate functions (Fenchel-Legendre transform)

1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an (arbitrary) function. Its conjugate (or Fenchel-Legendre conjugate) is defined to be
$$f^*(s) := \sup_t \left\{ \langle t, s \rangle - f(t) \right\}.$$
   (Picture here) Note that we always have $f^*(s) + f(t) \ge \langle s, t \rangle$, or $f(t) \ge \langle s, t \rangle - f^*(s)$
2. The Fenchel biconjugate is defined to be $f^{**}(t) = \sup_s \{ \langle t, s \rangle - f^*(s) \}$ (Picture here, noting that $f'(t) = -s$ implies $f^*(t) = ts - f(t)$)
3. In fact, the biconjugate is the largest closed convex function smaller than f:

   Lemma 13.0.5. *We have*
$$f^{**}(x) = \sup_{a \in \mathbb{R}^d, b \in \mathbb{R}} \left\{ \langle a, x \rangle - b : \langle a, t \rangle - b \le f(t) \text{ for all } t \right\}.$$

   Proof Let $A \subset \mathbb{R}^d \times \mathbb{R}$ denote all the pairs (a, b) minorizing f, that is, those pairs such that $f(t) \ge \langle a, t \rangle - b$ for all t. Then we have
$$\begin{aligned} (a, b) \in A &\Leftrightarrow f(t) \ge \langle a, t \rangle - b \text{ for all } t \\ &\Leftrightarrow b \ge \langle a, t \rangle - f(t) \text{ all } t \\ &\Leftrightarrow b \ge f^*(a) \text{ and } a \in \operatorname{dom} f^*. \end{aligned}$$
   Thus we obtain the following sequence of equalities:
$$\begin{aligned} \sup_{(a,b) \in A} \left\{ \langle a, t \rangle - b \right\} &= \sup \left\{ \langle a, t \rangle - b : a \in \operatorname{dom} f^*, -b \le -f^*(a) \right\} \\ &= \sup \left\{ \langle a, t \rangle - f^*(a) \right\}. \end{aligned}$$
   So we have all the supporting hyperplanes to the graph of f as desired. □

4. Other interesting lemma:

   Lemma 13.0.6. *Let h be either (i) continuous on $[0,1]$ or (ii) non-decreasing on $[0,1]$. (And set $h(1+\delta) = +\infty$ for $\delta > 0$.) If h satisfies $h(t) > 0$ for $t > 0$ and $h(0) = 0$, then $f(t) = h^{**}(t)$ satisfies $f(t) > 0$ for any $t > 0$.*

   (Proof by picture)

IV. Classification calibration results:

(a) Getting quantitative bounds on risk: define the ψ-transform via

$$\psi(\delta) := H^{**}(\delta). \tag{13.0.1}$$

(b) Main theorem for today:

Theorem 13.0.7. *Let ϕ be a margin-based loss function and ψ the associated ψ-transform. Then for any $f : \mathcal{X} \to \mathbb{R}$,*

$$\psi(R(f) - R^*) \le R_\phi(f) - R_\phi^*. \tag{13.0.2}$$

Moreover, the following three are equivalent:

1. *The loss ϕ is classification-calibrated*
2. *For any sequence $\delta_n \in [0,1]$,*

$$\psi(\delta_n) \to 0 \quad \Leftrightarrow \quad \delta_n \to 0.$$

3. *For any sequence of measurable functions $f_n : \mathcal{X} \to \mathbb{R}$,*

$$R_\phi(f_n) \to R_\phi^* \quad \textit{implies} \quad R(f_n) \to R^*.$$

1. Some insights from theorem. Recall examples 13.0.1 and 13.0.2. For both of these, we have that $\psi(\delta) = H(\delta)$, as H is convex. For the hinge loss, $\phi(s) = [1-s]_+$, we obtain for any f that

$$\mathbb{P}(Yf(X) \le 0) - \inf_f \mathbb{P}(Yf(X) \le 0) \le \mathbb{E}\left[[1 - Yf(X)]_+\right] - \inf_f \mathbb{E}\left[[1 - Yf(X)]_+\right].$$

   On the other hand, for the exponential loss, we have

$$\frac{1}{2}\left(\mathbb{P}(Yf(X) \le 0) - \inf_f \mathbb{P}(Yf(X) \le 0)\right)^2 \le \mathbb{E}\left[\exp(-Yf(X))\right] - \inf_f \mathbb{E}\left[\exp(-Yf(X))\right].$$

   The hinge loss is sharper.

2. **Example 13.0.8** (Regression for classification)**:** What about the surrogate loss $\frac{1}{2}(f(x) - y)^2$? In the homework, show which margin ϕ this corresponds to, and moreover, $H(\delta) = \frac{1}{2}\delta^2$. So regressing on the labels is consistent. $\diamond$

(c) Proof of Theorem 13.0.7 The proof of the theorem proceeds in several parts.

1. We first state a lemma, which follows from the results on convex functions we have already proved. The lemma is useful for several different parts of our proof.

   Lemma 13.0.9. *We have the following.*

   a. The functions H and ψ are continuous.

b. We have $H \geq 0$ and $H(0) = 0$.

c. If $H(\delta) > 0$ for all $\delta > 0$, then $\psi(\delta) > 0$ for all $\delta > 0$.

Because $H(0) = 0$ and $H \geq 0$: we have

$$\ell_\phi^{\rm wrong}(1/2) := \inf_{s(1-1)\leq 0} \ell_\phi(s, 1/2) = \inf_s \ell_\phi(s, 1/2) = \ell_\phi^*(1/2),$$

so $H(0) = \ell_\phi^*(1/2) - \ell_\phi^*(1/2) = 0$. (It is clear that the sub-optimality gap $H \geq 0$ by construction.)

2. We begin with the first statement of the theorem, inequality (13.0.2). Consider first the gap (for a fixed margin s) in conditional 0-1 risk,

$$\begin{aligned}\ell(s,\eta) - \inf_s \ell(s,\eta) &= \eta \mathbf{1}\left\{s \leq 0\right\} + (1-\eta)\mathbf{1}\left\{s \geq 0\right\} - \eta \mathbf{1}\left\{\eta \leq 1/2\right\} - (1-\eta)\mathbf{1}\left\{\eta \geq 1/2\right\} \\ &= \begin{cases} 0 & \text{if } \mathrm{sign}(s) = \mathrm{sign}(\eta - \frac{1}{2}) \\ \eta \vee (1-\eta) - \eta \wedge (1-\eta) = |2\eta - 1| & \text{if } \mathrm{sign}(s) \neq \mathrm{sign}(\eta - \frac{1}{2}). \end{cases}\end{aligned}$$

In particular, we obtain that the gap in risks is

$$R(f) - R^* = \mathbb{E}\left[\mathbf{1}\left\{\mathrm{sign}(f(X)) \neq \mathrm{sign}(2\eta(X) - 1)\right\} |2\eta(X) - 1|\right]. \tag{13.0.3}$$

Now we use expression (13.0.3) to get an upper bound on $R(f) - R^*$ via the ϕ-risk. Indeed, consider the ψ-transform (13.0.1). By Jensen's inequality, we have that

$$\psi(R(f) - R^*) \leq \mathbb{E}\left[\psi(\mathbf{1}\left\{\mathrm{sign}(f(X)) \neq \mathrm{sign}(2\eta(X) - 1)\right\} |2\eta(X) - 1|)\right].$$

Now we recall from Lemma 13.0.9 that $\psi(0) = 0$. Thus we have

$$\begin{aligned}\psi(R(f) - R^*) &\leq \mathbb{E}\left[\psi(\mathbf{1}\left\{\mathrm{sign}(f(X)) \neq \mathrm{sign}(2\eta(X) - 1)\right\} |2\eta(X) - 1|)\right] \\ &= \mathbb{E}\left[\mathbf{1}\left\{\mathrm{sign}(f(X)) \neq \mathrm{sign}(2\eta(X) - 1)\right\} \psi(|2\eta(X) - 1|)\right] \end{aligned} \tag{13.0.4}$$

Now we use the special structure of the suboptimality function we have constructed. Note that $\psi \leq H$, and moreover, we have for any $s \in \mathbb{R}$ that

$$\begin{aligned}\mathbf{1}\left\{\mathrm{sign}(s) \neq \mathrm{sign}(2\eta - 1)\right\} H(|2\eta - 1|) &= \mathbf{1}\left\{\mathrm{sign}(s) \neq \mathrm{sign}(2\eta - 1)\right\} \left[\inf_{s(2\eta-1)\leq 0} \ell_\phi(s,\eta) - \ell_\phi^*(\eta)\right] \\ &\leq \ell_\phi(s,\eta) - \ell_\phi^*(\eta), \end{aligned} \tag{13.0.5}$$

because $(1 + |2\eta - 1|)/2 = \max\{\eta, 1-\eta\}$.
Combining inequalities (13.0.4) and (13.0.5), we see that

$$\begin{aligned}\psi(R(f) - R^*) &\leq \mathbb{E}\left[\mathbf{1}\left\{\mathrm{sign}(f(X)) \neq \mathrm{sign}(2\eta(X) - 1)\right\} H(|2\eta(X) - 1|)\right] \\ &\leq \mathbb{E}\left[\ell_\phi(f(X), \eta(X)) - \ell_\phi^*(\eta(X))\right] \\ &= R_\phi(f) - R_\phi^*, \end{aligned}$$

which is our desired result.

3. Having proved the quantitative bound (13.0.2), we now turn to proving the second part of Theorem 13.0.7. Using Lemma 13.0.9, we can prove the equivalence of all three items.

We begin by showing that IV(b)1 implies IV(b)2. If ϕ is classification calibrated, we have $H(\delta) > 0$ for all $\delta > 0$. Because ψ is continuous and $\psi(0) = 0$, if $\delta \to 0$, then $\psi(\delta) \to 0$. It remains to show that $\psi(\delta) \to 0$ implies that $\delta \to 0$. But this is clear because we know that $\psi(0) = 0$ and$\psi(\delta) > 0$ whenever $\delta > 0$, and the convexity of ψ implies that ψ is increasing.
To obtain IV(b)3 from IV(b)2, note that by inequality (13.0.2), we have

$$\psi(R(f_n) - R^*) \le R_\phi(f_n) - R_\phi^* \to 0,$$

so we must have that $\delta_n = R(f_n) - R^* \to 0$.
Finally, we show that IV(b)1 follows from IV(b)3. Assume for the sake of contradiction that IV(b)3 holds but IV(b)1 fails, that is, ϕ is not classification calibrated. Then there must exist $\eta < 1/2$ and a sequence $s_n \ge 0$ (i.e. a sequence of predictions with incorrect sign) satisfying

$$\ell_\phi(s_n, \eta) \to \ell_\phi^*(\eta).$$

Construct the classification problem with a singleton $\mathcal{X} = \{x\}$, and set $\mathbb{P}(Y = 1) = \eta$. Then the sequence $f_n(x) = s_n$ satisfies $R_\phi(f_n) \to R_\phi^*$ but the true 0-1 risk $R(f_n) \not\to R^*$.

V. Classification calibration in the convex case

a. Suppose that ϕ is *convex*, which we often use for computational reasons

b.

Theorem 13.0.10 (Bartlett, Jordan, McAuliffe [19])**.** *If ϕ is convex, then ϕ is classification calibrated if and only if $\phi'(0)$ exists and $\phi'(0) < 0$.*

Proof First, suppose that ϕ is differentiable at 0 and $\phi'(0) < 0$. Then

$$\ell_\phi(s, \eta) = \eta\phi(s) + (1 - \eta)\phi(-s)$$

satisfies $\ell_\phi'(0, \eta) = (2\eta - 1)\phi'(0)$, and if $\phi'(0) < 0$, this quantity is negative for $\eta > 1/2$. Thus the minimizing $s(\eta) \in (0, \infty]$. (Proof by picture, but formalize in full notes.)

For the other direction assume that ϕ is classification calibrated. Recall the definition of a subgradient g_s of the function ϕ at $s \in \mathbb{R}$ is any g_s such that $\phi(t) \ge \phi(s) + g_s(t - s)$ for all $t \in \mathbb{R}$. (Picture.) Let g_1, g_2 be such that $\ell(s) \ge \ell(0) + g_1 s$ and $\ell(s) \ge \ell(0) + g_2 s$, which exist by convexity. We show that both $g_1, g_2 < 0$ and $g_1 = g_2$. By convexity we have

$$\begin{aligned} \ell_\phi(s, \eta) &\ge \eta(\phi(0) + g_1 s) + (1 - \eta)(\phi(0) - g_2 s) \\ &= [\eta g_1 - (1 - \eta)g_2]\, s + \phi(0). \end{aligned} \tag{13.0.6}$$

We first show that $g_1 = g_2$, meaning that ϕ is differentiable. Without loss of generality, assume $g_1 > g_2$. Then for $\eta > 1/2$, we would have $\eta g_1 - (1 - \eta)g_2 > 0$, which would imply that

$$\ell_\phi(s, \eta) \ge \phi(0) \ge \inf_{s' \le 0} \left\{\eta\phi(s') + (1 - \eta)\phi(-s')\right\} = \ell_\phi^{\text{wrong}}(\eta),$$

for all $s \ge 0$ by (13.0.6), by taking $s' = 0$ in the second inequality. By our assumption of classification calibration, for $\eta > 1/2$ we know that

$$\inf_s \ell_\phi(s, \eta) < \inf_{s \le 0} \ell_\phi(s, \eta) = \ell_\phi^{\text{wrong}}(\eta) \quad \text{so} \quad \ell_\phi^*(\eta) = \inf_{s \ge 0} \ell_\phi(s, \eta),$$

and under the assumption that $g_1 > g_2$ we obtain $\ell_\phi^*(\eta) = \inf_{s \ge 0} \ell_\phi(s, \eta) > \ell_\phi^{\text{wrong}}(\eta)$, which is a contradiction to classification calibration. We thus obtain $g_1 = g_2$, so that the function ϕ has a unique subderivative at $s = 0$ and is thus differentiable.

Now that we know ϕ is differentiable at 0, consider

$$\eta\phi(s) + (1 - \eta)\phi(-s) \ge (2\eta - 1)\phi'(0)s + \phi(0).$$

If $\phi'(0) \ge 0$, then for $s \ge 0$ and $\eta > 1/2$ we must have the right hand side is at least $\phi(0)$, which contradicts classification calibration, because we know that $\ell_\phi^*(\eta) < \ell_\phi^{\text{wrong}}(\eta)$ exactly as in the preceding argument. □

13.1 General results

JCD Comment: Here we should have some more general results on surrogate risk consistency.

I. Setting: we have a loss (risk) $L : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}_+$ and instead wish to minimize a surrogate $\varphi : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ for it

a. Say it's *Fisher consistent* (or infinite sample consistent) if $R_\varphi(f_n) \to R_\varphi^\star$ implies $R(f_n) \to R^\star$

b. Reduce to pointwise cases, compare non-uniform to uniform results (noting that in cases where L is discrete, they are the same—requires a proof)

c. Basically, this is Question 13.4, except we will use finite $\mathcal{Y}$ I think (can still leave the super general version in)

13.2 Proofs of convex analytic results

13.2.1 Proof of Lemma 13.0.4

First, let $(a, b) \subset \operatorname{dom} f$ and fix $x_0 \in (a, b)$. Let $x \uparrow x_0$, which is no loss of generality, and we may also assume $x \in (a, b)$. Then we have

$$x = sa + (1 - s)x_0 \quad \text{and} \quad x_0 = \beta b + (1 - \beta)x$$

for some $s, \beta \in [0, 1]$. Rearranging by convexity,

$$f(x) \le sf(a) + (1 - s)f(x_0) = f(x_0) + s(f(a) - f(x_0))$$

and

$$f(x_0) \le \beta f(b) + (1 - \beta)f(x), \quad \text{or} \quad \frac{1}{1 - \beta} f(x_0) \le f(x) + \frac{\beta}{1 - \beta} f(b).$$

Taking $s, \beta \to 0$, we obtain

$$\liminf_{x \to x_0} f(x) \ge f(x_0) \quad \text{and} \quad \limsup_{x \to x_0} f(x) \le f(x_0)$$

as desired.

13.2.2 Proof of Lemma 13.0.4

We need only consider the endpoints of the domain by Lemma 13.0.3, and we only need to show that $\limsup_{x \to x_0} f(x) \le f(x_0)$. But this is obvious by convexity: let $x = ty + (1-t)x_0$ for any $y \in \dom f$, and taking $t \to 0$, we have $f(x) \le tf(y) + (1-t)f(x_0) \to f(x_0)$.

13.2.3 Proof of Lemma 13.0.6

We begin with the case (i). Define the function $h_{\low}(t) := \inf_{s \ge t} h(s)$. Then because h is continuous, we know that over any compact set it attains its infimum, and thus (by assumption on h) $h_{\low}(t) > 0$ for all $t > 0$. Moreover, $h_{\low}$ is non-decreasing. Now define $f_{\low}(t) = h_{\low}^{**}(t)$ to be the biconjugate of $h_{\low}$; it is clear that $f \ge f_{\low}$ as $h \ge h_{\low}$. Thus we see that case (ii) implies case (i), so we turn to the more general result to see that $f_{\low}(t) > 0$ for all $t > 0$.

For the result in case (ii), assume for the sake of contradiction there is some $z \in (0,1)$ satisfying $h^{**}(z) = 0$. It is clear that $h^{**}(0) = 0$ and $h^{**} \ge 0$, so we must have $h^{**}(z/2) = 0$. Now, by assumption we have $h(z/2) = b > 0$, whence we have $h(1) \ge b > 0$. In particular, the piecewise linear function defined by

$$g(t) = \begin{cases} 0 & \text{if } t \le z/2 \\ \frac{b}{1-z/2}(t - z/2) & \text{if } t > z/2 \end{cases}$$

is closed, convex, and satisfies $g \le h$. But $g(z) > 0 = h^{**}(z)$, a contradiction to the fact that h^{**} is the largest (closed) convex function below h.

13.3 Exercises

Exercise 13.1: Find the suboptimality function H_ϕ and ψ-transform for the binary classification problem with the following losses.

(a) Logistic loss. That is,

$$\phi(s) = \log(1 + e^{-s})$$

(b) Squared error (ordinary regression). The surrogate loss in this case for the pair (x, y) is $\frac{1}{2}(f(x) - y)^2$. Show that for $y \in \{-1, 1\}$, this can be written as a margin-based loss, and compute the associated suboptimality function H_ϕ and ψ-transform. Is the squared error classification calibrated?

Exercise 13.2: Suppose we have a regression problem with data (independent variables) $x \in \mathcal{X}$ and $y \in \mathbb{R}$. We wish to find a predictor $f : \mathcal{X} \to \mathbb{R}$ minimizing the probability of being far away from the true y, that is, for some $c > 0$, our loss is of the form

$$L(f(x), y) = \mathbf{1}\left\{|y - f(x)| \ge c\right\}.$$

Show that no loss of the form $\varphi(s, y) = |s - y|^p$, where $p \ge 1$, is Fisher consistent for the loss L, even if the distribution of Y conditioned on $X = x$ is symmetric about its mean $\mathbb{E}[Y \mid X]$. That is, show there exists a distribution on pairs X, Y such that the set of minimizers of the surrogate

$$R_\varphi(f) := \mathbb{E}[\varphi(f(X), Y)]$$

is not included in the set of minimizers of the true risk, $R(f) = \mathbb{P}(|Y - f(X)| \ge c)$, even if the distribution of Y (conditional on X) is symmetric.

Exercise 13.3 (Empirics of classification calibration)**:** In this problem you will compare the performance of hinge loss minimization and an ordinary linear regression in terms of classification performance. Specifically, we compare the performance of the hinge surrogate loss with the regression surrogate when the data is generated according to the model

$$y = \sign(\langle \theta^*, x \rangle + \sigma Z), \quad Z \sim \mathsf{N}(0, 1) \tag{13.3.1}$$

where $\theta^* \in \mathbb{R}^d$ is a fixed vector, $\sigma \geq 0$ is an error magnitude, and Z is a standard normal random variable. We investigate the model (13.3.1) with a simulation study.

Specifically, we consider the following set of steps:

(i) Generate two collections of n datapoints in d dimensions according to the model (13.3.1), where $\theta \in \mathbb{R}^d$ is chosen (ahead of time) uniformly at random from the sphere $\{\theta \in \mathbb{R}^d : \|\theta\|_2 = R\}$, and where each $x_i \in \mathbb{R}^d$ is chosen as $\mathsf{N}(0, I_{d \times d})$. Let (x_i, y_i) denote pairs from the first collection and $(x_i^{\test}, y_i^{\test})$ pairs from the second.

(ii) Set

$$\widehat{\theta}_{\hinge} = \argmin_{\theta : \|\theta\|_2 \leq R} \frac{1}{n} \sum_{i=1}^n [1 - y_i \langle x_i, \theta \rangle]_+$$

and

$$\widehat{\theta}_{\reg} = \argmin_{\theta} \frac{1}{2n} \sum_{i=1}^n (y_i - \langle x_i, \theta \rangle)^2 = \argmin_{\theta} \|X\theta - y\|_2^2 .$$

(iii) Evaluate the 0-1 error rate of the vectors $\widehat{\theta}_{\hinge}$ and $\widehat{\theta}_{\reg}$ on the held-out data points $\{(x_i^{\test}, y_i^{\test})\}_{i=1}^n$.

Perform the preceding steps (i)–(iii), using any $n \geq 100$ and $d \geq 10$ and a radius $R = 5$, for different standard deviations $\sigma = \{0, 1, \ldots, 10\}$; perform the experiment a number of times. Give a plot or table exhibiting the performance of the classifiers learned on the held-out data. How do the two compare? Given that for the hinge loss we know $H_\phi(\delta) = \delta$ (as presented in class), what would you expect based on the answer to Question 13.1?

I have implemented (in the `julia` language; see `http://julialang.org/`) methods for solving the hinge loss minimization problem with stochastic gradient descent so that you do not need to. The file is available at this link. The code should (hopefully) be interpretable enough that if `julia` is not your language of choice, you can re-implement the method in an alternative language.

Exercise 13.4: In this question, we generalize our results on classification calibration and surrogate risk consistency to a much broader supervised learning setting. Consider the following general supervised learning problem, where we assume that we have data in pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are general spaces.

Let $L : \mathbb{R}^m \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function we wish to minimize, so that the loss of a prediction function $f : \mathcal{X} \to \mathbb{R}^m$ for the pair (x, y) is $L(f(x), y)$. Let $\varphi : \mathbb{R}^m \times \mathcal{Y} \to \mathbb{R}$ be an arbitrary surrogate, where $\varphi(f(x), y)$ is the surrogate loss. Define the risk and φ-risk

$$R(f) := \mathbb{E}[L(f(X), Y)] \quad \text{and} \quad R_\varphi(f) := \mathbb{E}[\varphi(f(X), Y)].$$

Let $\mathcal{P}_{\mathcal{Y}}$ denote the space of all probability distributions on $\mathcal{Y}$, and define the conditional (pointwise) risks $\ell : \mathbb{R}^m \times \mathcal{P}_{\mathcal{Y}} \to \mathbb{R}$ and $\ell_\varphi : \mathbb{R}^m \times \mathcal{P}_{\mathcal{Y}} \to \mathbb{R}$ by

$$\ell(s, P) = \int_{\mathcal{Y}} L(s, y) p(y) dy \quad \text{and} \quad \ell_\varphi(s, P) = \int_{\mathcal{Y}} \ell(s, y) p(y) dy.$$

(Here for simplicity we simply write integration against dy; you may make this fully general if you wish.) Let $\ell^*(P) = \inf_s \ell(s, P)$ denote the minimal conditional risk, and similarly for $\ell^*_\varphi(P)$, when Y has distribution P. If P_x denotes the distribution of Y conditioned on $X = x$, then we may rewrite the risk functionals as

$$R(f) = \mathbb{E}[\ell(f(X), P_X)] \quad \text{and} \quad R_\varphi(f) = \mathbb{E}[\ell_\varphi(f(X), P_X)].$$

We will show that the same machinery we developed for classification calibration extends to this general supervised learning setting.

For $\epsilon \geq 0$, define the suboptimality gap function

$$\Delta_\varphi(\epsilon, P) := \inf_{s \in \mathbb{R}^m} \left\{ \ell_\varphi(s, P) - \ell^*_\varphi(P) : \ell(s, P) - \ell^*(P) \geq \epsilon \right\}, \tag{13.3.2}$$

which measures the gap between achievable (pointwise) risk and the best surrogate risk when we enforce that the true loss is not minimized. Also define the uniform suboptimality function

$$\Delta_\varphi(\epsilon) := \inf_{s \in \mathbb{R}^m, P \in \mathcal{P}_\mathcal{Y}} \left\{ \ell_\varphi(s, P) - \ell^*_\varphi(P) : \ell(s, P) - \ell^*(P) \geq \epsilon \right\}.$$

(Compare this with the definition of Δ for the classification case to gain intuition.)

(a) A uniform result: let $\Delta^{**}_\varphi(\epsilon)$ be the biconjugate of Δ_φ (that is, Δ^{**}_φ is the largest convex function below Δ_φ). Show that

$$\Delta^{**}_\varphi(R(f) - R^*) \leq R_\varphi(f) - R^*_\varphi.$$

Prove that $\Delta_\varphi(\epsilon) > 0$ for all $\epsilon > 0$ implies if $R_\varphi(f_n) \to R^*_\varphi$, then $R(f_n) \to R^*$.

(b) We say that the loss φ is *uniformly calibrated* if $\Delta_\varphi(\epsilon) > 0$ for all $\epsilon > 0$. Show that, in the margin-based binary classification case with loss $\phi : \mathbb{R} \to \mathbb{R}$, uniform calibration as defined here is equivalent to classification-calibration as defined in class. You may assume that the margin-based loss ϕ is continuous.

(c) A non-uniform result: assume that for all distributions $P \in \mathcal{P}_\mathcal{Y}$ on the set $\mathcal{Y}$, we have

$$\Delta_\varphi(\epsilon, P) > 0$$

if $\epsilon > 0$. (We call this *calibration*.) Assume that there exists an upper bound function $B : \mathcal{X} \to \mathbb{R}_+$ such that $\mathbb{E}[B(X)] < \infty$ and $\ell(s, P_x) \leq \ell^*(P_x) + B(x)$ for all x and $s \in \mathbb{R}^m$. For example, if the loss L is bounded, this holds. Show that if the sequence of functions $f_n : \mathcal{X} \to \mathbb{R}^m$ satisfies

$$R_\varphi(f_n) \to R^*_\varphi \quad \text{then} \quad R(f_n) \to R^*.$$

Equivalently, show that for any distribution P on $\mathcal{X} \times \mathcal{Y}$, for all $\epsilon > 0$ there exists a $\delta > 0$ such that

$$R_\varphi(f) \leq R^*_\varphi + \delta \quad \text{implies} \quad R(f) \leq R^* + \epsilon.$$

(You may ignore any measurability issues that come up.)

Chapter 14

Divergences, classification, and risk

JCD Comment: There is so much to do in this section.

1. Change entropies to all be $H(Y)$ and $H(Y \mid X)$ or $H(Y \mid \mathsf{q}(X))$
2. For losses or risks, probably ℓ would be better and L for population loss (risk), but not sure
3. connect information to amount of entropy left, so there are alternative informations
4. Give proof of universal equivalence for the *binary* case, which is "easy" (at least, easier...) because we can just use binary entropies of the form $h(p) = \inf_\alpha\{p\ell(\alpha) + (1-p)\ell(-\alpha)\}$, choosing distributions in a fairly transparent way to get them. (Will probably write this down in afternoon.)

New outline:

I. Generalized entropies
   (a) Definitions as infima of losses
   (b) Gaps in prior and posterior risk become statistical information

II. From entropy to losses
   (a) Basically that each entropy gives rise to a loss
   (b) Generalized version of this: structured prediction problems (with an example)

III. Predictions with entropies and scoring rules
   (a) Some similarity to the ideas in the Fenchel-Young losses paper, where given a vector s of scores, we make predictions

$$\mathsf{pred}_\Omega(s) := \operatorname{argmax}\{\langle p, s\rangle - \Omega(p)\}$$

   (b) If loss is generalized entropy loss, then there is duality in that loss minimizers s give calibrated p when Ω is strictly (or perhaps uniformly?) convex

IV. Surrogate risk consistency with convex entropy-based losses

(a) Multiclass case: any time we have a uniformly convex loss, we get consistency (or uniformly convex entropy I guess)

(b) The discrete losses for structured prediction

V. Loss equivalence

(a)

I. Bayes risk in classification problems

a. Recall definition (2.2.3) of f-divergence between two distributions P and Q as

$$D_f\left(P\|Q\right) := \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx,$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex function satisfying $f(1) = 0$. If f is not linear, then $D_f\left(P\|Q\right) > 0$ unless $P = Q$.

b. Focusing on binary classification case, let us consider some example risks and see what connections they have to f-divergences. (Recall we have $X \in \mathcal{X}$ and $Y \in \{-1, 1\}$ we would like to classify.)

1. We require a few definitions to understand the performance of different classification strategies. In particular, we consider the difference between the risk possible when we see a point to classify and when we do not.
2. The prior risk is the risk attainable *without* seeing x, we have for a fixed sign $\alpha \in \mathbb{R}$ the definition

$$R_{\rm prior}(\alpha) := P(Y = 1)\mathbf{1}\left\{\alpha \le 0\right\} + P(Y = -1)\mathbf{1}\left\{\alpha \ge 0\right\}, \tag{14.0.1}$$

and similarly the minimal prior risk, defined as

$$R^*_{\rm prior} := \inf_\alpha \left\{P(Y = 1)\mathbf{1}\left\{\alpha \le 0\right\} + P(Y = -1)\mathbf{1}\left\{\alpha \ge 0\right\}\right\} = \min\{P(Y = 1), P(Y = -1)\}. \tag{14.0.2}$$

3. Also have the prior ϕ-risk, defined as

$$R_{\phi,\rm prior}(\alpha) := P(Y = 1)\phi(\alpha) + P(Y = -1)\phi(-\alpha), \tag{14.0.3}$$

and the minimal prior ϕ-risk, defined as

$$R^*_{\phi,\rm prior} := \inf_\alpha \left\{P(Y = 1)\phi(\alpha) + P(Y = -1)\phi(-\alpha)\right\}. \tag{14.0.4}$$

c. Examples of 0-1 loss and its friends: have $X \in \mathcal{X}$ and $Y \in \{-1, 1\}$.

1. **Example 14.0.1** (Binary classification with 0-1 loss)**:** What is Bayes risk of binary classifier? Let

$$p_{+1}(x) = p(x \mid Y = 1) = \frac{P(Y = 1 \mid X = x)p(x)}{P(Y = 1)}$$

be the density of X conditional on $Y = 1$ and similarly for $p_{-1}(x)$, and assume that each class occurs with probability $1/2$. Then

$$R^* = \inf_\gamma \int \left[\mathbf{1}\left\{\gamma(x) \le 0\right\} P(Y = 1 \mid X = x) + \mathbf{1}\left\{\gamma(x) \ge 0\right\} P(Y = -1 \mid X = x)\right] p(x)dx$$
$$= \frac{1}{2} \inf_\gamma \int \left[\mathbf{1}\left\{\gamma(x) \le 0\right\} p_{+1}(x) + \mathbf{1}\left\{\gamma(x) \ge 0\right\} p_{-1}(x)\right] dx = \frac{1}{2} \int \min\{p_{+1}(x), p_{-1}(x)\}dx.$$

Similarly, we may compute the minimal prior risk, which is simply $\frac{1}{2}$ by definition (14.0.2). Looking at the gap between the two, we obtain

$$R^*_{\rm prior} - R^* = \frac{1}{2} - \frac{1}{2} \int \min\{p_{+1}(x), p_{-1}(x)\}dx = \frac{1}{2} \int \left[p_1 - p_{-1}\right]_+ = \frac{1}{2} \left\|P_1 - P_{-1}\right\|_{\rm TV}.$$

That is, the difference is half the variation distance between P_1 and P_{-1}, the distributions of x conditional on the label Y. $\diamond$

2. **Example 14.0.2** (Binary classification with hinge loss)**:** We now repeat precisely the same calculations as in Example 14.0.1, but using as our loss the hinge loss (recall Example 13.0.2). In this case, the minimal ϕ-risk is

$$R^*_\phi = \int \inf_\alpha \left[\left[1 - \alpha\right]_+ P(Y = 1 \mid X = x) + \left[1 + \alpha\right]_+ P(Y = -1 \mid X = x)\right] p(x)dx$$
$$= \frac{1}{2} \int \inf_\alpha \left[\left[1 - \alpha\right]_+ p_1(x) + \left[1 + \alpha\right]_+ p_{-1}(x)\right] dx = \int \min\{p_1(x), p_{-1}(x)\}dx.$$

We can similarly compute the prior risk as $R^*_{\phi,\rm prior} = 1$. Now, when we calculate the improvement available via observing $X = x$, we find that

$$R^*_{\phi,\rm prior} - R^*_\phi = 1 - \int \min\{p_1(x), p_{-1}(x)\}dx = \left\|P_1 - P_{-1}\right\|_{\rm TV},$$

which is suggestively similar to Example 14.0.1. $\diamond$

d. Is there anything more we can say about this?

II. Statistical information, f-divergences, and classification problems

a. Statistical information

1. Suppose we have a classification problem with data $X \in \mathcal{X}$ and labels $Y \in \{-1, 1\}$. A natural notion of information that X carries about Y is the gap

$$R^*_{\rm prior} - R^*, \tag{14.0.5}$$

that between the prior risk and the risk attainable after viewing $x \in \mathcal{X}$.

2. **Didn't present this.** True definition of *statistical information*: suppose class 1 has prior probability π and class -1 has prior $1 - \pi$, and let P_1 and P_{-1} be the distributions of $X \in \mathcal{X}$ given $Y = 1$ and $Y = -1$, respectively. The *Bayes risk* associated with the problem is then

$$B_\pi(P_1, P_{-1}) := \inf_\gamma \int \left[\mathbf{1}\left\{\gamma(x) \le 0\right\} p_1(x)\pi + \mathbf{1}\left\{\gamma(x) \ge 0\right\} p_{-1}(x)(1 - \pi)\right] dx \tag{14.0.6}$$
$$= \int p_1(x)\pi \wedge p_{-1}(x)(1 - \pi)dx$$

and similarly, the prior Bayes risk is

$$B_\pi := \inf_\alpha \left\{ \mathbf{1}\left\{\alpha \le 0\right\} \pi + \mathbf{1}\left\{\alpha \ge 0\right\} (1-\pi) \right\} = \pi \wedge (1-\pi). \tag{14.0.7}$$

Then statistical information is

$$B_\pi - B_\pi(P_1, P_{-1}). \tag{14.0.8}$$

3. Measure proposed by DeGroot [60] in experimental design problem; goal is to infer state of world based on further experiments, want to measure quality of measurement.
4. Saw that for 0-1 loss, when *a-priori* each class was equally likely, then $R^*_{\text{prior}} - R^* = \frac{1}{2} \|P_1 - P_{-1}\|_{\text{TV}}$, and similarly for hinge loss (Example 14.0.2) that $R^*_{\phi,\text{prior}} - R^*_\phi = \|P_1 - P_{-1}\|_{\text{TV}}$.
5. Note that if $P_1 \neq P_{-1}$, then the statistical information is positive.

b. **Did present this.** More general story? Yes.

1. Consider any margin-based surrogate loss ϕ, and look at the difference between

$$\begin{aligned} B_{\phi,\pi}(P_1, P_{-1}) &:= \inf_\gamma \int \left[\phi(\gamma(x)) p_1(x) \pi + \phi(-\gamma(x)) p_{-1}(x)(1-\pi)\right] dx \\ &= \int \inf_\alpha \left[\phi(\alpha) p_1(x) \pi + \phi(-\alpha) p_{-1}(x)(1-\pi)\right] dx \end{aligned}$$

and the prior ϕ-risk, $B_{\phi,\pi}$.

2. Note that

$$B_{\phi,\pi} - B_{\phi,\pi}(P_1, P_{-1})$$

is simply gap in ϕ-risk $R^*_{\phi,\text{prior}} - R^*_\phi$ for distribution with $P(Y = 1) = \pi$ and

$$P(Y = y \mid X = x) = \frac{p(x \mid Y = y) P(Y = y)}{p(x)} = \frac{p_y(x) \pi^{\mathbf{1}\{y=1\}} (1-\pi)^{\mathbf{1}\{y=-1\}}}{\pi p_1(x) + (1-\pi) p_{-1}(x)}. \tag{14.0.9}$$

c. Have theorem (see, for example, Liese and Vajda [131], or Reid and Williamson [150]):

Theorem 14.0.3. *Let P_1 and P_{-1} be arbitrary distributions on $\mathcal{X}$, and let $\pi \in [0,1]$ be a prior probability of a class label. Then there is a convex function $f_{\pi,\phi} : \mathbb{R}_+ \to \mathbb{R}$ satisfying $f_{\pi,\phi}(1) = 0$ such that*

$$B_{\phi,\pi} - B_{\phi,\pi}(P_1, P_{-1}) = D_{f_{\pi,\phi}}\left(P_{-1} \| P_1\right).$$

Moreover, this function $f_{\pi,\phi}$ is

$$f_{\pi,\phi}(t) = \sup_\alpha \left[\ell^*_\phi(\pi) - \frac{\pi \phi(\alpha) t + (1-\pi)\phi(-\alpha)}{\pi t + (1-\pi)} \right] (t\pi + (1-\pi)). \tag{14.0.10}$$

Proof First, consider the integrated Bayes risk. Recalling the definition of the conditional distribution $\eta(x) = P(Y = 1 \mid X = x)$, we have

$$\begin{aligned} B_{\phi,\pi} - B_{\phi,\pi}(P_1, P_{-1}) &= \int \left[\ell^*_\phi(\pi) - \ell^*_\phi(\eta(x))\right] p(x) dx \\ &= \int \sup_\alpha \left[\ell^*_\phi(\pi) - \phi(\alpha) P(Y = 1 \mid x) - \phi(-\alpha) P(Y = -1 \mid x)\right] p(x) dx \\ &= \int \sup_\alpha \left[\ell^*_\phi(\pi) - \phi(\alpha) \frac{p_1(x)\pi}{p(x)} - \phi(-\alpha) \frac{p_{-1}(x)(1-\pi)}{p(x)}\right] p(x) dx, \end{aligned}$$

where we have used Bayes rule as in (14.0.9). Let us now divide all appearances of the density p_1 by p_{-1}, which yields

$$B_{\phi,\pi} - B_{\phi,\pi}(P_1, P_{-1}) = \int \sup_\alpha \left[\ell_\phi^*(\pi) - \frac{\phi(\alpha)\frac{p_1(x)}{p_{-1}(x)}\pi + \phi(-\alpha)(1-\pi)}{\frac{p_1(x)}{p_{-1}(x)}\pi + (1-\pi)}\right] \left(\frac{p_1(x)}{p_{-1}(x)}\pi + (1-\pi)\right) p_{-1}(x)dx. \tag{14.0.11}$$

By inspection, the representation (14.0.11) gives the result of the theorem if we can argue that the function f_π is convex, where we substitute $p_1(x)/p_{-1}(x)$ for t in $f_\pi(t)$.

To see that the function f_π is convex, consider the intermediate function

$$s_\pi(u) := \sup_\alpha \left\{-\pi\phi(\alpha)u - (1-\pi)\phi(-\alpha)\right\}.$$

This is the supremum of a family of linear functions in the variable u, so it is convex. Moreover, as we noted in the first exercise set, the perspective of a convex function g, defined by $h(u,t) = tg(u/t)$ for $t \geq 0$, is jointly convex in u and t. Thus, as

$$f_\pi(t) = \ell_\phi^*(\pi) + s_\pi\left(\frac{t}{\pi t + (1-\pi)}\right)(\pi t + (1-\pi)),$$

we have that f_π is convex. It is clear that $f_\pi(1) = 0$ by definition of $\ell_\phi^*(\pi)$. □

d. Take-home: any loss function induces an associated f-divergence. (There is a complete converse, in that any f-divergence can be realized as the difference in prior and posterior Bayes risk for some loss function; see, for example, Liese and Vajda [131] for results of this type.)

III. Quantization and other types of empirical minimization

a. Do these equivalences mean anything? What about the fact that the suboptimality function H_ϕ was linear for the hinge loss?

b. Consider problems with *quantization*: we must jointly learn a classifier (prediction or discriminant function) γ and a quantizer $\mathsf{q} : \mathcal{X} \to \{1, \ldots, k\}$, where k is fixed and we wish to find an optimal quantizer $\mathsf{q} \in \mathsf{Q}$, where Q is some family of quantizers. Recall the notation (2.2.1) of quantization of f-divergence, so

$$D_f\left(P_0 \| P_1 \mid \mathsf{q}\right) = \sum_{i=1}^k P_1(\mathsf{q}^{-1}(i)) f\left(\frac{P_0(\mathsf{q}^{-1}(i))}{P_1(\mathsf{q}^{-1}(i))}\right) = \sum_{i=1}^k P_1(A_i) f\left(\frac{P_0(A_i)}{P_1(A_i)}\right)$$

where the A_i are the quantization regions of $\mathcal{X}$.

c. Using Theorem 14.0.3, we can show how quantization and learning can be unified.

1. Quantized version of risk: for $\mathsf{q} : \mathcal{X} \to \{1, \ldots, k\}$ and $\gamma : [k] \to \mathbb{R}$,

$$R_\phi(\gamma \mid \mathsf{q}) = \mathbb{E}[\phi(Y\gamma(\mathsf{q}(X)))]$$

2. Rearranging and using integration,

$$R_\phi(\gamma \mid \mathsf{q}) = \mathbb{E}[\phi(Y\gamma(\mathsf{q}(X)))] = \sum_{z=1}^{k} \mathbb{E}[\phi(Y\gamma(z)) \mid \mathsf{q}(X) = z] P(\mathsf{q}(X) = z)$$
$$= \sum_{z=1}^{k} \left[\phi(\gamma(z)) P(Y = 1 \mid \mathsf{q}(X) = z) + \phi(-\gamma(z)) P(Y = -1 \mid \mathsf{q}(X) = z)\right] P(\mathsf{q}(X) = z)$$
$$= \sum_{z=1}^{k} \left[\phi(\gamma(z)) \frac{P(\mathsf{q}(X) = z \mid Y = 1) P(Y = 1)}{P(\mathsf{q}(X) = z)} + \phi(-\gamma(z)) \frac{P(\mathsf{q}(X) = z \mid Y = -1) P(Y = -1)}{P(\mathsf{q}(X) = z)}\right] P(\mathsf{q}(X)$$
$$= \sum_{z=1}^{k} \left[\phi(\gamma(z)) P_1(\mathsf{q}(X) = z)\pi + \phi(-\gamma(z)) P_{-1}(\mathsf{q}(X) = z)(1 - \pi)\right].$$

3. Let P^{q} denote the distribution with probability mass function

$$P^{\mathsf{q}}(z) = P(\mathsf{q}(X) = z) = P(\mathsf{q}^{-1}(\{z\})),$$

and define quantized Bayes ϕ-risk

$$R^*_\phi(\mathsf{q}) = \inf_\gamma R_\phi(\gamma \mid \mathsf{q})$$

Then for problem with $P(Y = 1) = \pi$, we have

$$R^*_{\phi,\text{prior}} - R^*_\phi(\mathsf{q}) = B_{\phi,\pi} - B_{\phi,\pi}(P_1^{\mathsf{q}}, P_{-1}^{\mathsf{q}}) = D_{f_{\pi,\phi}}\left(P_{-1} \| P_1 \mid \mathsf{q}\right). \tag{14.0.12}$$

d. Result unifying quantization and learning: we say that loss functions ϕ_1 and ϕ_2 are *universally equivalent* if they induce the same f divergence (14.0.10), that is, there is a constant $c > 0$ and $a, b \in \mathbb{R}$ such that

$$f_{\pi,\phi_1}(t) = c f_{\pi,\phi_2}(t) + at + b \text{ for all } t. \tag{14.0.13}$$

Theorem 14.0.4. *Let ϕ_1 and ϕ_2 be equivalent margin-based surrogate loss functions. Then for any quantizers q_1 and q_2,*

$$R^*_{\phi_1}(\mathsf{q}_1) \le R^*_{\phi_1}(\mathsf{q}_2) \text{ if and only if } R^*_{\phi_2}(\mathsf{q}_1) \le c R^*_{\phi_2}(\mathsf{q}_2).$$

Proof The proof follows straightforwardly via the representation (14.0.12). If ϕ_1 and ϕ_2 are equivalent, then we have that

$$R^*_{\phi_1,\text{prior}} - R^*_{\phi_1}(\mathsf{q}) = D_{f_{\pi,\phi_1}}\left(P_{-1} \| P_1 \mid \mathsf{q}\right) = c D_{f_{\pi,\phi_2}}\left(P_{-1} \| P_1 \mid \mathsf{q}\right) + a + b$$
$$= c\left[R^*_{\phi_2,\text{prior}} - R^*_{\phi_2}(\mathsf{q})\right] + a + b$$

for *any* quantizer q. In particular, we have

$$\begin{aligned} R^*_{\phi_1}(\mathsf{q}_1) \le R^*_{\phi_1}(\mathsf{q}_2) \text{ if and only if } & R^*_{\phi_1,\text{prior}} - R^*_{\phi_1}(\mathsf{q}_1) \ge R^*_{\phi_1,\text{prior}} - R^*_{\phi_1}(\mathsf{q}_2) \\ \text{if and only if } & D_{f_{\pi,\phi_1}}\left(P_{-1} \| P_1 \mid \mathsf{q}_1\right) \ge D_{f_{\pi,\phi_1}}\left(P_{-1} \| P_1 \mid \mathsf{q}_2\right) \\ \text{if and only if } & D_{f_{\pi,\phi_2}}\left(P_{-1} \| P_1 \mid \mathsf{q}_1\right) \ge D_{f_{\pi,\phi_2}}\left(P_{-1} \| P_1 \mid \mathsf{q}_2\right) \\ \text{if and only if } & R^*_{\phi_2,\text{prior}} - R^*_{\phi_2}(\mathsf{q}_1) \ge R^*_{\phi_2,\text{prior}} - R^*_{\phi_2}(\mathsf{q}_2). \end{aligned}$$

Subtracting $R^*_{\phi_2,\text{prior}}$ from both sides gives our desired result. □

e. Some comments:

1. We have an interesting thing: if we wish to learn a quantizer and a classifier jointly, then this is possible by using any loss equivalent to the true loss we care about.
2. Example: hinge loss and 0-1 loss are equivalent.
3. Turns out that the condition that the losses ϕ_1 and ϕ_2 be equivalent is (essentially) necessary and sufficient for two quantizers to induce the same ordering [144]. This equivalence is necessary and sufficient for the ordering conclusion of Theorem 14.0.4.

14.1 Generalized entropies

14.2 From entropy to losses

14.2.1 Classification case

JCD Comment: Notation: let P be the distribution on Y and let p the associated p.m.f., with $p_y = P(Y = y)$ for notational simplicity.

Say we have a (generalized) entropy $H : \Delta_k \to \mathbb{R}$, a concave function with $H(p) > -\infty$ except (potentially) when one of the $p_j = 0$. From any generalized entropy H, we can define a *convex* loss φ for which

$$H(p) = \inf_s \sum_{y=1}^k p_y L(s, y),$$

and this loss is

$$\varphi(s, y) = -s_y + (-H)^*(s).$$

14.2.2 Structured prediction case

In the structured prediction case, where we represent y by a statistic $\tau(y) \in \mathbb{R}^m$ so that

$$L(y', y) := \tau(y')^\top A \tau(y),$$

we can define the generalized entropy (with some abuse of notation)

$$H_L(P) := \min_y \mathbb{E}_P \left[\tau(y)^\top A \tau(Y) \right].$$

We define the *marginal polytope*

$$\mathcal{M} := \mathrm{Conv}(\{\tau(y)\}_{y \in \mathcal{Y}}) = \left\{ \sum_{y \in \mathcal{Y}} p_y \tau(y) \mid p \in \Delta_{\mathcal{Y}} \right\}$$

and if we define the mean mapping $\mu : \Delta_{\mathcal{Y}} \to \mathcal{M}$ by

$$\mu(P) := \mathbb{E}_P[\tau(Y)] = \sum_{y \in \mathcal{Y}} p_y \tau(y),$$

we see that

$$H_L(P) = \min_{y \in \mathcal{Y}} \tau(y)^\top A \mu(P) = \inf_{\nu \in \mathcal{M}} \nu^\top A \mu(P).$$

Notably, H_L is concave in p, as it is the infimum of linear functionals, and with some abuse of notation, we can also define the negative entropy mapping

$$\Omega(\mu) := -\min_{y \in \mathcal{Y}} \tau(y)^\top A\mu + \mathbf{I}_{\mathcal{M}}(\mu),$$

which evidently satisfies $\Omega(\nu) = H_L(P)$ for any $\nu \in \mathcal{M}$ satisfying $\nu = \mu(P)$ and is convex. The associated surrogate loss is

$$\varphi(s, y) := -s^\top \tau(y) + \Omega^*(s), \tag{14.2.1}$$

and we have

$$\mathbb{E}_p[\varphi(s, Y)] = -s^\top \mu(p) + \Omega^*(s),$$

so that

$$\inf_s \mathbb{E}_p[\varphi(s, Y)] = \inf_s \left\{-s^\top \mu(p) + \Omega^*(s)\right\} = -\sup_s \left\{s^\top \mu(p) - \Omega^*(s)\right\} = -\Omega(\mu(p)) = H_L(p).$$

14.3 Predictions, calibration, and scoring rules

14.4 Surrogate risk consistency

14.4.1 Uniformly convex case

14.4.2 Structured prediction (discrete) case

The amazing thing is that the construction (14.2.1) is surrogate-risk consistent under fairly weak conditions. We have the prediction function

$$\widehat{y}(s) \equiv \mathsf{pred}(s) := \operatorname*{argmax}_{y \in \mathcal{Y}} \left\{\tau(y)^\top s\right\},$$

where we choose the element arbitrarily if the maximizer is non-unique.

The first question is why this should be (surrogate-risk) consistent. To gain some intuition for this case, we present a few heuristic calculations that rely on convex analysis, before we move the the more sophisticated (and rigorous) argument to come. As always, we consider only pointwise versions of the risk—as surrogate risk consistency requires only this—and fix a $P \in \Delta_{\mathcal{Y}}$ and its induced $\mu = \mu(P)$. Consider the s minimizing the conditional surrogate risk

$$\ell_\varphi(s, P) = -s^\top \mu + \Omega^*(s).$$

As s minimizes ℓ, we have $\Omega(\mu) - s^\top \mu + \Omega^*(s) = 0$, and thus (using some duality results in the **appendices**) we necessarily have

$$s \in \partial\Omega(\mu).$$

As $\Omega(\mu) = \max_{y \in \mathcal{Y}} -\tau(y)^\top A\mu + \mathbf{I}_{\mathcal{M}}(\mu)$, we see that

$$\begin{aligned}\partial\Omega(\mu) &= \mathrm{Conv}\{-A^\top \tau(y) \mid \tau(y)^\top A\mu = \min_{y'} \tau(y')^\top A\mu\} + \mathcal{N}_{\mathcal{M}}(\mu) \\ &= \{-A^\top \nu \mid \nu^\top A\mu = H_L(P), \nu \in \mathcal{M}\} + \mathcal{N}_{\mathcal{M}}(\mu),\end{aligned} \tag{14.4.1}$$

where we recall that $\mu = \mu(P)$. If we make the (unrealistically) simplifying assumption that μ is interior to $\mathcal{M}$ (say, for example, if P assigns positive probability to all $y \in \mathcal{Y}$), i.e. $\mu \in \operatorname{int} \mathcal{M}$, then

$\mathcal{N}_{\mathcal{M}}(\mu) = \{\mathbf{0}\}$. If we also assume there is only a single label $y^\star \in \mathcal{Y}$ minimizing $\tau(y^\star)^\top A\mu$, that is, a single best prediction for the probabilities P on Y, then we obtain that $s = -A^\top \tau(y^\star)$. Then of course the prediction function becomes

$$\mathsf{pred}(s) = \operatorname*{argmax}_{y \in \mathcal{Y}} \left\{ -\tau(y)^\top A^\top \tau(y^\star) \right\} = y^\star$$

by the assumed identifiability condition on L.

We can actually make this fully rigorous under a few additional assumptions.

Assumption 14.1. *The loss L is symmetric, and if y minimizes $\mathbb{E}_P[L(y, Y)] = \mathbb{E}_P[\tau(y)^\top A\tau(Y)]$ then $P(Y = y) > 0$.*

We have the following theorem.

Theorem 14.4.1. *The surrogate φ is consistent for the discrete structured prediction loss L.*

14.4.3 Proof of Theorem 14.4.1

The proof proceeds similarly to the heuristic guarantee that $\mathsf{pred}(s)$ is correct in this case. Recall the gap functional (13.3.2),

$$\Delta_\varphi(\epsilon, P) := \inf_s \left\{ \ell_\varphi(s, P) - \ell_\varphi^\star(P) \mid \ell(s, P) - \ell^\star(P) \geq \epsilon \right\}.$$

In this case, we may simplify the quantities by writing out the entropy functionals explicitly as $\ell(s, P) = \tau(\widehat{y}(s))^\top A\mu - \inf_{\nu \in \mathcal{M}} \nu^\top A\mu$, where $\widehat{y}(s)$ is (an arbitrary) element of the prediction set $\operatorname{argmax}_y s^\top \tau(y)$. We need only show that $\Delta_\varphi(\epsilon, P) > 0$ whenever $\epsilon > 0$, which we prove by contradiction.

Thus, assume for the sake of contradiction that $\Delta_\varphi(\epsilon, P) = 0$. As the losses φ are piecewise linear and the set of s such that $\ell(s, P) - \ell^\star(P) \geq \epsilon$ is a union of polyhedra, there must be s achieving the infimum, and so for some vector of scores s, we have

$$\Omega^*(s) - s^\top \mu + \Omega(\mu) = 0$$

while $\widehat{y}(s)$ is incorrect. Following the calculation (14.4.1), we thus obtain that for some $\nu^\star \in \mathcal{M}$ satisfying $\langle \nu^\star, A\mu \rangle = \min_y \tau(y)^\top A\mu$ and a vector $w \in \mathcal{N}_{\mathcal{M}}(\mu)$, we have

$$s = -A^\top \nu^\star + w.$$

For any $\nu \in \mathcal{M}$, define the shorthand let $y^\star(\mu) = \operatorname{argmin}_y \tau(y)^\top A\mu$, which is a set-valued mapping, and let $y^\star(P) = y^\star(\mu(P))$ when there is no chance of notational confusion. If we can show the inclusions

$$\widehat{y}(s) \subset y^\star(\nu^\star) \subset y^\star(\mu(P)), \tag{14.4.2}$$

then the proof is complete, as we would evidently have our desired contradiction because necessarily $\tau(\widehat{y})^\top A\mu = \min_y \tau(y)^\top A\mu$ for any $\widehat{y} \in \widehat{y}(s)$.

To see the inclusion $y^\star(\nu^\star) \subset y^\star(\mu(P))$ is relatively straightforward. Let

$$\nu' \in \operatorname{Conv} \left\{ \tau(y) \mid \tau(y)^\top A\mu(P) = \min_{y'} \mathbb{E}_P[L(y', Y)] = H_L(P) \right\} \tag{14.4.3}$$

be otherwise arbitrary. For all P' such that $\nu' = \mu(P')$, the identifiability assumption 14.1 guarantees that if $y \in y^\star(\nu')$, we must have $P'(Y = y) > 0$. That is, we have

$$y^\star(\nu') \subset \cap_{P'} \left\{\operatorname{supp} P' \mid \nu' = \mu(P')\right\}.$$

In particular, Assumption 14.1 guarantees there is at least one P' satisfying $\operatorname{supp} P' \subset y^\star(\mu(P))$ and $\nu' = \mu(P')$, so that $y^\star(\nu') \subset y^\star(\mu(P))$ for all ν' in the convex hull (14.4.3), and in particular for $\nu^\star$.

The first inclusion in the chain (14.4.2) is more challenging. We begin a convex analytic result that allows us to simplify maximizers of $s^\top \tau(y)$ in y.

Lemma 14.4.2. *Let $w \in \mathcal{N}_{\mathcal{M}}(\mu)$ be the element satisfying $s = -A^T\nu^\star + w$. Then for any $y \in \mathcal{Y}$ and any $z \in \operatorname{supp} P$,*

$$\langle \tau(z) - \tau(y), w\rangle \geq 0.$$

Proof Fix any $y \in \mathcal{Y}$ and let $z \in \operatorname{supp} P$, so that $p_z > 0$. Then for a vector $\alpha \in \operatorname{Conv}(\tau(y') \mid y \notin \{y, z\})$, we can write $\mu(P) = \lambda_y\tau(y) + \lambda_z\tau(z) + (1 - \lambda_y - \lambda_z)\alpha$, where $\lambda_y \geq 0, \lambda_z \geq p_z > 0$, and $\lambda_y + \lambda_z \leq 1$. The vector $\nu = (\lambda_y + \lambda_z)\tau(y) + (1 - \lambda_y - \lambda_z)\alpha$ similarly satisfies $\nu \in \mathcal{M}$. By the definition of the normal cone $\mathcal{N}_{\mathcal{M}}(\mu)$, we know that $w^\top(\mu' - \mu) \leq 0$ for all $\mu' \in \mathcal{M}$, and in particular this holds for $\mu' = \nu$. As

$$\nu - \mu = \lambda_z(\tau(y) - \tau(z)),$$

we obtain

$$\lambda_z w^\top(\tau(y) - \tau(z)) \leq 0,$$

and as $\lambda_z > 0$ the lemma follows. □

With Lemma 14.4.2 in hand, we can consider the predictions $\mathsf{pred}(s) = \operatorname{argmax}_y s^\top\tau(y)$. As $s = -A^\top\nu^\star + w$, we have

$$\widehat{y}(s) = \operatorname*{argmax}_{y\in\mathcal{Y}} \left\{-\tau(y)^\top A^\top\nu^\star + \tau(y)^\top w\right\} = \operatorname*{argmax}_{y\in\mathcal{Y}} \left\{-\tau(y)^\top A\nu^\star + \tau(y)^\top w\right\},$$

where we have used the assumed symmetry of A. Let $y \in y^\star(\nu^\star)$ and $y' \notin y^\star(\nu^\star)$, so that $\tau(y')^\top A\nu^\star > \tau(y)^\top A\nu^\star$. Then by our earlier argument that $P(Y = y) > 0$, we obtain from Lemma 14.4.2 that

$$\tau(y')^\top w \leq \tau(y)^\top w.$$

We then see that

$$-\tau(y')^\top A\nu^\star + \tau(y')^\top w < -\tau(y)^\top A\nu^\star + \tau(y)^\top w,$$

and so $y' \notin \widehat{y}(s)$. In particular, $\widehat{y}(s) \subset y^\star(\nu^\star)$ as desired.

14.5 Loss equivalence

Definition 14.1. *The* generalized entropy *associated with a vector of losses $(\ell_y)_{y=1}^k$, $\ell_y : \mathbb{R}^k \to \mathbb{R}_+$ is*

$$H_\ell(Y) := \inf_s \left\{\sum_{y=1}^k P(Y = y)\ell_y(s)\right\}.$$

The associated conditional entropy of Y given $X = x$ *is* $H_\ell(Y \mid X = x) = \inf_s \sum_{y=1}^k P(Y = y \mid X = x)\ell_y(s)$, *and the* conditional entropy of Y given X *is*

$$H_\ell(Y \mid X) := \int_{\mathcal{X}} H_\ell(Y \mid X = x) dP(x).$$

From the definition, it is immediate that such entropies obey similar properties to the typical (Shannon) entropy $H(Y) = -\sum_y p(y) \log p(y)$ of a discrete random variable. Indeed, they are non-negative, and

JCD Comment: Fill this out in more detail, giving examples, motivation, etc. Some examples that would be worth doing (maybe earlier): multiclass with sums of hinge losses, and also the multiclass with order statistics (both from my papers). Can leave giving "gap" calculations as exercises.

We can associate a natural information measure to such entropies: the loss-based information that X carries about a target Y is

$$I_\ell(X; Y) := H_\ell(Y) - H_\ell(Y \mid X),$$

which is clearly nonnegative.

In cases such as margin-based binary classification, when it is more natural to think of prediction functions as mapping into $\mathbb{R}$, it is more convenient to work with a slight modification of these entropies, where for a margin-based loss ϕ and $Y \in \{\pm 1\}$, we define

$$H_\phi(Y) := \inf_{s \in \mathbb{R}} \left\{P(Y = 1)\phi(-s) + P(Y = -1)\phi(s)\right\},$$

so that H_ϕ is really a concave function on $p \in [0, 1]$ with $H_\phi(Y) = h_\phi(P(Y = 1))$, where the binary generalized entropy is

$$h_\phi(p) := \inf_{s \in \mathbb{R}} \left\{p\phi(-s) + (1 - p)\phi(s)\right\}.$$

Definition 14.2. *Losses ℓ_1 and ℓ_2 are* universally equivalent *if for all distributions on (X, Y) and all quantizers q_1 and q_2,*

$$I_{\ell_1}(\mathsf{q}_1(X); Y) \le I_{\ell_1}(\mathsf{q}_2(X); Y) \quad \text{if and only if} \quad I_{\ell_2}(\mathsf{q}_1(X); Y) \le I_{\ell_2}(\mathsf{q}_2(X); Y).$$

We note in passing that swapping the roles of q_1 and q_2 and taking contrapositives, we an equivalent formulation to Definition 14.2 is that

$$I_{\ell_1}(\mathsf{q}_1(X); Y) < I_{\ell_1}(\mathsf{q}_2(X); Y) \quad \text{if and only if} \quad I_{\ell_2}(\mathsf{q}_1(X); Y) < I_{\ell_2}(\mathsf{q}_2(X); Y).$$

Theorem 14.5.1. *Let the multiclass losses ℓ_1 and ℓ_2 be bounded below and H_1 and H_2 be the associated generalized entropies. Then ℓ_1 and ℓ_2 are universally equivalent if and only if there exist $a > 0$, $b \in \mathbb{R}^k$, and $c \in \mathbb{R}$ such that for all distributions on $Y \in [k]$,*

$$H_1(Y) = aH_2(Y) + b^\top p + c, \tag{14.5.1}$$

where $p = [P(Y = y)]_{y=1}^k$ is the p.m.f. of Y.

JCD Comment: Do the easy version here.

14.6 Proof of Theorem 14.5.1

We give the proof of the "hard" direction of Theorem 14.5.1: that is, that if ℓ_1 and ℓ_2 are universally equivalent, then their associated entropies are necessary linearly related (14.5.1). We prove the result in the slightly simpler case of binary classification, as the more general result introduces few new ideas except that it requires more technical care.

We thus work with margin-based losses $\phi_i : \mathbb{R} \to \mathbb{R}_+$, $i = 1, 2$, where without any loss of generality we assume $\inf_s \ell_i(s) = 0$ (as we may subtract a constant), and we have generalized (binary) entropies

$$h_i(p) := \inf_s \left\{p\phi_i(-s) + (1-p)\phi(s)\right\}.$$

By inspection, each h_i is a closed concave function (as it is the infimum of linear functions of p), and by symmetry they satisfy $h_i(0) = h_i(1) = 0$ and $h_i(\frac{1}{2}) = \sup_{p\in[0,1]} h_i(p)$. We show that these entropies satisfy a particular *order equivalence* property on $[0, 1]$, which will turn out to be sufficient to prove their equality.

To motivate what follows, recall that universal equivalence (Def. 14.2) must hold for *all* distributions on (X, Y), and hence all (measurable) spaces $\mathcal{X}$ and joint distributions on $\mathcal{X} \times \{\pm 1\}$. Thus, consider a space $\mathcal{X}$ that we can partition into sets $\{A, A^c\}$ or $\{B, B^c\}$, where we take the conditional distributions

$$Y \mid X \in A = \begin{cases} 1 & \text{w.p. } p_a \\ -1 & \text{w.p. } 1 - p_a, \end{cases} \quad Y \mid X \in A^c = \begin{cases} 1 & \text{w.p. } q_a \\ -1 & \text{w.p. } 1 - q_a, \end{cases}$$

and

$$Y \mid X \in B = \begin{cases} 1 & \text{w.p. } p_b \\ -1 & \text{w.p. } 1 - p_b, \end{cases} \quad Y \mid X \in B^c = \begin{cases} 1 & \text{w.p. } q_b \\ -1 & \text{w.p. } 1 - q_b, \end{cases}$$

where we require the consistency conditions that the marginals over Y remain constant, that is, if $P(A) = P(X \in A)$, we have

$$P(A)p_a + P(A^c)q_a = P(Y = 1) = P(B)p_b + P(B^c)q_b.$$

Then evidently by defining quantizers q_1 and q_2 such that $\mathsf{q}_1(x) = \mathbf{1}\{x \in A\}$ and $\mathsf{q}_2(x) = \mathbf{1}\{x \in B\}$, we have

$$\begin{aligned} I_{\phi_1}(\mathsf{q}_1(X); Y) &= h_1(P(Y = 1)) - P(A)h_1(p_a) - (1 - P(A))h_1(q_a), \\ I_{\phi_1}(\mathsf{q}_2(X); Y) &= h_1(P(Y = 1)) - P(B)h_1(p_b) - (1 - P(B))h_1(q_b), \end{aligned}$$

and similarly for I_{ϕ_2}. Then universal equivalence implies that

$$\begin{aligned} &P(A)h_1(p_a) + (1 - P(A))h_1(q_a) \le P(B)h_1(p_b) + (1 - P(B))h_1(q_b) \text{ if and only if} \\ &P(A)h_2(p_a) + (1 - P(A))h_2(q_a) \le P(B)h_2(p_b) + (1 - P(B))h_2(q_b) \end{aligned}$$

whenever the consistency condition $P(A)p_a + (1 - P(A))q_a = P(B)p_b + (1 - P(B))q_b$ holds. As we may choose $\mathcal{X}$ and the probabilities, we can take $P(A) = P(B) = \frac{1}{2}$ (so that their are two equiprobable partitions), and the preceding conditions become

$$h_1(p_a) + h_1(q_a) \le h_1(p_b) + h_1(q_b) \text{ if and only if } h_2(p_a) + h_2(q_a) \le h_2(p_b) + h_2(q_b)$$

whenever $p_a + q_a = p_b + q_b$.

Generalizing this construction by taking distributions over $\mathcal{X}$ that partition it into k equiprobable sets $\{A_1, \ldots, A_k\}$ or $\{B_1, \ldots, B_k\}$, each with $P(A_i) = P(B_i) = 1/k$, we see that universal equivalence implies that for any vectors $p \in [0,1]^k$ and $q \in [0,1]^k$ satisfies $\mathbf{1}^\top p = \mathbf{1}^\top q$,

$$\sum_{i=1}^k h_1(p_i) \le \sum_{i=1}^k h_1(q_i) \quad \text{if and only if} \quad \sum_{i=1}^k h_2(p_i) \le \sum_{i=1}^k h_2(q_i). \tag{14.6.1}$$

We shall give condition (14.6.1) a name, as it implies certain equivalence properties for convex functions (we can replace h_i with $-h_i$ and obtain convex functions).

Definition 14.3. *Let $\Omega \subset \mathbb{R}$ be a closed interval and let $f_1, f_2 : \Omega \to \mathbb{R}$ be closed convex functions. Then f_1 and f_2 are* order equivalent *if for any $k \in \mathbb{N}$ and vectors $s \in \Omega^k$ and $t \in \Omega^k$ satisfying $\mathbf{1}^\top s = \mathbf{1}^\top t$, we have*

$$\sum_{i=1}^k f_1(s_i) \le \sum_{i=1}^k f_1(t_i) \quad \textit{if and only if} \quad \sum_{i=1}^k f_2(s_i) \le \sum_{i=1}^k f_2(t_i)$$

As in the brief remark following Definition 14.2, by taking complements we have as well that

$$\sum_{i=1}^k f_1(s_i) < \sum_{i=1}^k f_1(t_i) \quad \text{if and only if} \quad \sum_{i=1}^k f_2(s_i) < \sum_{i=1}^k f_2(t_i)$$

The theorem will then be proved if we can show the following lemma.

Lemma 14.6.1. *Let f_1 and f_2 be order equivalent on Ω. Then there exist $a > 0$, and $b, c \in \mathbb{R}$ such that $f_1(t) = a f_2(t) + bt + c$ for all $t \in \Omega$.*

The proof of Lemma 14.6.1 is somewhat involved, and we proceed in three parts. The key is that order equivalence actually implies a strong relationship between *affine* combinations of points in the domain of the functions f_i, not just convex combinations of points, which guarantees that we can predict values of $f_2(v)$ for any $v \in \Omega$ by just three values of f_i evaluate in Ω. We state this as a lemma, whose proof we defer temporarily to Sec. 14.6.1

Lemma 14.6.2. *If f_1 and f_2 are order equivalent on Ω, then for any $\lambda \in \mathbb{R}^k$ satisfying $\lambda^\top \mathbf{1} = 1$ and any $u \in \Omega^k$, if $\lambda^\top u = v \in \Omega$ then*

$$\sum_{i=1}^k \lambda_i f_1(u_i) \le f_1(v) \quad \textit{if and only if} \quad \sum_{i=1}^k \lambda_i f_2(u_i) \le f_2(v),$$

and the statement still holds with both inequalities replaced with strict inequalities.

In particular, if

$$\sum_{i=1}^k \lambda_i f_1(u_i) = f_1(v) \;\; \text{then necessarily} \;\; \sum_{i=1}^k \lambda_i f_2(u_i) = f_2(v) \tag{14.6.2}$$

whenever $\lambda \in \mathbb{R}^k$ satisfies $\lambda^\top \mathbf{1} = 1$ and $u^\top \lambda = \sum_{i=1}^k \lambda_i u_i = v$.

Second, we recognize that we may assume both f_1 and f_2 are nonlinear in the lemma; otherwise, it is immediate. Nonlinearity guarantees that

Lemma 14.6.3. *Let f be convex on $\mathbb{R}$. Let $u_0 < u_1$ and for $\lambda \in [0,1]$, define $u_\lambda = (1-\lambda)u_0 + \lambda u_1$. If there exists any $\lambda \in (0,1)$ such that $f(u_\lambda) = \lambda f(u_0) + (1-\lambda) f(u_1)$, then f is linear on $[u_0, u_1]$.*

We leave the proof (an algebraic manipulation using the definitions of convexity) as Question 14.2.

The last intermediate step we require in the proof of Lemma 14.6.1 is that at three particular points in the domain Ω, we can satisfy Lemma 14.6.1.

Lemma 14.6.4. *Let f_1, f_2 be order equivalent on $\Omega = [u_0, u_1]$ and $u_c = \frac{1}{2}(u_0 + u_1)$. There are $a > 0$ and $b, c \in \mathbb{R}$ such that $f_1(t) = a f_2(t) + bt + c$ for $t \in \{u_0, u_c, u_1\}$.*

We can now finalize the proof of Lemma 14.6.1:

Proof Without loss of generality by an affine rescaling, we can assume that $f_1(t) = f_2(t)$ for $t \in \{u_0, u_c, u_1\}$, and our goal will be to show that $f_1(t) = f_2(t)$ for all $t \in \Omega$.

Let $v \in \Omega$ with $v \notin \{u_0, u_c, u_1\}$, and $u = [u_0 \; u_c \; u_1]^\top$ for shorthand. We seek $\lambda = (\lambda_0, \lambda_c, \lambda_1) \in \mathbb{R}^3$, where $\lambda^\top \mathbf{1} = 1$, such that both $\lambda^\top u = v$ and $\lambda_0 f_1(u_0) + \lambda_c f_1(u_c) + \lambda_1 f_1(u_1) = f_1(v)$. If we can find such a λ, then equality (14.6.2) guarantees that $f_1(v) = f_2(v)$, and we are done. As the points $(u_i, f_1(u_i))_{i=1}^3$ are not collinear (recall Lemma 14.6.3), the matrix

$$A := \begin{bmatrix} 1 & 1 & 1 \\ u_0 & u_c & u_1 \\ f_1(u_0) & f_1(u_c) & f_1(u_1) \end{bmatrix}$$

is full rank. In particular, there is a vector λ solving

$$A\lambda = [1 \; v \; f_1(v)]^\top, \quad \text{i.e.} \quad \lambda = A^{-1} [1 \; v \; f_1(v)]^\top,$$

which evidently satisfies our desiderata. Thus $f_1(v) = f_2(v)$, and as v was arbitrary, the proof is complete. □

14.6.1 Proof of Lemma 14.6.2

We prove the result first for λ with rational entries, as a continuity argument will give the rest. For each i, let $\alpha_i = [\lambda_i]_+$ and $\beta_i = [-\lambda_i]_+$ be the positive and negative parts of λ, so that $\lambda = \alpha - \beta$. Let $k \in \mathbb{N}$ be such that we can write $\alpha_i = \frac{a_i}{k}$ and $\beta_i = \frac{b_i}{k}$, where $a_i, b_i \in \mathbb{N}$. Then we have

$$\alpha^\top u = v + \beta^\top u \quad \text{or} \quad a^\top u = kv + b^\top u,$$

where $\mathbf{1}^\top a = k + \mathbf{1}^\top b$, as $\mathbf{1}^\top \lambda = \frac{1}{k} \mathbf{1}^\top (a - b) = 1$. Then we may define the two vectors

$$s = [\underbrace{u_1 \; \cdots \; u_1}_{a_1 \text{ times}} \; \cdots \; \underbrace{u_m \; \cdots \; u_m}_{a_m \text{ times}}]^\top \quad \text{and} \quad t = [\underbrace{v \; \cdots \; v}_{k \text{ times}} \; \underbrace{u_1 \; \cdots \; u_1}_{b_1 \text{ times}} \; \cdots \; \underbrace{u_m \; \cdots \; u_m}_{b_m \text{ times}}]^\top.$$

Then each has entries in Ω, and we have $\mathbf{1}^\top t = \mathbf{1}^\top s$. Then order equivalence (Def. 14.3) guarantees that

$$\sum_{i=1}^m a_i f_1(u_i) \le k f_1(v) + \sum_{i=1}^m b_i f_1(u_i) \quad \text{if and only if} \quad \sum_{i=1}^m a_i f_2(u_i) \le k f_2(v) + \sum_{i=1}^m b_i f_2(u_i)$$

and (as per the remark following the definition)

$$\sum_{i=1}^m a_i f_1(u_i) = k f_1(v) + \sum_{i=1}^m b_i f_1(u_i) \text{ if and only if } \sum_{i=1}^m a_i f_2(u_i) = k f_2(v) + \sum_{i=1}^m b_i f_2(u_i).$$

These two displays are equivalent to $\sum_{i=1}^m \lambda_i f_j(u_i) \le f_j(v)$ and $\sum_{i=1}^m \lambda_i f_j(u_i) = f_j(v)$, respectively, for $j = 1, 2$.

We have therefore proved Lemma 14.6.2 for λ taking rational values. Because closed convex functions on $\mathbb{R}$ are continuous on their domains (Lemma 13.0.4) the result extends to real-valued λ.

14.6.2 Proof of Lemma 14.6.4

If either of f_1 or f_2 is linear, the other is as well, and the proof becomes trivial, so we assume w.l.o.g. they are both nonlinear.

Without loss of generality, we take $u_0 = 0$, $u_1 = 1$, and $u_c = \frac{1}{2}$ by scaling. Then we must solve the three equations

$$f_1(0) = a f_2(0) + c, \quad f_1\left(\frac{1}{2}\right) = a f_2\left(\frac{1}{2}\right) + \frac{b}{2} + c, \quad f_1(1) = a f_2(1) + b + c.$$

From the first we obtain $c = f_1(0) - a f_2(0)$, and substituting this into the third yields $b = f_1(1) - f_1(0) - a(f_2(1) - f_2(0))$. Finally, substituting both equalities into the equality with $f_1(\frac{1}{2})$ yields that

$$\begin{aligned} f_1\left(\frac{1}{2}\right) &= a\left[f_2\left(\frac{1}{2}\right) - \frac{f_2(1) - f_2(0)}{2}\right] + \frac{f_1(1) - f_1(0)}{2} + f_1(0) - a f_2(0) \\ &= a\left[f_2(\frac{1}{2}) - \frac{f_2(1) + f_2(0)}{2}\right] + \frac{f_1(1) + f_1(0)}{2}. \end{aligned}$$

As we know that f_1, f_2 are nonlinear, Lemma 14.6.3 applies, so that the convexity gaps $f_1(\frac{1}{2}) - \frac{f_1(1)+f_1(0)}{2}$ and $f_2(\frac{1}{2}) - \frac{f_2(1)+f_2(0)}{2}$ are both positive, and thus we take

$$a = \frac{f_1(\frac{1}{2}) - \frac{f_1(0)+f_1(1)}{2}}{f_2(\frac{1}{2}) - \frac{f_2(0)+f_2(1)}{2}} > 0.$$

14.7 Bibliography

Point to full proof of Theorem 14.5.1.

14.8 Exercises

Exercise 14.1 (Bayes risk gaps)**:** Consider a general binary classification problem with $(X, Y) \in \mathcal{X} \times \{-1, 1\}$. Let $\phi(\alpha) = \log(1 + e^{-\alpha})$, so that we use the logistic loss. Show that the surrogate risk gap

$$L^*_{\phi,\text{prior}} - L^*_\phi = I(X; Y),$$

where I is the mutual information.

Exercise 14.2: Prove Lemma 14.6.3. *Hint:* without loss of generality, you may take $u_0 = 0$, $u_1 = 1$. Then for any $u \in [0, 1]$, write λ as a convex combination of either $\{0, u\}$ or $\{u, 1\}$ and use the definition of convexity.

Chapter 15

Fisher Information

Having explored the definitions associated with exponential families and their robustness properties, we now turn to a study of somewhat more general parameterized distributions, developing connections between divergence measures and other geometric ideas such as the Fisher information. After this, we illustrate a few consequences of Fisher information for optimal estimators, which gives a small taste of the deep connections between information geometry, Fisher information, exponential family models. In the coming chapters, we show how Fisher information measures come to play a central role in sequential (universal) prediction problems.

15.1 Fisher information: definitions and examples

We begin by defining the Fisher information. Let $\{P_\theta\}_{\theta \in \Theta}$ denote a parametric family of distributions on a space $\mathcal{X}$, each where $\theta \in \Theta \subset \mathbb{R}^d$ indexes the distribution. Throughout this lecture and the next, we assume (with no real loss of generality) that each P_θ has a density given by p_θ. Then the *Fisher information* associated with the model is the matrix given by

$$I_\theta := \mathbb{E}_\theta \left[\nabla_\theta \log p_\theta(X) \nabla \log p_\theta(X)^\top \right] = \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^\top], \tag{15.1.1}$$

where the score function $\dot{\ell}_\theta = \nabla_\theta \log p_\theta(x)$ is the gradient of the log likelihood at θ (implicitly depending on X) and the expectation $\mathbb{E}_\theta$ denotes expectation taken with respect to P_θ. Intuitively, the Fisher information captures the variability of the gradient $\nabla \log p_\theta$; in a family of distributions for which the score function $\dot{\ell}_\theta$ has high variability, we intuitively expect estimation of the parameter θ to be easier—different θ change the behavior of $\dot{\ell}_\theta$—though the log-likelihood functional $\theta \mapsto \mathbb{E}_{\theta_0}[\log p_\theta(X)]$ varies more in θ.

Under suitable smoothness conditions on the densities p_θ (roughly, that derivatives pass through expectations; see Remark 15.1 at the end of this chapter), there are a variety of alternate definitions of Fisher information. These smoothness conditions hold for exponential families, so at least in the exponential family case, everything in this chapter is rigorous. (We note in passing that there are more general definitions of Fisher information for more general families under quadratic mean differentiability; see, for example, van der Vaart [169].) First, we note that the score function has

mean zero under P_θ: we have

$$\begin{aligned}\mathbb{E}_\theta[\dot\ell_\theta] &= \int p_\theta(x)\nabla_\theta \log p_\theta(x)dx = \int \frac{\nabla p_\theta(x)}{p_\theta(x)} p_\theta(x)dx \\ &= \int \nabla p_\theta(x)dx \stackrel{(\star)}{=} \nabla \int p_\theta(x)dx = \nabla 1 = 0,\end{aligned}$$

where in equality $(\star)$ we have assumed that integration and derivation may be exchanged. Under similar conditions, we thus attain an alternate definition of Fisher information as the negative expected hessian of $\log p_\theta(X)$. Indeed,

$$\nabla^2 \log p_\theta(x) = \frac{\nabla^2 p_\theta(x)}{p_\theta(x)} - \frac{\nabla p_\theta(x) \nabla p_\theta(x)^\top}{p_\theta(x)^2} = \frac{\nabla^2 p_\theta(x)}{p_\theta(x)} - \dot\ell_\theta \dot\ell_\theta^\top,$$

so we have that the Fisher information is equal to

$$\begin{aligned}I_\theta = \mathbb{E}_\theta[\dot\ell_\theta \dot\ell_\theta^\top] &= -\int p_\theta(x) \nabla^2 \log p_\theta(x) dx + \int \nabla^2 p_\theta(x) dx \\ &= -\mathbb{E}[\nabla^2 \log p_\theta(x)] + \nabla^2 \underbrace{\int p_\theta(x)dx}_{=1} = -\mathbb{E}[\nabla^2 \log p_\theta(x)]. \end{aligned} \tag{15.1.2}$$

Summarizing, we have that

$$I_\theta = \mathbb{E}_\theta[\dot\ell_\theta \dot\ell_\theta] = -\mathbb{E}_\theta[\nabla^2 \log p_\theta(X)].$$

This representation also makes clear the additional fact that, if we have n i.i.d. observations from the model P_θ, then the information content similarly grows linearly, as $\log p_\theta(X_1^n) = \sum_{i=1}^n \log p_\theta(X_i)$.

We now give two examples of Fisher information, the first somewhat abstract and the second more concrete.

Example 15.1.1 (Canonical exponential family)**:** In a canonical exponential family model, we have $\log p_\theta(x) = \langle \theta, \phi(x)\rangle - A(\theta)$, where ϕ is the sufficient statistic and A is the log-partition function. Because $\dot\ell_\theta = \phi(x) - \nabla A(\theta)$ and $\nabla^2 \log p_\theta(x) = -\nabla^2 A(\theta)$ is a constant, we obtain

$$I_\theta = \nabla^2 A(\theta).$$

$\diamond$

Example 15.1.2 (Two parameterizations of a Bernoulli)**:** In the canonical parameterization of a Bernoulli as an exponential family model (Example 3.1.1), we had $p_\theta(x) = \exp(\theta x - \log(1 + e^\theta))$ for $x \in \{0, 1\}$, so by the preceding example the associated Fisher information is $\frac{e^\theta}{1+e^\theta}\frac{1}{1+e^\theta}$. If we make the change of variables $p = P_\theta(X = 1) = e^\theta/(1 + e^\theta)$, or $\theta = \log \frac{p}{1-p}$, we have $I_\theta = p(1 - p)$. On the other hand, if $P(X = x) = p^x(1 - p)^{1-x}$ for $p \in [0, 1]$, the standard formulation of the Bernoulli, then $\nabla \log P(X = x) = \frac{x}{p} - \frac{1-x}{1-p}$, so that

$$I_p = \mathbb{E}_p\left[\left(\frac{X}{p} - \frac{1 - X}{1 - p}\right)^2\right] = \frac{1}{p} + \frac{1}{1 - p} = \frac{1}{p(1 - p)}.$$

That is, the parameterization can change the Fisher information. $\diamond$

15.2 Estimation and Fisher information: elementary considerations

The Fisher information has intimate connections to estimation, both in terms of classical estimation and the information games that we discuss subsequently. As a motivating calculation, we consider estimation of the mean of a $\mathsf{Bernoulli}(p)$ random variable, where $p \in [0, 1]$, from a sample $X_1^n \stackrel{\text{iid}}{\sim} \mathsf{Bernoulli}(p)$. The sample mean $\widehat{p}$ satisfies

$$\mathbb{E}[(\widehat{p} - p)^2] = \frac{1}{n}\mathrm{Var}(X) = \frac{p(1-p)}{n} = \frac{1}{I_p} \cdot \frac{1}{n},$$

where I_p is the Fisher information for the single observation $\mathsf{Bernoulli}(p)$ family as in Example 15.1.2. In fact, this inverse dependence on Fisher information is unavoidable, as made clear by the Cramér Rao Bound, which provides lower bounds on the mean squared error of all unbiased estimators.

Proposition 15.2.1 (Cramér Rao Bound)**.** *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary differentiable function and assume that the random function (estimator) T is unbiased for $\phi(\theta)$ under P_θ. Then*

$$\mathrm{Var}(T) \geq \nabla\phi(\theta)^\top I_\theta^{-1} \nabla\phi(\theta).$$

As an immediate corollary to Proposition 15.2.1, we may take $\phi(\theta) = \langle \lambda, \theta \rangle$ for $\lambda \in \mathbb{R}^d$. Then varying λ over all of $\mathbb{R}^d$, and we obtain that for any unbiased estimator T for the parameter $\theta \in \mathbb{R}^d$, we have $\mathrm{Var}(\langle \lambda, T \rangle) \geq \lambda^\top I_\theta^{-1} \lambda$. That is, we have

Corollary 15.2.2. *Let T be unbiased for the parameter θ under the distribution P_θ. Then the covariance of T has lower bound*

$$\mathrm{Cov}(T) \succeq I_\theta^{-1}.$$

In fact, the Cramér-Rao bound and Corollary 15.2.2 hold, in an asymptotic sense, for substantially more general settings (without the unbiasedness requirement). For example, see the books of van der Vaart [169] or Le Cam and Yang [128, Chapters 6 & 7], which show that under appropriate conditions (known variously as quadratic mean differentiability and local asymptotic normality) that no estimator can have smaller mean squared error than Fisher information in any uniform sense.

We now prove the proposition, where, as usual, we assume that it is possible to exchange differentiation and integration.

Proof Throughout this proof, all expectations and variances are computed with respect to P_θ. The idea of the proof is to choose $\lambda \in \mathbb{R}^d$ to minimize the variance

$$\mathrm{Var}(T - \langle \lambda, \dot{\ell}_\theta \rangle) \geq 0,$$

then use this λ to provide a lower bound on $\mathrm{Var}(T)$.

To that end, let $\dot{\ell}_{\theta,j} = \frac{\partial}{\partial \theta_j} \log p_\theta(X)$ denote the jth component of the score vector. Because $\mathbb{E}_\theta[\dot{\ell}_\theta] = 0$, we have the covariance equality

$$\begin{aligned}\mathrm{Cov}(T - \phi(\theta), \dot{\ell}_{\theta,j}) &= \mathbb{E}[(T - \phi(\theta))\dot{\ell}_{\theta,j}] = \mathbb{E}[T\dot{\ell}_{\theta,j}] = \int T(x) \frac{\frac{\partial}{\partial \theta_j} p_\theta(x)}{p_\theta(x)} p_\theta(x) dx \\ &= \frac{\partial}{\partial \theta_j} \int T(x) p_\theta(x) dx = \frac{\partial}{\partial \theta_j} \phi(\theta),\end{aligned}$$

where in the final step we used that T is unbiased for $\phi(\theta)$. Using the preceding equality,

$$\mathrm{Var}(T - \langle \lambda, \dot{\ell}_\theta \rangle) = \mathrm{Var}(T) + \lambda^\top I_\theta \lambda - 2\mathbb{E}[(T - \phi(\theta))\langle \lambda, \dot{\ell}_\theta \rangle] = \mathrm{Var}(T) + \lambda^\top I_\theta \lambda - 2\langle \lambda, \nabla \phi(\theta) \rangle.$$

Taking $\lambda = I_\theta^{-1} \nabla \phi(\theta)$ gives $0 \le \mathrm{Var}(T - \langle \lambda, \dot{\ell}_\theta \rangle) = \mathrm{Var}(T) - \nabla \phi(\theta)^\top I_\theta^{-1} \nabla \phi(\theta)$, and rearranging gives the result. □

15.3 Connections between Fisher information and divergence measures

By making connections between Fisher information and certain divergence measures, such as KL-divergence and mutual (Shannon) information, we gain additional insights into the structure of distributions, as well as optimal estimation and encoding procedures. As a consequence of the asymptotic expansions we make here, we see that estimation of 1-dimensional parameters is governed (essentially) by moduli of continuity of the loss function with respect to the metric induced by Fisher information; in short, Fisher information is an unavoidable quantity in estimation. We motivate our subsequent development with the following example.

Example 15.3.1 (Divergences in exponential families)**:** Consider the exponential family density $p_\theta(x) = h(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta))$. Then a straightforward calculation implies that for any θ_1 and θ_2, the KL-divergence between distributions P_{θ_1} and P_{θ_2} is

$$D_{\rm kl}\left(P_{\theta_1} \| P_{\theta_2}\right) = A(\theta_2) - A(\theta_1) - \langle \nabla A(\theta_1), \theta_2 - \theta_1 \rangle.$$

That is, the divergence is simply the difference between $A(\theta_2)$ and its first order expansion around θ_1. This suggests that we may approximate the KL-divergence via the quadratic remainder in the first order expansion. Indeed, as A is infinitely differentiable (it is an exponential family model), the Taylor expansion becomes

$$\begin{aligned} D_{\rm kl}\left(P_{\theta_1} \| P_{\theta_2}\right) &= \frac{1}{2} \langle \theta_1 - \theta_2, \nabla^2 A(\theta_1)(\theta_1 - \theta_2) \rangle + O(\|\theta_1 - \theta_2\|^3) \\ &= \frac{1}{2} \langle \theta_1 - \theta_2, I_{\theta_1}(\theta_1 - \theta_2) \rangle + O(\|\theta_1 - \theta_2\|^3). \end{aligned}$$

◇

In particular, KL-divergence is roughly quadratic for exponential family models, where the quadratic form is given by the Fisher information matrix. We also remark in passing that for a convex function f, the Bregman divergence (associated with f) between points x and y is given by $D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$; such divergences are common in convex analysis, optimization, and differential geometry. Making such connections deeper and more rigorous is the goal of the field of information geometry (see the book of Amari and Nagaoka [5] for more).

We can generalize this example substantially under appropriate smoothness conditions. Indeed, we have

Proposition 15.3.2. *For appropriately smooth families of distributions* $\{P_\theta\}_{\theta \in \Theta}$*,*

$$D_{\rm kl}\left(P_{\theta_1} \| P_{\theta_2}\right) = \frac{1}{2} \langle \theta_1 - \theta_2, I_{\theta_1}(\theta_1 - \theta_2) \rangle + o(\|\theta_1 - \theta_2\|^2). \tag{15.3.1}$$

We only sketch the proof, as making it fully rigorous requires measure-theoretic arguments and Lebesgue's dominated convergence theorem.

Sketch of Proof By a Taylor expansion of the log density $\log p_{\theta_2}(x)$ about θ_1, we have

$$\begin{aligned}\log p_{\theta_2}(x) = \log p_{\theta_1}(x) + \langle \nabla \log p_{\theta_1}(x), \theta_1 - \theta_2 \rangle \\ + \frac{1}{2}(\theta_1 - \theta_2)^\top \nabla^2 \log p_{\theta_1}(x)(\theta_1 - \theta_2) + R(\theta_1, \theta_2, x),\end{aligned}$$

where $R(\theta_1, \theta_2, x) = O_x(\|\theta_1 - \theta_2\|^3)$ is the remainder term, where O_x denotes a hidden dependence on x. Taking expectations and assuming that we can interchange differentiation and expectation appropriately, we have

$$\begin{aligned}\mathbb{E}_{\theta_1}[\log p_{\theta_2}(X)] &= \mathbb{E}_{\theta_1}[\log p_{\theta_1}(X)] + \langle \mathbb{E}_{\theta_1}[\dot\ell_{\theta_1}], \theta_1 - \theta_2 \rangle \\ &\quad + \frac{1}{2}(\theta_1 - \theta_2)^\top \mathbb{E}_{\theta_1}[\nabla^2 \log p_{\theta_1}(X)](\theta_1 - \theta_2) + \mathbb{E}_{\theta_1}[R(\theta_1, \theta_2, X)] \\ &= \mathbb{E}_{\theta_1}[\log p_{\theta_1}(X)] - \frac{1}{2}(\theta_1 - \theta_2)^\top I_{\theta_1}(\theta_1 - \theta_2) + o(\|\theta_1 - \theta_2\|^2),\end{aligned}$$

where we have assumed that the $O(\|\theta_1 - \theta_2\|^3)$ remainder is uniform enough in X that $\mathbb{E}[R] = o(\|\theta_1 - \theta_2\|^2)$ and used that the score function $\dot\ell_\theta$ is mean zero under P_θ. □

We may use Proposition 15.3.2 to give a somewhat more general version of the Cramér-Rao bound (Proposition 15.2.1) that applies to more general (sufficiently smooth) estimation problems. Indeed, we will show that Le Cam's method (recall Chapter 8.3) is (roughly) performing a type of discrete second-order approximation to the KL-divergence, then using this to provide lower bounds. More concretely, suppose we are attempting to estimate a parameter θ parameterizing the family $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$, and assume that $\Theta \subset \mathbb{R}^d$ and $\theta_0 \in \operatorname{int} \Theta$. Consider the minimax rate of estimation of θ_0 in a neighborhood around θ_0; that is, consider

$$\inf_{\widehat\theta} \sup_{\theta = \theta_0 + v \in \Theta} \mathbb{E}_\theta[\|\widehat\theta(X_1^n) - \theta\|^2],$$

where the observations X_i are drawn i.i.d. P_θ. Fixing $v \in \mathbb{R}^d$ and setting $\theta = \theta_0 + \delta v$ for some $\delta > 0$, Le Cam's method (8.3.3) then implies that

$$\inf_{\widehat\theta} \max_{\theta \in \{\theta_0, \theta + \delta v\}} \mathbb{E}_\theta[\|\widehat\theta(X_1^n) - \theta\|^2] \ge \frac{\delta^2 \|v\|^2}{8}\left[1 - \left\|P_{\theta_0}^n - P_{\theta_0 + \delta v}^n\right\|_{\mathrm{TV}}\right].$$

Using Pinsker's inequality that $2\|P - Q\|_{\mathrm{TV}}^2 \le D_{\mathrm{kl}}(P\|Q)$ and the asymptotic quadratic approximation (15.3.1), we have

$$\left\|P_{\theta_0}^n - P_{\theta_0 + \delta v}^n\right\|_{\mathrm{TV}} \le \sqrt{\frac{n}{2} D_{\mathrm{kl}}(P_{\theta_0}\|P_{\theta_0 + \delta v})} = \frac{\sqrt{n}}{2}\left(\delta^2 v^\top I_{\theta_0} v + o(\delta^2\|v\|^2)\right)^{\frac{1}{2}}.$$

By taking $\delta^2 = (nv^\top I_{\theta_0} v)^{-1}$, for large enough v and n we know that $\theta_0 + \delta v \in \operatorname{int}\Theta$ (so that the distribution $P_{\theta_0 + \delta v}$ exists), and for large n, the remainder term $o(\delta^2\|v\|^2)$ becomes negligible. Thus we obtain

$$\inf_{\widehat\theta} \max_{\theta \in \{\theta_0, \theta + \delta v\}} \mathbb{E}_\theta[\|\widehat\theta(X_1^n) - \theta\|^2] \gtrsim \frac{\delta^2\|v\|^2}{16} = \frac{1}{16}\frac{\|v\|^2}{nv^\top I_{\theta_0} v}. \tag{15.3.2}$$

In particular, in one-dimension, inequality (15.3.2) implies a result generalizing the Cramér-Rao bound. We have the following asymptotic local minimax result:

Corollary 15.3.3. *Let $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$, where $\Theta \subset \mathbb{R}$, be a family of distributions satisfying the quadratic approximation condition of Proposition 15.3.2. Then there exists a constant $c > 0$ such that*

$$\lim_{v \to \infty} \lim_{n \to \infty} \inf_{\widehat{\theta}_n} \sup_{\theta : |\theta - \theta_0| \le v/\sqrt{n}} \mathbb{E}_\theta \left[(\widehat{\theta}_n(X_1^n) - \theta)^2 \right] \ge c \frac{1}{n} I_{\theta_0}^{-1}.$$

Written differently (and with minor extension), Corollary 15.3.3 gives a lower bound based on a local modulus of continuity of the loss function with respect to the metric induced by the Fisher information. Indeed, suppose we wish to estimate a parameter θ in the neighborhood of θ_0 (where the neighborhood size decreases as $1/\sqrt{n}$) according to some loss function $\ell : \Theta \times \Theta \to \mathbb{R}$. Then if we define the modulus of continuity of ℓ with respect to the Fisher information metric as

$$\omega_\ell(\delta, \theta_0) := \sup_{v : \|v\| \le 1} \frac{\ell(\theta_0, \theta_0 + \delta v)}{\delta^2 v^\top I_{\theta_0} v},$$

the combination of Corollary 15.3.3 and inequality (15.3.2) shows that the local minimax rate of estimating $\mathbb{E}_\theta[\ell(\widehat{\theta}_n, \theta)]$ for θ near θ_0 must be at least $\omega_\ell(n^{-1/2}, \theta_0)$. For more on connections between moduli of continuity and estimation, see, for example, Donoho and Liu [64].

Remark 15.1: In order to make all of our exchanges of differentiation and expectation rigorous, we must have some conditions on the densities we consider. One simple condition sufficient to make this work is via Lebesgue's dominated convergence theorem. Let $f : \mathcal{X} \times \Theta \to \mathbb{R}$ be a differentiable function. For a fixed base measure ν assume there exists a function g such that $g(x) \ge \|\nabla_\theta f(x, \theta)\|$ for all θ, where

$$\int_{\mathcal{X}} g(x) d\mu(x) < \infty.$$

Then in this case, we have $\nabla_\theta \int f(x, \theta) d\mu(x) = \int \nabla_\theta f(x, \theta) d\mu(x)$ by the mean-value theorem and definition of a derivative. (Note that for all θ_0 we have $\sup_{v : \|v\|_2 \le \delta} \|\nabla_\theta f(x, \theta)\|_2 \big|_{\theta = \theta_0 + v} \le g(x)$.) More generally, this type of argument can handle absolutely continuous functions, which are differentiable almost everywhere. $\diamond$

Part IV

Online game playing and compression

Chapter 16

Universal prediction and coding

In this chapter, we explore sequential game playing and online probabilistic prediction schemes. These have applications in coding when the true distribution of the data is unknown, biological algorithms (encoding genomic data, for example), control, and a variety of other areas. The field of universal prediction is broad; in addition to this chapter touching briefly on a few of the techniques therein and their relationships with statistical modeling and inference procedures, relevant reading includes the survey by Merhav and Feder [140], the more recent book of Grünwald [95], and Tsachy Weissman's EE376c course at Stanford.

JCD Comment: Check out the below stuff a bit more carefully

16.1 Basics of minimax game playing with log loss

The final set of problems we consider in which exponential families make a natural appearance are in so-called minimax games under the log loss. In particular, we consider the following general formulation of a two-player minimax game. First, we choose a distribution Q on a set $\mathcal{X}$ (with density q). Then nature (or our adversary) chooses a distribution $P \in \mathcal{P}$ on the set $\mathcal{X}$, where $\mathcal{P}$ is a collection of distributions on $\mathcal{X}$, so we suffer loss

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[-\log q(X)] = \sup_{P \in \mathcal{P}} \int p(x) \log \frac{1}{q(x)} dx. \tag{16.1.1}$$

In particular, we would like to solve the minimax problem

$$\operatorname*{minimize}_Q \sup_{P \in \mathcal{P}} \mathbb{E}[-\log q(X)].$$

To motivate this abstract setting we give two examples, the first abstract and the second somewhat more concrete.

Example 16.1.1: Suppose that receive n random variables $X_i \stackrel{\text{iid}}{\sim} P$; in this case, we have the sequential prediction loss

$$\mathbb{E}_P[-\log q(X_1^n)] = \sum_{i=1}^n \mathbb{E}_P \left[\log \frac{1}{q(X_i \mid X_1^{i-1})} \right],$$

which corresponds to predicting X_i given X_1^{i-1} as well as possible, when the X_i follow an (unknown or adversarially chosen) distribution P. $\diamond$

Example 16.1.2 (Coding)**:** Expanding on the preceding example, suppose that the set $\mathcal{X}$ is finite, and we wish to encode $\mathcal{X}$ into $\{0,1\}$-valued sequences using as few bits as possible. In this case, the Kraft inequality (recall Theorem 2.4.2) tells us that if $C : \mathcal{X} \to \{0,1\}^*$ is an uniquely decodable code, and $\ell_C(x)$ denotes the length of the encoding for the symbol $x \in \mathcal{X}$, then

$$\sum_x 2^{-\ell_C(x)} \le 1.$$

Conversely, given any length function $\ell : \mathcal{X} \to \mathbb{N}$ satisfying $\sum_x 2^{-\ell(x)} \le 1$, there exists an instantaneous (prefix) code C with the given length function. Thus, if we define the p.m.f. $q_C(x) = 2^{-\ell_C(x)} / \sum_x 2^{-\ell_C(x)}$, we have

$$-\log_2 q_C(x_1^n) = \sum_{i=1}^n \left[\ell_C(x_i) + \log \sum_x 2^{-\ell_C(x)}\right] \le \sum_{i=1}^n \ell_C(x_i).$$

In particular, we have a coding game where we attempt to choose a distribution Q (or sequential coding scheme C) that has as small an expected length as possible, uniformly over distributions P. (The field of universal coding studies such questions in depth; see Tsachy Weissman's course EE376b.) $\diamond$

We now show how the minimax game (16.1.1) naturally gives rise to exponential family models, so that exponential family distributions are so-called robust Bayes procedures (cf. Grünwald and Dawid [96]). Specifically, we say that Q is a robust Bayes procedure for the class $\mathcal{P}$ of distributions if it minimizes the supremum risk (16.1.1) taken over the family $\mathcal{P}$; that is, it is uniformly good for all distributions $P \in \mathcal{P}$. If we restrict our class $\mathcal{P}$ to be a linearly constrained family of distributions, then we see that the exponential family distributions are natural robust Bayes procedures: they uniquely solve the minimax game. More concretely, assume that $\mathcal{P} = \mathcal{P}_\alpha^{\rm lin}$ and that P_θ denotes the exponential family distribution with density $p_\theta(x) = p(x)\exp(\langle\theta, \phi(x)\rangle - A(\theta))$, where p denotes the base density. We have the following.

Proposition 16.1.3. *If $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$, then*

$$\inf_Q \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\log q(X)] = \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\log p_\theta(X)] = \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \inf_Q \mathbb{E}_P[-\log q(X)].$$

Proof This is a standard saddle-point argument (cf. [153, 104, 35]). First, note that

$$\begin{aligned}\sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\log p_\theta(X)] &= \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\langle\phi(X), \theta\rangle + A(\theta)] \\ &= -\langle\alpha, \theta\rangle + A(\theta) = \mathbb{E}_{P_\theta}[-\langle\theta, \phi(X)\rangle + A(\theta)] = H(P_\theta),\end{aligned}$$

where H denotes the Shannon entropy, for any distribution $P \in \mathcal{P}_\alpha^{\rm lin}$. Moreover, for any $Q \neq P_\theta$, we have

$$\sup_P \mathbb{E}_P[-\log q(X)] \ge \mathbb{E}_{P_\theta}[-\log q(X)] > \mathbb{E}_{P_\theta}[-\log p_\theta(X)] = H(P_\theta),$$

where the inequality follows because $D_{\rm kl}\left(P_\theta \| Q\right) = \int p_\theta(x) \log \frac{p_\theta(x)}{q(x)} dx > 0$. This shows the first equality in the proposition.

For the second equality, note that

$$\inf_Q \mathbb{E}_P[-\log q(X)] = \underbrace{\inf_Q \mathbb{E}_P\left[\log \frac{p(X)}{q(X)}\right]}_{=0} - \mathbb{E}_P[\log p(x)] = H(P).$$

But we know from our standard maximum entropy results (Theorem 11.4.7) that P_θ maximizes the entropy over $\mathcal{P}_\alpha^{\rm lin}$, that is, $\sup_{P \in \mathcal{P}_\alpha^{\rm lin}} H(P) = H(P_\theta)$. □

In short: maximum entropy is equivalent to robust prediction procedures for linear families of distributions $\mathcal{P}_\alpha^{\rm lin}$, which is equivalent to maximum likelihood in exponential families, which in turn is equivalent to I-projection.

JCD Comment: Here we are back to the original stuff

16.2 Universal and sequential prediction

We begin by defining the universal prediction (and universal coding) problems. In this setting, we assume we are playing a game in which given a sequence X_1^n of data, we would like to predict the data (which, as we saw in Example 16.1.2, is the same as encoding the data) as as if we *knew* the true distribution of the data. Or, in more general settings, we would like to predict the data as well as all predictive distributions P from some family of distributions $\mathcal{P}$, even if *a priori* we know little about the coming sequence of data.

We consider two versions of this game: the probabilistic version and the adversarial version. We shall see that they have similarities, but there are also a few important distinctions between the two. For both of the following definitions of sequential prediction games, we assume that p and q are densities or probability mass functions in the case that $\mathcal{X}$ is continuous or discrete (this is no real loss of generality) for distributions P and Q.

We begin with the adversarial case. Given a sequence $x_1^n \in \mathcal{X}^n$, the *regret* of the distribution Q for the sequence x_1^n with respect to the distribution P is

$$\mathsf{Reg}(Q, P, x_1^n) := \log \frac{1}{q(x_1^n)} - \log \frac{1}{p(x_1^n)} = \sum_{i=1}^n \log \frac{1}{q(x_i \mid x_1^{i-1})} - \log \frac{1}{p(x_i \mid x_1^{i-1})}, \tag{16.2.1}$$

where we have written it as the sum over $q(x_i \mid x_1^{i-1})$ to emphasize the sequential nature of the game. Associated with the regret of the sequence x_1^n is the *adversarial regret* (usually simply called the regret) of Q with respect to the family $\mathcal{P}$ of distributions, which is

$$\mathfrak{R}_n^{\mathcal{X}}(Q, \mathcal{P}) := \sup_{P \in \mathcal{P}, x_1^n \in \mathcal{X}^n} \mathsf{Reg}(Q, P, x_1^n). \tag{16.2.2}$$

In more generality, we may which to use a loss function ℓ different than the log loss; that is, we might wish to measure a loss-based version the regret as

$$\sum_{i=1}^n \ell(x_i, Q(\cdot \mid x_1^{i-1})) - \ell(x_i, P(\cdot \mid x_1^{i-1})),$$

where $\ell(x_i, P)$ indicates the loss suffered on the point x_i when the distribution P over X_i is played, and $P(\cdot \mid x_1^{i-1})$ denotes the conditional distribution of X_i given x_1^{i-1} according to P. We defer discussion of such extensions later, focusing on the log loss for now because of its natural connections with maximum likelihood and coding.

A less adversarial problem is to minimize the *redundancy*, which is the expected regret under a distribution P. In this case, we define the redunancy of Q with respect to P as the expected regret of Q with respect to P under the distribution P, that is,

$$\mathsf{Red}_n(Q, P) := \mathbb{E}_P \left[\log \frac{1}{q(X_1^n)} - \log \frac{1}{p(X_1^n)} \right] = D_{\rm kl} \left(P \| Q \right), \tag{16.2.3}$$

where the dependence on n is implicit in the KL-divergence. The worst-case redundancy with respect to a class $\mathcal{P}$ is then

$$\mathfrak{R}_n(Q, \mathcal{P}) := \sup_{P \in \mathcal{P}} \mathsf{Red}_n(Q, P). \tag{16.2.4}$$

We now give two examples to illustrate the redundancy.

Example 16.2.1 (Example 16.1.2 on coding, continued)**:** We noted in Example 16.1.2 that for any p.m.f.s p and q on the set $\mathcal{X}$, it is possible to define coding schemes C_p and C_q with code lengths

$$\ell_{C_p}(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil \quad \text{and} \quad \ell_{C_q}(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil.$$

Conversely, given (uniquely decodable) encoding schemes C_p and $C_q : \mathcal{X} \to \{0,1\}^*$, the functions $p_{C_p}(x) = 2^{-\ell_{C_p}(x)}$ and $q_{C_q}(x) = 2^{-\ell_{C_q}(x)}$ satisfy $\sum_x p_{C_p}(x) \le 1$ and $\sum_x q_{C_q}(x) \le 1$. Thus, the redundancy of Q with respect to P is the additional number of bits required to encode variables distributed according to P when we assume they have distribution Q:

$$\begin{aligned} \mathsf{Red}_n(Q, P) &= \sum_{i=1}^n \mathbb{E}_P \left[\log \frac{1}{q(X_i \mid X_1^{i-1})} - \log \frac{1}{p(X_i \mid X_1^{i-1})} \right] \\ &= \sum_{i=1}^n \mathbb{E}_P[\ell_{C_q}(X_i)] - \mathbb{E}_P[\ell_{C_p}(X_i)], \end{aligned}$$

where $\ell_C(x)$ denotes the number of bits C uses to encode x. Note that, as in Section 2.4.1, the code $\lceil -\log p(x) \rceil$ is (essentially) optimal. $\diamond$

As another example, we may consider a filtering or prediction problem for a linear system.

Example 16.2.2 (Prediction in a linear system)**:** Suppose we believe that a sequence of random variables $X_i \in \mathbb{R}^d$ are Markovian, where X_i given X_{i-1} is normally distributed with mean $AX_{i-1}+g$, where A is an unknown matrix and $g \in \mathbb{R}^d$ is a constant drift term. Concretely, we assume $X_i \sim \mathsf{N}(AX_{i-1} + g, \sigma^2 I_{d \times d})$, where we assume σ^2 is fixed and known. For our class of predicting distributions Q, we may look at those that at iteration i predict $X_i \sim \mathsf{N}(\mu_i, \sigma^2 I)$. In this case, the regret is given by

$$\mathsf{Reg}(Q, P, x_1^n) = \sum_{i=1}^n \frac{1}{2\sigma^2} \left\| \mu_i - x_i \right\|_2^2 - \frac{1}{2\sigma^2} \left\| Ax_{i-1} + g - x_i \right\|_2^2,$$

while the redundancy is

$$\mathsf{Red}_n(Q, P) = \frac{1}{2\sigma^2} \sum_{i=1}^n \mathbb{E}[\|AX_{i-1} + g - \mu_i(X_1^{i-1})\|_2^2],$$

assuming that P is the linear Gaussian Markov chain specified. $\diamond$

16.3 Minimax strategies for regret

Our definitions in place, we now turn to strategies for attaining the optimal regret in the adversarial setting. We discuss this only briefly, as optimal strategies are somewhat difficult to implement, and the redundancy setting allows (for us) easier exploration.

We begin by describing a notion of complexity that captures the best possible regret in the adversarial setting. In particular, assume without loss of generality that we have a set of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ parameterized by $\theta \in \Theta$, where the distributions are supported on $\mathcal{X}^n$. We define the complexity of the set $\mathcal{P}$ (viz. the complexity of Θ) as

$$\mathsf{Comp}_n(\Theta) := \log \int_{\mathcal{X}^n} \sup_{\theta \in \Theta} p_\theta(x_1^n) dx_1^n \quad \text{or generally} \quad \mathsf{Comp}_n(\Theta) := \log \int_{\mathcal{X}^n} \sup_{\theta \in \Theta} p_\theta(x_1^n) d\nu(x_1^n), \tag{16.3.1}$$

where ν is some base measure on $\mathcal{X}^n$. Note that we may have $\mathsf{Comp}_n(\Theta) = +\infty$, especially when Θ is non-compact. This is not particularly uncommon, for example, consider the case of a normal location family model over $\mathcal{X} = \mathbb{R}$ with $\Theta = \mathbb{R}$.

It turns out that the complexity is precisely the minimax regret in the adversarial setting.

Proposition 16.3.1. *The minimax regret*

$$\inf_Q \mathfrak{R}^{\mathcal{X}}(Q, \mathcal{P}) = \mathsf{Comp}_n(\Theta).$$

Moreover, if $\mathsf{Comp}_n(\Theta) < +\infty$*, then the* normalized maximum likelihood *distribution (also known as the* Shtarkov *distribution)* $\overline{Q}$*, defined with density*

$$\overline{q}(x_1^n) = \frac{\sup_{\theta \in \Theta} p_\theta(x_1^n)}{\int \sup_\theta p_\theta(x_1^n) dx_1^n},$$

is uniquely minimax optimal.

The proposition completely characterizes the minimax regret in the adversarial setting, and it gives the unique distribution achieving the regret. Unfortunately, in most cases it is challenging to compute the minimax optimal distribution $\overline{Q}$, so we must make approximations of some type. One approach is to make Bayesian approximations to $\overline{Q}$, as we do in the sequel when we consider redundancy rather than adversarial regret. See also the book of Grünwald [95] for more discussion of this and other issues.

Proof We begin by proving the result in the case that $\mathsf{Comp}_n < +\infty$. First, note that the normalized maximum likelihood distribution $\overline{Q}$ has constant regret:

$$\begin{aligned}
\mathfrak{R}_n^{\mathcal{X}}(\overline{Q}, \mathcal{P}) &= \sup_{x_1^n \in \mathcal{X}^n} \left[\log \frac{1}{\overline{q}(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)} \right] \\
&= \sup_{x_1^n} \left[\log \frac{\int \sup_\theta p_\theta(x_1^n) dx_1^n}{\sup_\theta p_\theta(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)} \right] = \mathsf{Comp}_n(\mathcal{P}).
\end{aligned}$$

Moreover, for any distribution Q on $\mathcal{X}^n$ we have

$$\begin{aligned}\mathfrak{R}_n^{\mathcal{X}}(Q, \mathcal{P}) &\geq \int \left[\log \frac{1}{q(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)}\right] \overline{q}(x_1^n) dx_1^n \\ &= \int \left[\log \frac{\overline{q}(x_1^n)}{q(x_1^n)} + \mathsf{Comp}_n(\Theta)\right] \overline{q}(x_1^n) dx_1^n \\ &= D_{\rm kl}\left(\overline{Q} \| Q\right) + \mathsf{Comp}_n(\Theta), \end{aligned} \tag{16.3.2}$$

so that $\overline{Q}$ is uniquely minimax optimal, as $D_{\rm kl}\left(\overline{Q}\|Q\right) > 0$ unless $\overline{Q} = Q$.

Now we show how to extend the lower bound (16.3.2) to the case when $\mathsf{Comp}_n(\Theta) = +\infty$. Let us assume without loss of generality that $\mathcal{X}$ is countable and consists of points $x_1, x_2, \ldots$ (we can discretize $\mathcal{X}$ otherwise) and assume we have $n = 1$. Fix any $\epsilon \in (0, 1)$ and construct the sequence $\theta_1, \theta_2, \ldots$ so that $p_{\theta_j}(x_j) \geq (1 - \epsilon) \sup_{\theta \in \Theta} p_\theta(x)$, and define the sets $\Theta_j = \{\theta_1, \ldots, \theta_j\}$. Clearly we have $\mathsf{Comp}(\Theta_j) \leq \log j$, and if we define $\overline{q}_j(x) = \max_{\theta \in \Theta_j} p_\theta(x) / \sum_{x \in \mathcal{X}} \max_{\theta \in \Theta_j} p_\theta(x)$, we may extend the reasoning yielding inequality (16.3.2) to obtain

$$\begin{aligned}\mathfrak{R}^{\mathcal{X}}(Q, \mathcal{P}) &= \sup_{x \in \mathcal{X}} \left[\log \frac{1}{q(x)} - \log \frac{1}{\sup_{\theta \in \Theta} p_\theta(x)}\right] \\ &\geq \sum_x \overline{q}_j(x) \left[\log \frac{1}{q(x)} - \log \frac{1}{\max_{\theta \in \Theta_j} p_\theta(x)}\right] \\ &= \sum_x \overline{q}_j(x) \left[\log \frac{\overline{q}_j(x)}{q(x)} + \log \sum_{x'} \max_{\theta \in \Theta_j} p_\theta(x')\right] = D_{\rm kl}\left(\overline{Q}_j \| Q\right) + \mathsf{Comp}(\Theta_j). \end{aligned}$$

But of course, by noting that

$$\mathsf{Comp}(\Theta_j) \geq (1 - \epsilon) \sum_{i=1}^{j} \sup_\theta p_\theta(x_i) + \sum_{i > j} \max_{\theta \in \Theta_j} p_\theta(x_i) \to +\infty$$

as $j \to \infty$, we obtain the result when $\mathsf{Comp}_n(\Theta) = \infty$. □

We now give an example where (up to constant factor terms) we can explicitly calculate the minimax regret in the adversarial setting. In this case, we compete with the family of i.i.d. Bernoulli distributions.

Example 16.3.2 (Complexity of the Bernoulli distribution)**:** In this example, we consider competing against the family of Bernoulli distributions $\{P_\theta\}_{\theta \in [0,1]}$, where for a point $x \in \{0, 1\}$, we have $P_\theta(x) = \theta^x (1 - \theta)^{1 - x}$. For a sequence $x_1^n \in \{0, 1\}^n$ with m non-zeros, we thus have for $\widehat{\theta} = m/n$ that

$$\sup_{\theta \in [0,1]} P_\theta(x_1^n) = P_{\widehat{\theta}}(x_1^n) = \widehat{\theta}^m (1 - \widehat{\theta})^{n - m} = \exp(-n h_2(\widehat{\theta})),$$

where $h_2(p) = -p \log p - (1 - p) \log(1 - p)$ is the binary entropy. Using this representation, we find that the complexity of the Bernoulli family is

$$\mathsf{Comp}_n([0, 1]) = \log \sum_{m=0}^{n} \binom{n}{m} e^{-n h_2(\frac{m}{n})}.$$

Rather than explicitly compute with this, we now use Stirling's approximation (cf. Cover and Thomas [53, Chapter 17]): for any $p \in (0,1)$ with $np \in \mathbb{N}$, we have

$$\binom{n}{np} \in \frac{1}{\sqrt{n}} \left[\frac{1}{\sqrt{8p(1-p)}}, \frac{1}{\sqrt{\pi p(1-p)}} \right] \exp(nh_2(p)).$$

Thus, by dealing with the boundary cases $m = n$ and $m = 0$ explicitly, we obtain

$$\begin{aligned} \sum_{m=0}^{n} \binom{n}{m} \exp(-nh_2(\frac{m}{n})) &= 2 + \sum_{m=1}^{n-1} \binom{n}{m} \exp(-nh_2(\frac{m}{n})) \\ &\in 2 + \left[\frac{1}{\sqrt{8}}, \frac{1}{\sqrt{\pi}} \right] \frac{1}{\sqrt{n}} \underbrace{\sum_{m=1}^{n-1} \frac{1}{\sqrt{\frac{m}{n}(1-\frac{m}{n})}}}_{\to n \int_0^1 (\theta(1-\theta))^{-\frac{1}{2}}}, \end{aligned}$$

the noted asymptote occuring as $n \to \infty$ by the fact that this sum is a Riemann sum for the integral $\int_0^1 \theta^{-1/2}(1-\theta)^{-1/2} d\theta$. In particular, we have that as $n \to \infty$,

$$\begin{aligned} \inf_Q \mathfrak{R}_n^{\mathcal{X}}(Q, \mathcal{P}) = \mathsf{Comp}_n([0,1]) &= \log \left(2 + [8^{-1/2}, \pi^{-1/2}] n^{1/2} \int_0^1 \frac{1}{\sqrt{\theta(1-\theta)}} d\theta \right) + o(1) \\ &= \frac{1}{2} \log n + \log \int_0^1 \frac{1}{\sqrt{\theta(1-\theta)}} d\theta + O(1). \end{aligned}$$

We remark in passing that this is equal to $\frac{1}{2} \log n + \log \int_0^1 \sqrt{I_\theta} d\theta$, where I_θ denotes the Fisher information of the Bernoulli family (recall Example 15.1.2). We will see that this holds in more generality, at least for redundancy, in the sequel. $\diamond$

16.4 Mixture (Bayesian) strategies and redundancy

We now turn to a slightly less adversarial setting, where we assume that we compete against a random sequence X_1^n of data, drawn from some fixed distribution P, rather than an adversarially chosen sequence x_1^n. Thinking of this problem as a game, we choose a distribution Q according to which we make predictions (based on previous data), and nature chooses a distribution $P_\theta \in \mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$. In the simplest case—upon which we focus—the data X_1^n are then generated i.i.d. according to P_θ, and we suffer expected regret (or redundancy)

$$\mathsf{Red}_n(Q, P_\theta) = \mathbb{E}_\theta \left[\log \frac{1}{q(X_1^n)} \right] - \mathbb{E}_\theta \left[\log \frac{1}{p_\theta(X_1^n)} \right] = D_{\rm kl} \left(P_\theta^n \| Q_n \right), \tag{16.4.1}$$

where we use Q_n to denote that Q is applied on all n data points (in a sequential fashion, as $Q(\cdot \mid X_1^{i-1})$). In this expression, q and p denote the densities of Q and P, respectively. In a slightly more general setting, we may consider the expected regret of Q with respect to a distribution P_θ even under model mis-specification, meaning that the data is generated according to an alternate distribution P. In this case, the (more general) redundancy becomes

$$\mathbb{E}_P \left[\log \frac{1}{q(X_1^n)} - \log \frac{1}{p_\theta(X_1^n)} \right]. \tag{16.4.2}$$

In both cases (16.4.1) and (16.4.2), we would like to be able to guarantee that the redundancy grows more slowly than n as $n \to \infty$. That is, we would like to find distributions Q such that, for any $\theta_0 \in \Theta$, we have $\frac{1}{n} D_{\rm kl}\left(P_{\theta_0}^n \| Q_n\right) \to 0$ as $n \to \infty$. Assuming we could actually obtain such a distribution in general, this is interesting because (even in the i.i.d. case) for *any* fixed distribution $P_\theta \neq P_{\theta_0}$, we must have $D_{\rm kl}\left(P_{\theta_0}^n \| P_\theta^n\right) = n D_{\rm kl}\left(P_{\theta_0} \| P_\theta^n\right) = \Omega(n)$. A standard approach to attaining such guarantees is the *mixture approach*, which is based on choosing Q as a convex combination (mixture) of all the possible source distributions P_θ for $\theta \in \Theta$.

In particular, given a prior distribution π (weighting function integrating to 1) over Θ, we define the mixture distribution

$$Q_n^\pi(A) = \int_\Theta \pi(\theta) P_\theta(A) d\theta \quad \text{for} \quad A \subset \mathcal{X}^n. \tag{16.4.3}$$

Rewriting this in terms of densities p_θ, we have

$$q_n^\pi(x_1^n) = \int_\Theta \pi(\theta) p_\theta(x_1^n) d\theta.$$

Conceptually, this gives a simple prediction scheme, where at iteration i we play the density

$$q^\pi(x_i \mid x_1^{i-1}) = \frac{q^\pi(x_1^i)}{q^\pi(x_1^{i-1})},$$

which is equivalent to playing

$$q^\pi(x_i \mid x_1^{i-1}) = \int_\Theta q(x_i, \theta \mid x_1^{i-1}) d\theta = \int_\Theta p_\theta(x_i) \pi(\theta \mid x_1^{i-1}) d\theta,$$

by construction of the distributions Q^π as mixtures of i.i.d. P_θ. Here the posterior distribution $\pi(\theta \mid x_1^{i-1})$ is given by

$$\pi(\theta \mid x_1^{i-1}) = \frac{\pi(\theta) p_\theta(x_1^{i-1})}{\int_\Theta \pi(\theta') p_{\theta'}(x_1^{i-1}) d\theta'} = \frac{\pi(\theta) \exp\left(-\log \frac{1}{p_\theta(x_1^{i-1})}\right)}{\int_\Theta \pi(\theta') p_{\theta'}(x_1^{i-1}) d\theta'}, \tag{16.4.4}$$

where we have emphasized that this strategy exhibits an *exponential weighting* approach, where distribution weights are scaled exponentially by their previous loss performance of $\log 1/p_\theta(x_1^{i-1})$.

This mixture construction (16.4.3), with the weighting scheme (16.4.4), enjoys very good performance. In fact, we say that so long as the prior π puts non-zero mass over all of Θ, under some appropriate smoothness conditions, the scheme Q^π is universal, meaning that $D_{\rm kl}\left(P_\theta^n \| Q_n^\pi\right) = o(n)$. We have the following theorem illustrating this effect. In the theorem, we let π be a density on Θ, and we assume the Fisher information I_θ for the family $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ exists in a neighborhood of $\theta_0 \in \operatorname{int} \Theta$, and that the distributions P_θ are sufficiently regular that differentiation and integration can be interchanged. (See Clarke and Barron [49] for precise conditions.) We have

Theorem 16.4.1 (Clarke and Barron [49])**.** *Under the above conditions, if $Q_n^\pi = \int P_\theta^n \pi(\theta) d\theta$ is the mixture (16.4.3), then*

$$D_{\rm kl}\left(P_{\theta_0}^n \| Q_n^\pi\right) - \frac{d}{2} \log \frac{n}{2\pi e} \to \log \frac{1}{\pi(\theta_0)} + \frac{1}{2} \log \det(I_{\theta_0}) \quad \text{as } n \to \infty. \tag{16.4.5}$$

While we do not rigorously prove the theorem, we give a sketch showing the main components of the result based on asymptotic normality arguments for the maximum likelihood estimator in Section 16.5. See Clarke and Barron [49] for a full proof.

Example 16.4.2 (Bernoulli distributions with a Beta prior)**:** Consider the class of binary (i.i.d. or memoryless) Bernoulli sources, that is, the X_i are i.i.d $\mathsf{Bernoulli}(\theta)$, where $\theta = P_\theta(X = 1) \in [0, 1]$. The $\mathsf{Beta}(\alpha, \beta)$-distribution prior on θ is the mixture π with density

$$\pi(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}$$

on $[0, 1]$, where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ denotes the gamma function. We remark that that under the $\mathsf{Beta}(\alpha, \beta)$ distribution, we have $\mathbb{E}_\pi[\theta] = \frac{\alpha}{\alpha+\beta}$. (See any undergraduate probability text for such results.)

If we play via a mixture of Bernoulli distributions under such a Beta-prior for θ, by Theorem 16.4.1 we have a universal prediction scheme. We may also explicitly calculate the predictive distribution Q. To do so, we first compute the posterior $\pi(\theta \mid X_1^i)$ as in expression (16.4.4). Let $S_i = \sum_{j=1}^i X_j$ be partial sum of the Xs up to iteration i. Then

$$\pi(\theta \mid x_1^i) = \frac{p_\theta(x_1^i)\pi(\theta)}{q(x_1^i)} \propto \theta^{S_i}(1 - \theta)^{i-S_i}\theta^{\alpha-1}\theta^{\beta-1} = \theta^{\alpha+S_i-1}(1 - \theta)^{\beta+i-S_i-1},$$

where we have ignored the denominator as we must simply normalize the above quantity in θ. But by inspection, the posterior density of $\theta \mid X_1^i$ is a $\mathsf{Beta}(\alpha + S_i, \beta + i - S_i)$ distribution. Thus to compute the predictive distribution, we note that $\mathbb{E}_\theta[X_i] = \theta$, so we have

$$Q(X_i = 1 \mid X_1^i) = \mathbb{E}_\pi[\theta \mid X_1^i] = \frac{S_i + \alpha}{i + \alpha + \beta}.$$

Moreover, Theorem 16.4.1 shows that when we play the prediction game with a $\mathsf{Beta}(\alpha, \beta)$-prior, we have redundancy scaling as

$$D_{\rm kl}\left(P_{\theta_0}^n \| Q_n^\pi\right) = \frac{1}{2}\log\frac{n}{2\pi e} + \log\left[\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}\frac{1}{\theta_0^{\alpha-1}(1 - \theta_0)^{\beta-1}}\right] + \frac{1}{2}\log\frac{1}{\theta_0(1 - \theta_0)} + o(1)$$

for $\theta_0 \in (0, 1)$. $\diamond$

As one additional interesting result, we show that mixture models are actually quite robust, even under model mis-specification, that is, when the true distribution generating the data does not belong to the class $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$. That is, mixtures can give good performance for the generalized redundancy quantity (16.4.2). For this next result, we as usual define the mixture distribution Q^π over the set $\mathcal{X}$ via $Q^\pi(A) = \int_\Theta P_\theta(A)d\pi(\theta)$. We may also restrict this mixture distribution to a subset $\Theta_0 \subset \Theta$ by defining

$$Q_{\Theta_0}^\pi(A) = \frac{1}{\pi(\Theta_0)}\int_{\Theta_0} P_\theta(A)d\pi(\theta).$$

Then we obtain the following robustness result.

Proposition 16.4.3. *Assume that P_θ have densities p_θ over $\mathcal{X}$, let P be any distribution having density p over $\mathcal{X}$, and let q^π be the density associated with Q^π. Then for any $\Theta_0 \subset \Theta$,*

$$\mathbb{E}_P\left[\log\frac{1}{q^\pi(X)} - \log\frac{1}{p_\theta(X)}\right] \le \log\frac{1}{\pi(\Theta_0)} + D_{\rm kl}\left(P\|Q_{\Theta_0}^\pi\right) - D_{\rm kl}\left(P\|P_\theta\right).$$

In particular, Proposition 16.4.3 shows that so long as the mixture distributions $Q^\pi_{\Theta_0}$ can closely approximate P_θ, then we attain a convergence guarantee nearly as good as any in the family $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$. (This result is similar in flavor to the mutual information bound (8.7.2), Corollary 8.7.2, and the *index of resolvability* quantity.)

Proof Fix any $\Theta_0 \subset \Theta$. Then we have $q^\pi(x) = \int_\Theta p_\theta(x)d\pi(\theta) \geq \int_{\Theta_0} p_\theta(x)d\pi(\theta)$. Thus we have

$$\mathbb{E}_P\left[\log \frac{p(X)}{q^\pi(X)}\right] \leq \mathbb{E}_P\left[\inf_{\Theta_0\subset\Theta} \log \frac{p(X)}{\int_{\Theta_0} p_\theta(x)d\pi(\theta)}\right]$$
$$= \mathbb{E}_P\left[\inf_{\Theta_0} \log \frac{p(X)\pi(\Theta_0)}{\pi(\Theta_0)\int_{\Theta_0} p_\theta(x)d\pi(\theta)}\right] = \mathbb{E}_P\left[\inf_{\Theta_0} \log \frac{p(X)}{\pi(\Theta_0)q^\pi_{\Theta_0}(X)}\right].$$

This is certainly smaller than the same quantity with the infimum outside the expectation, and noting that

$$\mathbb{E}_P\left[\log \frac{1}{q^\pi(X)} - \log \frac{1}{p_\theta(X)}\right] = \mathbb{E}_P\left[\log \frac{p(X)}{q^\pi(X)}\right] - \mathbb{E}_P\left[\log \frac{p(X)}{p_\theta(X)}\right]$$

gives the result. □

16.4.1 Bayesian redundancy and objective, reference, and Jeffreys priors

We can also imagine a slight variant of the redundancy game we have described to this point. Instead of choosing a distribution Q and allowing nature to choose a distribution P_θ, we could switch the order of the game. In particular, we could assume that nature first chooses prior distribution π on θ, and without seeing θ (but with knowledge of the distribution π) we choose the predictive distribution Q. This leads to the *Bayesian redundancy*, which is simply the expected redundancy we suffer:

$$\int_\Theta \pi(\theta) D_{\rm kl}\left(P^n_\theta \| Q_n\right) d\theta.$$

However, recalling our calculations with mutual information (equations (8.4.4) and (8.7.3)), we know that the Bayes-optimal prediction distribution is Q^π_n. In particular, if we let T denote a random variable distributed according to π, and conditional on $T = \theta$ assume that the X_i are drawn according to P_θ, we have that the mutual information between T and X_1^n is

$$I_\pi(T; X_1^n) = \int \pi(\theta) D_{\rm kl}\left(P^n_\theta \| Q^\pi_n\right) d\theta = \inf_Q \int \pi(\theta) D_{\rm kl}\left(P^n_\theta \| Q\right) d\theta. \tag{16.4.6}$$

With Theorem 16.4.1 in hand, we can give a somewhat more nuanced picture of this mutual information quantity. As a first consequence of Theorem 16.4.1, we have that

$$I_\pi(T; X_1^n) = \frac{d}{2}\log\frac{n}{2\pi e} + \int \log \frac{\sqrt{\det I_\theta}}{\pi(\theta)}\pi(\theta)d\theta + o(1), \tag{16.4.7}$$

where I_θ denotes the Fisher information matrix for the family $\{P_\theta\}_{\theta\in\Theta}$. One strand of Bayesian statistics—we will not delve too deeply into this now, instead referring to the survey by Bernardo [26]—known as reference analysis, advocates that in performing a Bayesian analysis, we should choose the prior π that maximizes the mutual information between the parameters θ about which we wish to make inferences and any observations X_1^n available. Moreover, in this set of strategies,

one allows n to tend to ∞, as we wish to take advantage of any data we might actually see. The asymptotic formula (16.4.7) allows us to choose such a prior.

In a different vein, Jeffreys [116] proposed that if the square root of the determinant of the Fisher information was integrable, then one should take π as

$$\pi_{\text{jeffreys}}(\theta) = \frac{\sqrt{\det I_\theta}}{\int_\Theta \sqrt{\det I_\theta} d\theta}$$

known as the *Jeffreys prior*. Jeffreys originally proposed this for invariance reasons, as the inferences made on the parameter θ under the prior π_{jeffreys} are identical to those made on a transformed parameter $\phi(\theta)$ under the appropriately transformed Jeffreys prior. The asymptotic expression (16.4.7), however, shows that the Jeffreys prior is the asymptotic reference prior. Indeed, computing the integral in (16.4.7), we have

$$\begin{aligned}\int_\Theta \pi(\theta) \log \frac{\sqrt{\det I_\theta}}{\pi(\theta)} d\theta &= \int_\Theta \pi(\theta) \log \frac{\pi_{\text{jeffreys}}(\theta)}{\pi(\theta)} d\theta + \log \int \sqrt{\det I_\theta} d\theta \\ &= -D_{\rm kl}\left(\pi \| \pi_{\text{jeffreys}}\right) + \log \int \sqrt{\det I_\theta} d\theta,\end{aligned}$$

whenever the Jeffreys prior exists. Moreover, we see that in an asymptotic sense, the worst-case prior distribution π for nature to play is given by the Jeffreys prior, as otherwise the $-D_{\rm kl}\left(\pi \| \pi_{\text{jeffreys}}\right)$ term in the expected (Bayesian) redundancy is negative.

Example 16.4.4 (Jeffreys priors and the exponential distribution)**:** Let us now assume that our source distributions P_θ are exponential distributions, meaning that $\theta \in (0, \infty)$ and we have density $p_\theta(x) = \exp(-\theta x - \log \frac{1}{\theta})$ for $x \in [0, \infty)$. This is clearly an exponential family model, and the Fisher information is easy to compute as $I_\theta = \frac{\partial^2}{\partial \theta^2} \log \frac{1}{\theta} = 1/\theta^2$ (cf. Example 15.1.1). In this case, the Jeffreys prior is $\pi_{\text{jeffreys}}(\theta) \propto \sqrt{I} = 1/\theta$, but this "density" does not integrate over $[0, \infty)$. One approach to this difficulty, advocated by Bernardo [26, Definition 3] (among others) is to just proceed formally and notice that after observing a single datapoint, the "posterior" distribution $\pi(\theta \mid X)$ is well-defined. Following this idea, note that after seeing some data $X_1, \ldots, X_i$, with $S_i = \sum_{j=1}^i X_j$ as the partial sum, we have

$$\pi(\theta \mid x_1^i) \propto p_\theta(x_1^i) \pi_{\text{jeffreys}}(\theta) = \theta^i \exp\left(-\theta \sum_{j=1}^i x_j\right) \frac{1}{\theta} = \theta^{i-1} \exp(-\theta S_i).$$

Integrating, we have for $s_i = \sum_{j=1}^i x_j$

$$q(x \mid x_1^i) = \int_0^\infty p_\theta(x) \pi(\theta \mid x_1^i) d\theta \propto \int_0^\infty \theta e^{-\theta x} \theta^{i-1} e^{-\theta s_i} d\theta = \frac{1}{(s_i + x)^{i+1}} \int_0^\infty u^i e^{-u} du,$$

where we made the change of variables $u = \theta(s_i + x)$. This is at least a distribution that normalizes, so often one simply assumes the existence of a piece of fake data. For example, by saying we "observe" $x_0 = 1$, we have prior proportional to $\pi(\theta) = e^{-\theta}$, which yields redundancy

$$D_{\rm kl}\left(P_{\theta_0}^n \| Q_n^\pi\right) = \frac{1}{2} \log \frac{n}{2\pi e} + \theta_0 + \log \frac{1}{\theta_0} + o(1).$$

The difference is that, in this case, the redundancy bound is no longer uniform in θ_0, as it would be for the true reference (or Jeffreys, if it exists) prior. $\diamond$

16.4.2 Redundancy capacity duality

Let us discuss Bayesian redundancy versus worst-case redundancy in somewhat more depth. If we play a game where nature chooses T according to the known prior π, and draws data $X_1^n \sim P_\theta$ conditional on $T = \theta$, then we know that as in expression (16.4.7), we have

$$\inf_Q \mathbb{E}_\pi \left[D_{\rm kl} \left(P_T^n \| Q \right) \right] = \int D_{\rm kl} \left(P_\theta^n \| Q_n^\pi \right) \pi(\theta) d\theta = I_\pi(T; X_1^n).$$

A natural question that arises from this expression is the following: if nature chooses a worst-case prior, can we swap the order of maximization and minimization? That is, do we ever have the equality

$$\sup_\pi I_\pi(T; X_1^n) = \inf_Q \sup_\theta D_{\rm kl} \left(P_\theta^n \| Q \right),$$

so that the worst-case Bayesian redundancy is actually the minimax redundancy? It is clear that if nature can choose the worst case P_θ after we choose Q, the redundancy must be at least as bad as the Bayesian redundancy, so

$$\sup_\pi I_\pi(T; X_1^n) \le \inf_Q \sup_\theta D_{\rm kl} \left(P_\theta^n \| Q \right) = \inf_Q \mathfrak{R}_n(Q, \mathcal{P}).$$

Indeed, if this inequality were an equality, then for the worst-case prior π^*, the mixture $Q_n^{\pi^*}$ would be minimax optimal.

In fact, the redundancy-capacity theorem, first proved by Gallager [88], and extended by Haussler [101] (among others) allows us to do just that. That is, if we must choose a distribution Q and then nature chooses P_θ adversarially, we can guarantee to worse redundancy than in the (worst-case) Bayesian setting. We state a simpler version of the result that holds when the random variables X take values in finite spaces; Haussler's more general version shows that the next theorem holds whenever $X \in \mathcal{X}$ and $\mathcal{X}$ is a complete separable metric space.

Theorem 16.4.5 (Gallager [88])**.** *Let X be a random variable taking on a finite number of values and Θ be a measurable space. Then*

$$\sup_\pi \inf_Q \int D_{\rm kl} \left(P_\theta \| Q \right) d\pi(\theta) = \sup_\pi I_\pi(T; X) = \inf_Q \sup_{\theta \in \Theta} D_{\rm kl} \left(P_\theta \| Q \right).$$

Moreover, the infimum on the right is uniquely attained by some distribution Q^*, and if π^* attains the supremum on the left, then $Q^* = \int P_\theta d\pi^*(\theta)$.

See Section 16.6 for a proof of Theorem 16.4.5.

This theorem is known as the *redundancy-capacity* theorem in the literature, because in classical information theory, the capacity of a noisy channel $T \to X_1^n$ is the maximal mutual informationx $\sup_\pi I_\pi(T; X_1^n)$. In the exercises, you explore some robustness properties of the optimal distribution Q^π in relation to this theorem. In short, though, we see that if there is a capacity achieving prior, then the associated mixture distribution Q^π is minimax optimal and attains the minimax redundancy for the game.

16.5 Asymptotic normality and Theorem 16.4.1

In this section, we very briefly (and very hand-wavily) justify the asymptotic expression (16.4.5). To do this, we argue that (roughly) the posterior distribution $\pi(\theta \mid X_1^n)$ should be roughly normally

distributed with appropriate variance measure, which gives the result. We now give the intuition for this statement, first by heuristically deriving the asymptotics of a maximum likelihood estimator, then by looking at the Bayesian case. (Clarke and Barron [49] provide a fully rigorous proof.)

16.5.1 Heuristic justification of asymptotic normality

First, we sketch the asymptotic normality of the maximum likelihood estimator $\widehat{\theta}$, that is, $\widehat{\theta}$ is chosen to maximize $\log p_\theta(X_1^n)$. (See, for example, Lehmann and Casella [130] for more rigorous arguments.) Assume that the data are generated i.i.d. according to P_{θ_0}. Then by assumption that $\widehat{\theta}$ maximizes the log-likelihood, we have the stationary condition $0 = \nabla \log p_{\widehat{\theta}}(X_1^n)$. Performing a Taylor expansion of this quantity about θ_0, we have

$$0 = \nabla \log p_{\widehat{\theta}}(X_1^n) = \nabla \log p_{\theta_0}(X_1^n) + \nabla^2 \log p_{\theta_0}(X_1^n)(\widehat{\theta} - \theta_0) + R$$

where R is a remainder term. Assuming that $\widehat{\theta} \to \theta_0$ at any reasonable rate (this can be made rigorous), this remainder is negligible asymptotically.

Rearranging this equality, we obtain

$$\begin{aligned}
\widehat{\theta} - \theta_0 &\approx (-\nabla^2 \log p_{\theta_0}(X_1^n))^{-1} \nabla \log p_{\theta_0}(X_1^n) \\
&= \frac{1}{n} \underbrace{\left(-\frac{1}{n} \sum_{i=1}^n \nabla^2 \log p_{\theta_0}(X_i) \right)}_{\approx I_{\theta_0}}{}^{-1} \sum_{i=1}^n \nabla \log p_{\theta_0}(X_i) \\
&\approx \frac{1}{n} I_{\theta_0}^{-1} \sum_{i=1}^n \nabla \log p_{\theta_0}(X_i),
\end{aligned}$$

where we have used that the Fisher information $I_\theta = -\mathbb{E}_\theta[\nabla^2 \log p_\theta(X)]$ and the law of large numbers. By the (multivariate) central limit theorem, we then obtain the asymptotic normality result

$$\sqrt{n}(\widehat{\theta} - \theta_0) \approx \frac{1}{\sqrt{n}} I_{\theta_0}^{-1} \sum_{i=1}^n \nabla \log p_{\theta_0}(X_i) \xrightarrow{d} \mathsf{N}(0, I_{\theta_0}^{-1}),$$

where $\xrightarrow{d}$ denotes convergence in distribution, with asymptotic variance

$$I_{\theta_0}^{-1} \mathbb{E}_{\theta_0}[\nabla \log p_{\theta_0}(X) \nabla \log p_{\theta_0}(X)^\top] I_{\theta_0}^{-1} = I_{\theta_0}^{-1} I_{\theta_0} I_{\theta_0}^{-1} = I_{\theta_0}^{-1}.$$

Completely heuristically, we also write

$$\widehat{\theta} \text{ “}\sim\text{” } \mathsf{N}(\theta_0, (nI_{\theta_0})^{-1}). \tag{16.5.1}$$

16.5.2 Heuristic calculations of posterior distributions and redundancy

With the asymptotic distributional heuristic (16.5.1), we now look at the redundancy and posterior distribution of θ conditioned on the data X_1^n when the data are drawn i.i.d. P_{θ_0}. When Q_n^π is the mixture distribution associated with π, the posterior density of $\theta \mid X_1^n$ is

$$\pi(\theta \mid X_1^n) = \frac{p_\theta(X_1^n)\pi(\theta)}{q_n(X_1^n)}.$$

By our heuristic calculation of the MLE, this density (assuming the data overwhelms the prior) is approximately a normal density with mean θ_0 and variance $(nI_{\theta_0})^{-1}$, where we have used expression (16.5.1). Expanding the redundancy, we obtain

$$\mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_1^n)}{q_n(X_1^n)}\right] = \mathbb{E}_{\theta_0}\left[\log \frac{p_{\widehat{\theta}}(X_1^n)\pi(\widehat{\theta})}{q_n(X_1^n)}\right] + \mathbb{E}_{\theta_0}\left[\log \frac{1}{\pi(\widehat{\theta})}\right] + \mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_1^n)}{p_{\widehat{\theta}}(X_1^n)}\right]. \tag{16.5.2}$$

Now we use our heuristic. We have that

$$\mathbb{E}_{\theta_0}\left[\log \frac{p_{\widehat{\theta}}(X_1^n)\pi(\widehat{\theta})}{q_n(X_1^n)}\right] \approx \log \frac{1}{(2\pi)^{d/2}\det(nI_{\theta_0})^{-1/2}} + \mathbb{E}_{\theta_0}\left[-\frac{1}{2}(\widehat{\theta}-\theta_0)^\top (nI_{\theta_0})^{-1}(\widehat{\theta}-\theta_0)\right],$$

by the asymptotic normality result, $\pi(\widehat{\theta}) = \pi(\theta_0) + O(1/\sqrt{n})$ again by the asymptotic normality result, and

$$\begin{aligned}
\log p_{\widehat{\theta}}(X_1^n) &\approx \log p_{\theta_0}(X_1^n) + \left(\sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)\right)^\top (\widehat{\theta}-\theta_0) \\
&\approx \log p_{\theta_0}(X_1^n) + \left(\sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)\right)^\top I_{\theta_0}^{-1}\left(\frac{1}{n}\sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)\right).
\end{aligned}$$

Substituting these three into the redundancy expression (16.5.2), we obtain

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_1^n)}{q_n(X_1^n)}\right] &\approx \log \frac{1}{(2\pi)^{d/2}\det(nI_{\theta_0})^{-1/2}} + \mathbb{E}_{\theta_0}\left[-\frac{1}{2}(\widehat{\theta}-\theta_0)^\top (nI_{\theta_0})^{-1}(\widehat{\theta}-\theta_0)\right] \\
&\quad + \log \frac{1}{\pi(\theta_0)} - \mathbb{E}_{\theta_0}\left[\left(\sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)\right)^\top I_{\theta_0}^{-1}\left(\frac{1}{n}\sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)\right)\right] \\
&= \frac{d}{2}\log\frac{n}{2\pi} + \frac{1}{2}\log\det(I_{\theta_0}) + \log\frac{1}{\pi(\theta_0)} - d + R,
\end{aligned}$$

where R is a remainder term. This gives the major terms in the asymptotic result in Theorem 16.4.1.

16.6 Proof of Theorem 16.4.5

In this section, we prove one version of the strong saddle point results associated with the universal prediction game as given by Theorem 16.4.5 (in the case that X belongs to a finite set). For shorthand, we recall the definition of the redundancy

$$\mathsf{Red}(Q,\theta) := \mathbb{E}_{P_\theta}\left[-\log Q(X) + \log P_\theta(X)\right] = D_{\rm kl}\left(P_\theta \| Q\right),$$

where we have assumed that X belongs to a finite set, so that $Q(X)$ is simply the probability of X. For a given prior distribution π on θ, we define the expected redundancy as

$$\mathsf{Red}(Q,\pi) = \int D_{\rm kl}\left(P_\theta \| Q\right) d\pi(\theta).$$

Our goal is to show that the max-min value of the prediction game is the same as the min-max value of the game, that is,

$$\sup_\pi I_\pi(T;X) = \sup_\pi \inf_Q \mathsf{Red}(Q,\pi) = \inf_Q \sup_{\theta\in\Theta} \mathsf{Red}(Q,\theta).$$

Proof We know that the max-min risk (worst-case Bayes risk) of the game is $\sup_\pi I_\pi(T; X)$; it remains to show that this is the min-max risk. To that end, define the *capacity* of the family $\{P_\theta\}_{\theta \in \Theta}$ as

$$C := \sup_\pi I_\pi(T; X). \tag{16.6.1}$$

Notably, this constant is finite (because $I_\pi(T; X) \le \log |\mathcal{X}|$), and there exists a sequence π_n of prior probabilities such that $I_{\pi_n}(T; X) \to C$. Now, let $\bar{Q}$ be any cluster point of the sequence of mixtures $Q^{\pi_n} = \int P_\theta d\pi_n(\theta)$; such a point exists because the space of probability distributions on the finite set $\mathcal{X}$ is compact. We will show that

$$\sum_x P_\theta(x) \log \frac{P_\theta(x)}{\bar{Q}(x)} \le C \quad \text{for all } \theta \in \Theta, \tag{16.6.2}$$

and we claim this is sufficient for the theorem. Indeed, suppose that inequality (16.6.2) holds. Then in this case, we have

$$\inf_Q \sup_{\theta \in \Theta} \mathsf{Red}(Q, \theta) \le \sup_{\theta \in \Theta} \mathsf{Red}(\bar{Q}, \theta) = \sup_{\theta \in \Theta} D_{\rm kl}\left(P_\theta \| \bar{Q}\right) \le C,$$

which implies the theorem, because it is always the case that

$$\sup_\pi \inf_Q \mathsf{Red}(Q, \theta) \le \inf_Q \sup_\pi \mathsf{Red}(Q, \pi) = \inf_Q \sup_{\theta \in \Theta} \mathsf{Red}(Q, \theta).$$

For the sake of contradiction, let us assume that there exists some $\theta \in \Theta$ such that inequality (16.6.2) fails, call it θ^*. We will then show that suitable mixtures $(1 - \lambda)\pi + \lambda\delta_{\theta^*}$, where δ_{θ^*} is the point mass on θ^*, could increase the capacity (16.6.1). To that end, for shorthand define the mixtures

$$\pi_{n,\lambda} = (1 - \lambda)\pi_n + \lambda\delta_{\theta^*} \quad \text{and} \quad Q^{\pi_n, \lambda} = (1 - \lambda)Q^{\pi_n} + \lambda P_{\theta^*}$$

for $\lambda \in [0, 1]$. Let us also use the notation $H_w(X \mid T)$ to denote the conditionaly entropy of the random variable X on T (when T is distributed as w), and we abuse notation by writing $H(X) = H(P)$ when X is distributed as P. In this case, it is clear that we have

$$H_{\pi_{n,\lambda}}(X \mid T) = (1 - \lambda)H_{\pi_n}(X \mid T) + \lambda H(X \mid T = \theta^*),$$

and by definition of the mutual information we have

$$\begin{aligned} I_{\pi_{n,\lambda}}(T; X) &= H_{\pi_{n,\lambda}}(X) - H_{\pi_{n,\lambda}}(X \mid T) \\ &= H((1 - \lambda)Q^{\pi_n} + \lambda P_{\theta^*}) - (1 - \lambda)H_{\pi_n}(X \mid T) - \lambda H(X \mid T = \theta^*). \end{aligned}$$

To demonstrate our contradiction, we will show two things: first, that at $\lambda = 0$ the limits of both sides of the preceding display are equal to the capacity C, and second, that the derivative of the right hand side is positive. This will contradict the definition (16.6.1) of the capacity.

To that end, note that

$$\lim_n H_{\pi_n}(X \mid T) = \lim_n H_{\pi_n}(X) - I_{\pi_n}(T; X) = H(\bar{Q}) - C,$$

by the continuity of the entropy function. Thus, we have

$$\lim_n I_{\pi_n,\lambda}(T; X) = H((1 - \lambda)\bar{Q} + \lambda P_{\theta^*}) - (1 - \lambda)(H(\bar{Q}) - C) - \lambda H(P_\theta). \tag{16.6.3}$$

It is clear that at $\lambda = 0$, both sides are equal to the capacity C, while taking derivatives with respect to λ we have

$$\frac{\partial}{\partial \lambda} H((1-\lambda)\bar{Q} + \lambda P_{\theta^*}) = -\sum_x (P_{\theta^*}(x) - \bar{Q}(x)) \log \left((1-\lambda)\bar{Q}(x) + \lambda P_{\theta^*}(x)\right).$$

Evaluating this derivative at $\lambda = 0$, we find

$$\begin{aligned}
&\frac{\partial}{\partial \lambda} \lim_n I_{\pi_n,\lambda}(T;X)\Big|_{\lambda=0} \\
&\quad = -\sum_x P_{\theta^*}(x) \log \bar{Q}(x) + \sum_x \bar{Q}(x) \log \bar{Q}(x) + H(\bar{Q}) - C + \sum_x P_{\theta^*}(x) \log P_{\theta^*}(x) \\
&\quad = \sum_x P_{\theta^*}(x) \log \frac{P_{\theta^*}(x)}{\bar{Q}(x)} - C.
\end{aligned}$$

In particular, if inequality (16.6.2) fails to hold, then $\frac{\partial}{\partial \lambda} \lim_n I_{\pi_n,\lambda}(T;X)|_{\lambda=0} > 0$, contradicting the definition (16.6.1) of the channel capacity.

The uniqueness of the result follows from the strict convexity of the mutual information I in the mixture channel $\bar{Q}$. □

16.7 Exercises

Exercise 16.1 (Minimax redundancy and different loss functions)**:** In this question, we consider expected losses under the Bernoulli distribution. Assume that $X_i \stackrel{\text{iid}}{\sim} \mathsf{Bernoulli}(p)$, meaning that $X_i = 1$ with probability p and $X_i = 0$ with probability $1-p$. We consider four different loss functions, and their associated expected regret, for measuring the accuracy of our predictions of such X_i. For each of the four choices below, we prove expected regret bounds on

$$\mathsf{Red}_n(\widehat{\theta}, P, \ell) := \sum_{i=1}^n \mathbb{E}_P[\ell(\widehat{\theta}(X_1^{i-1}), X_i)] - \inf_\theta \sum_{i=1}^n \mathbb{E}_P[\ell(\theta, X_i)], \tag{16.7.1}$$

where $\widehat{\theta}$ is a predictor based on $X_1, \ldots, X_{i-1}$ at time i. Define $S_i = \sum_{j=1}^i X_j$ to be the partial sum up to time i. For each of parts (a)–(c), at time i use the predictor

$$\widehat{\theta}_i = \widehat{\theta}(X_1^{i-1}) = \frac{S_{i-1} + \frac{1}{2}}{i}.$$

(a) Loss function: $\ell(\theta, x) = \frac{1}{2}(x-\theta)^2$. Show that $\mathsf{Red}_n(\widehat{\theta}, P, \ell) \le C \cdot \log n$ where C is a constant.

(b) Loss function: $\ell(\theta, x) = x \log \frac{1}{\theta} + (1-x) \log \frac{1}{1-\theta}$, the usual log loss for predicting probabilities. Show that $\mathsf{Red}_n(\widehat{\theta}, P, \ell) \le C \cdot \log n$ whenever the true probability $p \in (0,1)$, where C is a constant. *Hint: Note that there exists a prior π for which $\widehat{\theta}$ is a Bayes strategy. What is this prior?*

(c) Loss function: $\ell(\theta, x) = |x - \theta|$. Show that $\mathsf{Red}_n(\widehat{\theta}, P, \ell) \ge c \cdot n$, where $c > 0$ is a constant, whenever the true probability $p \notin \{0, \frac{1}{2}, 1\}$.

(d) **Extra credit:** Show that there is a numerical constant $c > 0$ such that for any procedure $\widehat{\theta}$, the worst-case redundancy $\sup_{p\in[0,1]} \mathsf{Red}_n(\widehat{\theta}, \mathsf{Bernoulli}(p), \ell) \geq c\sqrt{n}$ for the absolute loss ℓ in part (c). Give a strategy attaining this redundancy.

Exercise 16.2 (Strong versions of redundancy)**:** Assume that for a given $\theta \in \Theta$ we draw $X_1^n \sim P_\theta$. We define the Bayes redundancy for a family of distributions $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ as

$$C_n^\pi := \inf_Q \int D_{\rm kl}\left(P_\theta \| Q\right) d\pi(\theta) = I_\pi(T; X_1^n),$$

where π is a probability measure on Θ, T is distributed according to π, and conditional on $T = \theta$, we draw $X_1^n \sim P_\theta$, and I_π denotes the mutual information when T is drawn according to π. Define the maximin redundancy $C_n^* := \sup_\pi C_n^\pi$ as the worst-case Bayes redundancy. We show that for "most" points θ under the prior π, if $\overline{Q} = \int P_\theta d\pi(\theta)$ is the mixture of all the P_θ under the prior π, then no distribution Q can have subtantially better redundancy that $\overline{Q}$.

Consider any distribution Q on the set $\mathcal{X}$ and let $\epsilon \in [0, 1]$, and define the set of points θ where Q is ϵ-better than the worst case redundancy as

$$B_\epsilon := \left\{\theta \in \Theta : D_{\rm kl}\left(P_\theta \| Q\right) \leq (1 - \epsilon)C_n^*\right\}.$$

(a) Show that for any prior π, we have

$$\pi(B_\epsilon) \leq \frac{\log 2 + C_n^* - I_\pi(T; X_1^n)}{\epsilon C_n^*}.$$

As an aside, note this implies that if π_i is a sequence of priors tending to $\sup_\pi I_\pi(T; X_1^n)$ and the redundancy $C_n^* \to \infty$, then so long as $C_n^* - I_{\pi_i}(T; X_1^n) \ll \epsilon C_n^*$, we have $\pi_i(B_\epsilon) \approx 0$.

(b) Assume that π attains the supremum in the definition of C_n^*. Show that

$$\pi(B_\epsilon) \leq O(1) \cdot \exp(-\epsilon C_n^*).$$

Hint: Introduce the random variable Z to be 1 if the random variable $T \in B_\epsilon$ and 0 otherwise, then use that $Z \to T \to X_1^n$ forms a Markov chain, and expand the mutual information. For part (b), the inequality $\frac{1-x}{x} \log \frac{1}{1-x} \leq 1$ for all $x \in [0, 1]$ may be useful.

Exercise 16.3 (Mixtures are as good as point distributions)**:** Let P be a $\mathsf{Laplace}(\lambda)$ distribution on $\mathbb{R}$, meaning that $X \sim P$ has density

$$p(x) = \frac{\lambda}{2} \exp(-\lambda|x|).$$

Assume that $X_1, \ldots, X_n \stackrel{\rm iid}{\sim} P$, and let P^n denote the n-fold product of P. In this problem, we compare the predictive performance of distributions from the normal location family $\mathcal{P} = \{\mathsf{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$ with the mixture distribution Q^π over $\mathcal{P}$ defined by the normal prior distribution $\mathsf{N}(\mu, \tau^2)$, that is, $\pi(\theta) = (2\pi\tau^2)^{-1/2} \exp(-(\theta - \mu)^2/2\tau^2)$.

(a) Let $P_{\theta,\Sigma}$ be the multivariate normal distribution with mean $\theta \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n\times n}$. What is $D_{\rm kl}\left(P^n \| P_{\theta,\Sigma}\right)$?

(b) Show that $\inf_{\theta\in\mathbb{R}^n} D_{\rm kl}\left(P^n \| P_{\theta,\Sigma}\right) = D_{\rm kl}\left(P^n \| P_{0,\Sigma}\right)$, that is, the mean-zero normal distribution has the smallest KL-divergence from the Laplace distribution.

(c) Let Q_n^π be the mixture of the n-fold products in $\mathcal{P}$, that is, Q_n^π has density

$$q_n^\pi(x_1^n) = \int_{-\infty}^{\infty} \pi(\theta) p_\theta(x_1) \cdots p_\theta(x_n) d\theta,$$

where π is $\mathsf{N}(0, \tau^2)$. What is $D_{\rm kl}\left(P^n \| Q_n^\pi\right)$?

(d) Show that the redundancy of Q_n^π under the distribution P is asymptotically nearly as good as the redundancy of any $P_\theta \in \mathcal{P}$, the normal location family (so P_θ has density $p_\theta(x) = (2\pi\sigma^2)^{-1/2} \exp(-(x-\theta)^2/2\sigma^2)$). That is, show that

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_P \left[\log \frac{1}{q_n^\pi(X_1^n)} - \log \frac{1}{p_\theta(X_1^n)} \right] = \mathcal{O}(\log n)$$

for any prior variance $\tau^2 > 0$ and any prior mean $\mu \in \mathbb{R}$, where the big-Oh hides terms dependent on τ^2, σ^2, μ^2.

(e) **Extra credit:** Can you give an interesting condition under which such redundancy guarantees hold more generally? That is, using Proposition 16.4.3 in the notes, give a general condition under which

$$\mathbb{E}_P \left[\log \frac{1}{q^\pi(X_1^n)} - \log \frac{1}{p_\theta(X_1^n)} \right] = o(n)$$

as $n \to \infty$, for all $\theta \in \Theta$.

Chapter 17

Universal prediction with other losses

Thus far, in our discussion of universal prediction and related ideas, we have focused (essentially) exclusively on making predictions with the logarithmic loss, so that we play a full distribution over the set $\mathcal{X}$ as our prediction at each time step in the procedure. This is natural in settings, such as coding (recall examples 16.1.2 and 16.2.1), in which the log loss corresponds to a quantity we directly care about, or when we do not necessarily know much about the task at hand but rather wish to simply model a process. (We will see this more shortly.) In many cases, however, we have a natural task-specific loss. The natural question that follows, then, is to what extent it is possible to extend the results of Chapter 16 to different settings in which we do not necessarily care about prediction of an entire distribution. (Relevant references include the paper of Cesa-Bianchi and Lugosi [46], which shows how complexity measures known as Rademacher complexity govern the regret in online prediction games; the book by the same authors [47], which gives results covering a wide variety of online learning, prediction, and other games; the survey by Merhav and Feder [140]; and the study of consequences of the choice of loss for universal prediction problems by Haussler et al. [102].)

17.1 Redudancy and expected regret

We begin by considering a generalization of the redundancy (16.2.3) to the case in which we do not use the log loss. In particular, we have as usual a space $\mathcal{X}$ and a loss function $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $\ell(\widehat{x}, x)$ is the penalty we suffer for playing $\widehat{x}$ when the instantaneous data is x. (In somewhat more generality, we may allow the loss to act on $\widehat{\mathcal{X}} \times \mathcal{X}$, where the prediction space $\widehat{\mathcal{X}}$ may be different from $\mathcal{X}$.) As a simple example, consider a weather prediction problem, where $X_i \in \{0, 1\}$ indicates whether it rained on day i and $\widehat{X}_i$ denotes our prediction of whether it will rain. Then a natural loss includes $\ell(\widehat{x}, x) = \mathbf{1}\left\{\widehat{x} \cdot x \le 0\right\}$, which simply counts the number of mistaken predictions.

Given the loss ℓ, our goal is to minimize the expected cumulative loss

$$\sum_{i=1}^{n} \mathbb{E}_P[\ell(\widehat{X}_i, X_i)],$$

where $\widehat{X}_i$ are the predictions of the procedure we use and P is the distribution generating the data X_1^n. In this case, if the distribution P is known, it is clear that the optimal strategy is to play the Bayes-optimal prediction

$$X_i^* \in \operatorname*{argmin}_{x \in \widehat{\mathcal{X}}} \mathbb{E}_P[\ell(x, X_i) \mid X_1^{i-1}] = \operatorname*{argmin}_{x \in \widehat{\mathcal{X}}} \int_{\mathcal{X}} \ell(x, x_i) dP(x_i \mid X_1^{i-1}). \tag{17.1.1}$$

In many cases, however, we do not know the distribution P, and so our goal (as in the previous chapter) is to simultaneously minimize the cumulative loss simultaneously for all source distributions in a family $\mathcal{P}$.

17.1.1 Universal prediction via the log loss

As our first idea, we adapt the same strategies as those in the previous section, using a distribution Q that has redundancy growing only sub-linearly against the class $\mathcal{P}$, and making Bayes optimal predictions with Q. That is, at iteration i, we assume that $X_i \sim Q(\cdot \mid X_1^{i-1})$ and play

$$\widehat{X}_i \in \operatorname*{argmin}_{x \in \widehat{\mathcal{X}}} \mathbb{E}_Q[\ell(x, X_i) \mid X_1^{i-1}] = \int_{\mathcal{X}} \ell(x, x_i) dQ(x_i \mid X_1^{i-1}). \tag{17.1.2}$$

Given such a distribution Q, we measure its loss-based redundancy against P via

$$\mathsf{Red}_n(Q, P, \ell) := \mathbb{E}_P \left[\sum_{i=1}^n \ell(\widehat{X}_i, X_i) - \sum_{i=1}^n \ell(X_i^*, X_i) \right], \tag{17.1.3}$$

where $\widehat{X}_i$ chosen according to $Q(\cdot \mid X_1^{i-1})$ as in expression (17.1.2). The natural question now, of course, is whether the strategy (17.1.2) has redundancy growing more slowly than n.

It turns out that in some situations, this is the case: we have the following theorem [140, Section III.A.2], which only requires that the usual redundancy (16.2.3) (with log loss) is sub-linear and the loss is suitably bounded. In the theorem, we assume that the class of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ is indexed by $\theta \in \Theta$.

Theorem 17.1.1. *Assume that the redundancy* $\mathsf{Red}_n(Q, P_\theta) \le R_n(\theta)$ *and that* $|\ell(\widehat{x}, x) - \ell(x^*, x)| \le L$ *for all* x *and predictions* $\widehat{x}, x^*$. *Then we have*

$$\frac{1}{n} \mathsf{Red}_n(Q, P_\theta, \ell) \le L \sqrt{\frac{2}{n} R_n(\theta)}.$$

To attain vanishing expected regret under the loss ℓ, then, Theorem 17.1.1 requires only that we play a Bayes' strategy (17.1.2) with a distribution Q for which the average (over n) of the usual redundancy (16.2.3) tends to zero, so long as the loss is (roughly) bounded. We give two examples of bounded losses. First, we might consider the 0-1 loss, which clearly satisfies $|\ell(\widehat{x}, x) - \ell(x^*, x)| \le 1$. Second, the absolute value loss (which is used for robust estimation of location parameters [145, 108]), given by $\ell(\widehat{x}, x) = |x - \widehat{x}|$, satisfies $|\ell(\widehat{x}, x) - \ell(x^*, x)| \le |\widehat{x} - x^*|$. If the distribution P_θ has median θ and Θ is compact, then $\mathbb{E}[|\widehat{x} - X|]$ is minimized by its median, and $|\widehat{x} - x^*|$ is bounded by the diameter of Θ.

Proof The theorem is essentially a consequence of Pinsker's inequality (Proposition 2.2.8). By

expanding the loss-based redundancy, we have the following chain of equalities:

$$\begin{aligned}
\mathsf{Red}_n(Q, P_\theta, \ell) &= \sum_{i=1}^n \mathbb{E}_\theta[\ell(\widehat{X}_i, X_i)] - \mathbb{E}_\theta[\ell(X_i^*, X_i)] \\
&= \sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \int_{\mathcal{X}} p_\theta(x_i \mid x_1^{i-1}) \left[\ell(\widehat{X}_i, x_i) - \ell(X_i^*, x_i)\right] dx_i dx_1^{i-1} \\
&= \sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \int_{\mathcal{X}} (p_\theta(x_i \mid x_1^{i-1}) - q(x_i \mid x_1^{i-1})) \left[\ell(\widehat{X}_i, x_i) - \ell(X_i^*, x_i)\right] dx_i dx_1^{i-1} \\
&\quad + \sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \underbrace{\mathbb{E}_Q[\ell(\widehat{X}_i, X_i) - \ell(X_i^*, X_i) \mid x_1^{i-1}]}_{\le 0} dx_1^{i-1}, \qquad (17.1.4)
\end{aligned}$$

where for the inequality we used that the play $\widehat{X}_i$ minimizes

$$\mathbb{E}_Q[\ell(\widehat{X}_i, X_i) - \ell(X_i^*, X_i) \mid X_1^{i-1}]$$

by the construction (17.1.2).

Now, using Hölder's inequality on the innermost integral in the first sum of expression (17.1.4), we have

$$\begin{aligned}
&\int_{\mathcal{X}} (p_\theta(x_i \mid x_1^{i-1}) - q(x_i \mid x_1^{i-1})) \left[\ell(\widehat{X}_i, x_i) - \ell(X_i^*, x_i)\right] dx_i \\
&\qquad \le 2 \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\mathrm{TV}} \sup_{x \in \mathcal{X}} |\ell(\widehat{X}_i, x) - \ell(X_i^*, x)| \\
&\qquad \le 2L \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\mathrm{TV}},
\end{aligned}$$

where we have used the definition of total variation distance. Combining this inequality with (17.1.4), we obtain

$$\begin{aligned}
\mathsf{Red}_n(Q, P_\theta, \ell) &\le 2L \sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\mathrm{TV}} dx_1^{i-1} \\
&\overset{(\star)}{\le} 2L \sum_{i=1}^n \left(\int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) dx_1^{i-1}\right)^{\frac{1}{2}} \left(\int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\mathrm{TV}}^2\right)^{\frac{1}{2}} \\
&= 2L \sum_{i=1}^n \left(\int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\mathrm{TV}}^2\right)^{\frac{1}{2}},
\end{aligned}$$

where the inequality $(\star)$ follows by the Cauchy-Schwarz inequality applied to the integrands $\sqrt{p_\theta}$ and $\sqrt{p_\theta} \left\|P - Q\right\|_{\mathrm{TV}}$. Applying the Cauchy-Schwarz inequality to the final sum, we have

$$\begin{aligned}
\mathsf{Red}_n(Q, P_\theta, \ell) &\le 2L\sqrt{n} \left(\sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\mathrm{TV}}^2\right)^{\frac{1}{2}} \\
&\overset{(\star\star)}{\le} 2L\sqrt{n} \left(\frac{1}{2} \sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) D_{\mathrm{kl}}\left(P_\theta(\cdot \mid x_1^{i-1}) \| Q(\cdot \mid x_1^{i-1})\right) dx_1^{i-1}\right)^{\frac{1}{2}} \\
&= L\sqrt{2n} \sqrt{D_{\mathrm{kl}}\left(P_\theta^n \| Q\right)},
\end{aligned}$$

where inequality $(\star\star)$ is an application of Pinsker's inequality. But of course, we know by that $\mathsf{Red}_n(Q, P_\theta) = D_{\rm kl}\left(P_\theta^n \| Q\right)$ by definition (16.2.3) of the redundancy. $\square$

Before proceding to examples, we note that in a variety of cases the bounds of Theorem 17.1.1 are loose. For example, under mean-squared error, universal linear predictors [58, 151] have redundancy $\mathcal{O}(\log n)$, while Theorem 17.1.1 gives at best a bound of $\mathcal{O}(\sqrt{n})$.

TODO: Add material on redundancy/capacity (Theorem 16.4.5) analogue in general loss case, which allows playing mixture distributions based on mixture of $\{P_\theta\}_{\theta \in \Theta}$.

17.1.2 Examples

We now give an example application of Theorem 17.1.1 with an application to a classification problem with side information. In particular, let us consider the 0-1 loss $\ell_{0\text{-}1}(\hat{y}, y) = \mathbf{1}\left\{\hat{y} \cdot y \le 0\right\}$, and assume that we wish to predict y based on a vector $x \in \mathbb{R}^d$ of regressors that are fixed ahead of time. In addition, we assume that the "true" distribution (or competitor) P_θ is that given x and θ, Y has normal distribution with mean $\langle \theta, x \rangle$ and variance σ^2, that is,

$$Y_i = \langle \theta, x_i \rangle + \varepsilon_i, \quad \varepsilon_i \stackrel{\rm iid}{\sim} \mathsf{N}(0, \sigma^2).$$

Now, we consider playing according to a mixture distribution (16.4.3), and for our prior π we choose $\theta \sim \mathsf{N}(0, \tau^2 I_{d \times d})$, where $\tau > 0$ is some parameter we choose.

Let us first consider the case in which we observe $Y_1, \ldots, Y_n$ directly (rather than simply whether we classify correctly) and consider the prediction scheme this generates. First, we recall as in the posterior calculation (16.4.4) that we must calculate the posterior on θ given $Y_1, \ldots, Y_i$ at step $i+1$. Assuming we have computed this posterior, we play

$$\begin{aligned} \widehat{Y}_i := \operatorname*{argmin}_{y \in \mathbb{R}} \mathbb{E}_{Q^\pi}[\ell_{0\text{-}1}(y, Y_i) \mid Y_1^{i-1}] &= \operatorname*{argmin}_{y \in \mathbb{R}} Q^\pi(\operatorname{sign}(Y_i) \neq \operatorname{sign}(y) \mid Y_1^{i-1}) \\ &= \operatorname*{argmin}_{y \in \mathbb{R}} \int_{-\infty}^{\infty} P_\theta(\operatorname{sign}(Y_i) \neq \operatorname{sign}(y)) \pi(\theta \mid Y_1^{i-1}) d\theta. \end{aligned} \tag{17.1.5}$$

With this in mind, we begin by computing the posterior distribution on θ:

Lemma 17.1.2. *Assume that θ has prior $\mathsf{N}(0, \tau^2 I_{d \times d})$. Then conditional on $Y_1^i = y_1^i$ and the first i vectors $x_1^i = (x_1, \ldots, x_i) \subset \mathbb{R}^d$, we have*

$$\theta \mid y_1^i, x_1^i \sim \mathsf{N}\left(K_i^{-1} \sum_{j=1}^i x_j y_j, K_i^{-1}\right), \quad \text{where} \quad K_i = \frac{1}{\tau^2} I_{d \times d} + \frac{1}{\sigma^2} \sum_{j=1}^i x_j x_j^\top.$$

Deferring the proof of Lemma 17.1.2 temporarily, we note that under the distribution Q^π, as by assumption we have $Y_i = \langle \theta, x_i \rangle + \varepsilon_i$, the posterior distribution (under the prior π for θ) on Y_{i+1} conditional on $Y_1^i = y_i^i$ and $x_1, \ldots, x_{i+1}$ is

$$Y_{i+1} = \langle \theta, x_{i+1} \rangle + \varepsilon_{i+1} \mid y_1^i, x_i^1 \sim \mathsf{N}\left(\left\langle x_{i+1}, K_i^{-1} \sum_{j=1}^i x_j y_j \right\rangle, x_{i+1}^\top K_i^{-1} x_{i+1} + \sigma^2\right).$$

Consequently, if we let $\widehat{\theta}_{i+1}$ be the posterior mean of $\theta \mid y_1^i, x_i^i$ (as given by Lemma 17.1.2), the optimal prediction (17.1.5) is to choose any $\widehat{Y}_{i+1}$ satisfying $\sign(\widehat{Y}_{i+1}) = \sign(\langle x_{i+1}, \widehat{\theta}_{i+1}\rangle)$. Another option is to simply play

$$\widehat{Y}_{i+1} = x_{i+1}^\top K_i^{-1} \left(\sum_{j=1}^i y_j x_j \right), \tag{17.1.6}$$

which is $\mathbb{E}[\widehat{Y}_{i+1} \mid Y_1^i, X_1^{i+1}] = \mathbb{E}[\langle \theta, X_{i+1}\rangle \mid Y_1^i, X_1^i]$, because this $\widehat{Y}_{i+1}$ has sign that is most probable for Y_{i+1} (under the mixture Q^π).

Let us now evaluate the 0-1 redundancy of the prediction scheme (17.1.6). We first compute the Fisher information for the distribution $Y_i \sim \mathsf{N}(\langle \theta, x_i\rangle, \sigma^2)$. By a straightforward calculation, we have $I_\theta = \frac{1}{\sigma^2} X^\top X$, where the matrix $X \in \mathbb{R}^{n \times d}$ is the data matrix $X = [x_1 \; \cdots \; x_n]^\top$. Then for any $\theta_0 \in \mathbb{R}^d$, Theorem 16.4.1 implies that for the prior $\pi(\theta) = \frac{1}{(2\pi\tau^2)^{d/2}} \exp(-\frac{1}{2\tau^2} \|\theta\|_2^2)$, we have (up to constant factors) the redundancy bound

$$\mathsf{Red}_n(Q^\pi, P_{\theta_0}) \lesssim d \log n + d \log \tau + \frac{1}{\tau^2} \|\theta_0\|_2^2 + \log \det(\sigma^{-2} X^\top X).$$

Thus the expected regret under the 0-1 loss $\ell_{0\text{-}1}$ is

$$\mathsf{Red}_n(Q^\pi, P_{\theta_0}, \ell_{0\text{-}1}) \lesssim \sqrt{n} \sqrt{d \log n + d \log(\sigma\tau) + \frac{1}{\tau^2} \|\theta_0\|_2^2 + \log \det(X^\top X)} \tag{17.1.7}$$

by Theorem 17.1.1. We can provide some intuition for this expected regret bound. First, for any θ_0, we can asymptotically attain vanishing expected regret, though larger θ_0 require more information to identify. In addition, the less informative the prior is (by taking $\tau \uparrow +\infty$), the less we suffer by being universal to all θ_0, but there is logarithmic penalty in τ. We also note that the bound (17.1.7) is not strongly universal, because by taking $\|\theta_0\| \to \infty$ we can make the bound vacuous.

We remark in passing that we can play a similar game when all we observe are truncated (signed) normal random variables, that is, we see only $\sign(Y_i)$ rather than Y_i. Unfortunately, in this case, there is no closed form for the posterior updates as in Lemma 17.1.2. That said, it is possible to play the game using sampling (Monte Carlo) or other strategies.

Finally, we prove Lemma 17.1.2:

Proof We use Bayes rule, ignoring normalizing constants that do not depend on θ. In this case, we have the posterior distribution proportional to the prior times the likelihood, so

$$\pi(\theta \mid y_1^i, x_1^i) \propto \pi(\theta) \prod_{i=1}^n p_\theta(y_i \mid x_i) \propto \exp\left(-\frac{1}{2\tau^2} \|\theta\|_2^2 - \frac{1}{2\sigma^2} \sum_{j=1}^i (y_j - \langle x_j, \theta\rangle)^2 \right).$$

Now, we complete the square in the exponent above, which yields

$$\begin{aligned} \frac{1}{2\tau^2} \|\theta\|_2^2 + \frac{1}{2\sigma^2} \sum_{j=1}^i (y_j - \langle x_j, \theta\rangle)^2 &= \frac{1}{2} \theta^\top \left(\frac{1}{\tau^2} I_{d\times d} + \frac{1}{\sigma^2} \sum_{j=1}^i x_j x_j^\top \right) \theta - \theta^\top \sum_{j=1}^i y_j x_j + C \\ &= \frac{1}{2} \left(\theta - K_i^{-1} \sum_{j=1}^i y_j x_j \right)^\top K_i \left(\theta - K_i^{-1} \sum_{j=1}^i y_j x_j \right) + C', \end{aligned}$$

where C, C' are constants depending only on the y_1^i and not x_1^i or θ, and we have recalled the definition of $K_i = \tau^{-2} I_{d\times d} + \sigma^{-2} \sum_{j=1}^i x_j x_j^\top$. By inspection, this implies our desired result. $\square$

17.2 Individual sequence prediction and regret

Having discussed (in some minor detail) prediction games under more general losses in an expected sense, we now consider the more adversarial sense of Section 16.3, where we wish to compete against a family of prediction strategies and the data sequence observed is chosen adversarially. In this section, we look into the case in which the comparison class—set of strategies against which we wish to compete—is finite.

As a first observation, in the redundancy setting, we see that when the class $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ has $|\Theta| < \infty$, then the redundancy capacity theorem (Theorem 16.4.5) implies that

$$\inf_Q \sup_{\theta \in \Theta} \mathsf{Red}_n(Q, P_\theta) = \inf_Q \sup_{\theta \in \Theta} D_{\rm kl}\left(P_\theta^n \| Q\right) = \sup_\pi I_\pi(T; X_1^n) \le \log |\Theta|,$$

where $T \sim \pi$ and conditioned on $T = \theta$ we draw $X_1^n \sim P_\theta$. (Here we have used that $I(T; X_1^n) = H(T) - H(T \mid X_1^n) \le H(T) \le \log |\Theta|$, by definition (2.1.3) of the mutual information.) In particular, the redundancy is *constant* for any n.

Now we come to our question: is this possible in a purely sequential case? More precisely, suppose we wish to predict a sequence of variables $y_i \in \{-1, 1\}$, we have access to a finite collection of strategies, and we would like to guarantee that we perform as well in prediction as any single member of this class. Then, while it is not possible to achieve constant regret, it is possible to have regret that grows only logarithmically in the number of comparison strategies. To establish the setting, let us denote our collection of strategies, henceforth called "experts", by $\{x_{i,j}\}_{j=1}^d$, where i ranges in $1, \ldots, n$. Then at iteration i of the prediction game, we measure the loss of expert j by $\ell(x_{i,j}, y)$.

We begin by considering a mixture strategy that would be natural under the logarithmic loss, we assume the experts play points $x_{i,j} \in [0, 1]$, where $x_{i,j} = P(Y_i = 1)$ according to expert j. (We remark in passing that while the notation is perhaps not completely explicit about this, the experts may adapt to the sequence Y_1^n.) In this case, the loss we suffer is the usual log loss, $\ell(x_{i,j}, y) = y \log \frac{1}{x_{i,j}} + (1 - y) \log \frac{1}{1 - x_{i,j}}$. Now, if we assume we begin with the uniform prior distribution $\pi(j) = 1/d$ for all j, then the posterior distribution, denoted by $\pi_j^i = \pi(j \mid Y_1^{i-1})$, is

$$\begin{aligned}
\pi_j^i \propto \pi(j) \prod_{l=1}^i x_{l,j}^{y_l} (1 - x_{l,j})^{1 - y_l} &= \pi(j) \exp\left(- \sum_{l=1}^i \left[y_l \log \frac{1}{x_{l,j}} + (1 - y_l) \log \frac{1}{1 - x_{l,j}} \right] \right) \\
&= \pi(j) \exp\left(- \sum_{l=1}^i \ell(x_{l,j}, y_l) \right).
\end{aligned}$$

This strategy suggests what is known variously as the *multiplicative weights* strategy [8], exponentiated gradient descent method [119], or (after some massaging) a method known since the late 1970s as the mirror descent or non-Euclidean gradient descent method (entropic gradient descent) [142, 22].

In particular, we consider an algorithm for general losses where fix a stepsize $\eta > 0$ (as we cannot be as aggressive as in the probabilistic setting), and we then weight each of the experts j by exponentially decaying the weight assigned to the expert for the losses it has suffered. For the algorithm to work, unfortunately, we need a technical condition on the loss function and experts $x_{i,j}$. This loss function is analogous to a weakened version of exp-concavity, which is a common assumption in online game playing scenarios (see the logarithmic regret algorithms developed by Hazan et al. [103], as well as earlier work, for example, that by Kivinen and Warmuth [120] studying regression

problems for which the loss is strongly convex in one variable but not simultaneously in all). In particular, exp-concavity is the assumption that

$$x \mapsto \exp(-\ell(x, y))$$

is a concave function. Because the exponent of the logarithm is linear, the log loss is obviously exp-concave, but for alternate losses, we make a slightly weaker assumption. In particular, we assume there are constants c, η such that for any vector π in the d-simplex (i.e. $\pi \in \mathbb{R}_+^d$ satisfies $\sum_{j=1}^d \pi_j = 1$) there is some way to choose $\widehat{y}$ so that for any y (that can be played in the game)

$$\exp\left(-\frac{1}{c}\ell(\widehat{y}, y)\right) \geq \sum_{j=1}^d \pi_j \exp(-\eta \ell(x_{i,j}, y)) \quad \text{or} \quad \ell(\widehat{y}, y) \leq -c \log\left(\sum_{j=1}^d \pi_j \exp(-\eta \ell(x_{i,j}, y))\right). \tag{17.2.1}$$

By inspection, inequality (17.2.1) holds for the log loss with $c = \eta = 1$ and the choice $\widehat{y} = \sum_{j=1}^d \pi_j x_{i,j}$, because of the exp-concavity condition; any exp-concave loss also satisfies inequality (17.2.1) with $c = \eta = 1$ and the choice of the posterior mean $\widehat{y} = \sum_{j=1}^d \pi_j x_{i,j}$. The idea in this case is that losses satisfying inequality (17.2.1) behave enough like the logarithmic loss that a Bayesian updating of the experts works. (Condition (17.2.1) originates with the work of Haussler et al. [102], where they name such losses (c, η)-realizable.)

Example 17.2.1 (Squared error and exp-concavity)**:** Consider the squared error loss $\ell(\widehat{y}, y) = \frac{1}{2}(\widehat{y} - y)^2$, where $\widehat{y}, y \in \mathbb{R}$. We claim that if $x_j \in [0, 1]$ for each j, π is in the simplex, meaning $\sum_j \pi_j = 1$ and $\pi_j \geq 0$, and $y \in [0, 1]$, then the squared error $\pi \mapsto \ell(\langle \pi, x \rangle, y)$ is exp-concave, that is, inequality (17.2.1) holds with $c = \eta = 1$ and $\widehat{y} = \langle \pi, x \rangle$. Indeed, computing the Hessian of the exponent, we have

$$\begin{aligned}\nabla_\pi^2 \exp\left(-\frac{1}{2}(\langle \pi, x \rangle - y)^2\right) &= \nabla_\pi \left[-\exp\left(-\frac{1}{2}(\langle \pi, x \rangle - y)^2\right)(\langle \pi, x \rangle - y)x\right] \\ &= \exp\left(-\frac{1}{2}(\langle \pi, x \rangle - y)^2\right)\left((\langle \pi, x \rangle - y)^2 - 1\right)xx^\top.\end{aligned}$$

Noting that $|\langle \pi, x \rangle - y| \leq 1$ yields that $(\langle \pi, x \rangle - y)^2 - 1 \leq 0$, so we have

$$\nabla_\pi^2 \exp\left(-\frac{1}{2}(\langle \pi, x \rangle - y)^2\right) \preceq 0_{d \times d}$$

under the setting of the example. We thus have exp-concavity as desired. $\diamond$

We can also show that the 0-1 loss satisfies the weakened version of exp-concavity in inequality (17.2.1), but we have to take the constant c to be larger (or η to be smaller).

Example 17.2.2 (Zero-one loss and weak exp-concavity)**:** Now suppose that we use the 0-1 loss, that is, $\ell_{0\text{-}1}(\widehat{y}, y) = \mathbf{1}\{y \cdot \widehat{y} \leq 0\}$. We claim that if we take a weighted majority vote under the distribution π, meaning that we set $\widehat{y} = \sum_{j=1}^d \pi_j \operatorname{sign}(x_j)$ for a vector $x \in \mathbb{R}^d$, then inequality (17.2.1) holds with any c large enough that

$$c^{-1} \leq \log \frac{2}{1 + e^{-\eta}}. \tag{17.2.2}$$

Demonstrating inequality (17.2.2) is, by inspection, equivalent to showing that

$$\ell_{0\text{-}1}(\widehat{y}, y) \le -c \log \left(\sum_{j=1}^{d} \pi_j e^{-\eta \ell_{0\text{-}1}(x_j, y)} \right).$$

If $\widehat{y}$ has the correct sign, meaning that $\sign(\widehat{y}) = \sign(y)$, the result is trivial. If $\sign(\widehat{y})$ is not equal to $\sign(y) \in \{-1, 1\}$, then we know at least (by the weights π_j) half of the values x_j have incorrect sign. Thus

$$\sum_{j=1}^{d} \pi_j e^{-\eta \ell_{0\text{-}1}(x_j, y)} = \sum_{j : x_j y \le 0} \pi_j e^{-\eta} + \sum_{j : x_j y > 0} \pi_j \le \frac{1}{2} e^{-\eta} + \frac{1}{2}.$$

Thus, to attain

$$\ell_{0\text{-}1}(\widehat{y}, y) = 1 \le -c \log \left(\sum_{j=1}^{d} \pi_j e^{-\eta \ell_{0\text{-}1}(x_j, y)} \right)$$

it is sufficient that

$$1 \le -c \log \left(\frac{1 + e^{-\eta}}{2} \right) \le -c \log \left(\sum_{j=1}^{d} \pi_j e^{-\eta \ell_{0\text{-}1}(x_j, y)} \right), \quad \text{or} \quad c^{-1} \le \log \left(\frac{2}{1 + e^{-\eta}} \right).$$

This is our desired claim (17.2.2). $\diamond$

Having given general conditions and our motivation of exponential weighting scheme in the case of the logarithmic loss, we arrive at our algorithm. We simply weight the experts by exponentially decaying the losses they suffer. We begin the procedure by initializing a weight vector $w \in \mathbb{R}^d$ with $w_j = 1$ for $j = 1, \ldots, d$. After this, we repeat the following four steps at each time i, beginning with $i = 1$:

1. Set $w_j^i = \exp\left(-\eta \sum_{l=1}^{i-1} \ell(x_{l,j}, y_l)\right)$
2. Set $W^i = \sum_{j=1}^{d} w_j^i$ and $\pi_j^i = w_j^i / W^i$ for each $j \in \{1, \ldots, d\}$
3. Choose $\widehat{y}_i$ satisfying (17.2.1) for the weighting $\pi = \pi^i$ and expert values $\{x_{i,j}\}_{j=1}^d$
4. Observe y_i and suffer loss $\ell(\widehat{y}_i, y_i)$

With the scheme above, we have the following regret bound.

Theorem 17.2.3 (Haussler et al. [102])**.** *Assume condition (17.2.1) holds and that $\widehat{y}_i$ is chosen by the above scheme. Then for any $j \in \{1, \ldots, d\}$ and any sequence $y_1^n \in \mathbb{R}^n$,*

$$\sum_{i=1}^{n} \ell(\widehat{y}_i, y_i) \le c \log d + c\eta \sum_{i=1}^{n} \ell(x_{i,j}, y_i).$$

Proof This is an argument based on potentials. At each iteration, any loss we suffer implies that the potential W^i must decrease, but it cannot decrease too quickly (as otherwise the individual predictors $x_{i,j}$ would suffer too much loss). Beginning with condition (17.2.1), we observe that

$$\ell(\widehat{y}_i, y_i) \le -c \log \left(\sum_{j=1}^{d} \pi_j^i \exp(-\eta \ell(x_{i,j}, y_i)) \right) = -c \log \left(\frac{W^{i+1}}{W^i} \right)$$

Summing this inequality from $i = 1$ to n and using that $W^1 = d$, we have

$$\sum_{i=1}^n \ell(\widehat{y}_i, y_i) \le -c \log \left(\frac{W^{n+1}}{W^1} \right) = c \log d - c \log \left(\sum_{j=1}^d \exp \left(-\eta \sum_{i=1}^n \ell(x_{i,j}, y_i) \right) \right)$$
$$\le c \log d - c \log \exp \left(-\eta \sum_{i=1}^n \ell(x_{i,j}, y_i) \right),$$

where the inequality uses that $\exp(\cdot)$ is increasing. As $\log \exp(a) = a$, this is the desired result. □

We illustrate the theorem by continuing Example 17.2.2, showing how Theorem 17.2.3 gives a regret guarantee of at most $\sqrt{n \log d}$ for any set of at most d experts and any sequence $y_1^n \in \mathbb{R}^n$ under the zero-one loss.

Example (Example 17.2.2 continued)**:** By substituting the choice $c^{-1} = \log \frac{2}{1+e^{-\eta}}$ into the regret guarantee of Theorem 17.2.3 (which satisfies inequality (17.2.1) by our guarantee (17.2.2) from Example 17.2.2), we obtain

$$\sum_{i=1}^n \ell_{0\text{-}1}(\widehat{y}_i, y_i) - \ell_{0\text{-}1}(x_{i,j}, y_i) \le \frac{\log d}{\log \frac{2}{1+e^{-\eta}}} + \frac{\left(\eta - \log \frac{2}{1+e^{-\eta}} \right) \sum_{i=1}^n \ell_{0\text{-}1}(x_{i,j}, y_i)}{\log \frac{2}{1+e^{-\eta}}}.$$

Now, we make an asymptotic expansion to give the basic flavor of the result (this can be made rigorous, but it is sufficient). First, we note that

$$\log \frac{2}{1+e^{-\eta}} \approx \frac{\eta}{2} - \frac{\eta^2}{8},$$

and substituting this into the previous display, we have regret guarantee

$$\sum_{i=1}^n \ell_{0\text{-}1}(\widehat{y}_i, y_i) - \ell_{0\text{-}1}(x_{i,j}, y_i) \lesssim \frac{\log d}{\eta} + \eta \sum_{i=1}^n \ell_{0\text{-}1}(x_{i,j}, y_i). \tag{17.2.3}$$

By making the choice $\eta \approx \sqrt{\log d / n}$ and noting that $\ell_{0\text{-}1} \le 1$, we obtain

$$\sum_{i=1}^n \ell_{0\text{-}1}(\widehat{y}_i, y_i) - \ell_{0\text{-}1}(x_{i,j}, y_i) \lesssim \sqrt{n \log d}$$

for any collection of experts and any sequence y_1^n. ◇

We make a few remarks on the preceding example to close the chapter. First, ideally we would like to attain adaptive regret guarantees, meaning that the regret scales with the performance of the best predictor in inequality (17.2.3). In particular, we might expect that a good expert would satisfy $\sum_{i=1}^n \ell_{0\text{-}1}(x_{i,j}, y_i) \ll n$, which—if we could choose

$$\eta \approx \left(\frac{\log d}{\sum_{i=1}^n \ell_{0\text{-}1}(x_{i,j^*}, y_i)} \right)^{\frac{1}{2}},$$

where $j^* = \operatorname{argmin}_j \sum_{i=1}^n \ell_{0\text{-}1}(x_{i,j}, y_i)$—then we would attain regret bound

$$\sqrt{\log d \cdot \sum_{i=1}^n \ell_{0\text{-}1}(x_{i,j^*}, y_i)} \ll \sqrt{n \log d}.$$

For results of this form, see, for example, Cesa-Bianchi et al. [48] or the more recent work on mirror descent of Steinhardt and Liang [162].

Secondly, we note that it is actually possible to give a regret bound of the form (17.2.3) without relying on the near exp-concavity condition (17.2.1). In particular, performing mirror descent on the convex losses defined by

$$\pi \mapsto \left| \sum_{j=1}^d \operatorname{sign}(x_{i,j}) \pi_j - \operatorname{sign}(y_i) \right|,$$

which is convex, will give a regret bound of $\sqrt{n \log d}$ for the zero-one loss as well. We leave this exploration to the interested reader.

Chapter 18

Online convex optimization

A related notion to the universal prediction problem with alternate losses is that of *online learning* and *online convex optimization*, where we modify the requirements of Chapter 17 further. In the current setting, we essentially do away with distributional assumptions at all, including prediction with a distribution, and we consider the following two player sequential game: we have a space $\mathcal{W}$ in which we—the learner or first player—can play points $w_1, w_2, \ldots$, while nature plays a sequence of loss functions $\ell_t : \mathcal{W} \to \mathbb{R}$. The goal is to guarantee that the regret

$$\sum_{t=1}^{n} \left[\ell_t(w_t) - \ell_t(w^\star)\right] \tag{18.0.1}$$

grows at most sub-linearly with n, for any $w^\star \in \mathcal{W}$ (often, we desire this guarantee to be uniform). As stated, this goal is too broad, so in this chapter we focus on a few natural restrictions, namely, that the sequence of losses ℓ_t are convex, and $\mathcal{W}$ is a convex subset of $\mathbb{R}^d$. In this setting, the problem (18.0.1) is known as *online convex programming*.

18.1 The problem of online convex optimization

Before proceeding, we provide a few relevant definitions to make our discussion easier; we refer to Appendix B for an overview of convexity and proofs of a variety of useful properties of convex sets and functions. First, we recall that a set $\mathcal{W}$ is *convex* if for all $\lambda \in [0, 1]$ and $w, w' \in \mathcal{W}$, we have

$$\lambda w + (1 - \lambda)w' \in \mathcal{W}.$$

Similarly, a function f is *convex* if

$$f(\lambda w + (1 - \lambda)w') \le \lambda f(w) + (1 - \lambda)f(w')$$

for all $\lambda \in [0, 1]$ and w, w'. The *subgradient set*, or *subdifferential*, of a convex function f at the point w is defined to be

$$\partial f(w) := \{g \in \mathbb{R}^d : f(v) \ge f(w) + \langle g, v - w\rangle \text{ for all } v\},$$

and we say that any vector $g \in \mathbb{R}^d$ satisfying $f(v) \ge f(w) + \langle g, v - w\rangle$ for all v is a *subgradient*. For convex functions, the subdifferential set $\partial f(w)$ is essentially always non-empty for any $w \in \operatorname{dom} f$.[1]

[1] Rigorously, we are guaranteed that $\partial f(w) \neq \emptyset$ at all points w in the relative interior of the domain of f.

We now give several examples of convex functions, losses, and corresponding subgradients. The first two examples are for *classification problems*, in which we receive data points $x \in \mathbb{R}^d$ and wish to predict associated labels $y \in \{-1, 1\}$.

Example 18.1.1 (Support vector machines)**:** In the support vector machine problem, we receive data in pairs $(x_t, y_t) \in \mathbb{R}^d \times \{-1, 1\}$, and the loss function

$$\ell_t(w) = \left[1 - y_t \langle w, x_t \rangle\right]_+ = \max\{1 - y_t \langle w, x_t \rangle, 0\},$$

which is convex because it is the maximum of two linear functions. Moreover, the subgradient set is

$$\partial \ell_t(w) = \begin{cases} -y_t x_t & \text{if } y_t \langle w, x_t \rangle < 1 \\ -\lambda \cdot y_t x_t & \text{for } \lambda \in [0, 1] \text{ if } y_t \langle w, x_t \rangle = 1 \\ 0 & \text{otherwise.} \end{cases}$$

$\diamond$

Example 18.1.2 (Logistic regression)**:** As in the support vector machine, we receive data in pairs $(x_t, y_t) \in \mathbb{R}^d \times \{-1, 1\}$, and the loss function is

$$\ell_t(w) = \log(1 + \exp(-y_t \langle x_t, w \rangle)).$$

To see that this loss is convex, note that if $h(t) = \log(1 + e^t)$, then $h'(t) = \frac{1}{1+e^{-t}}$ and $h''(t) = \frac{e^{-t}}{(1+e^{-t})^2} \geq 0$, and ℓ_t is the composition of a linear transformation with h. In this case,

$$\partial \ell_t(w) = \nabla \ell_t(w) = -\frac{1}{1 + e^{y_t \langle x_t, w \rangle}} y_t x_t.$$

$\diamond$

Example 18.1.3 (Expert prediction and zero-one error)**:** By randomization, it is possible to cast certain non-convex optimization problems as convex. Indeed, let us assume that there are d experts, each of which makes a prediction $x_{t,j}$ (for $j = 1, \ldots, d$) at time t, represented by the vector $x_t \in \mathbb{R}^d$, of a label $y_t \in \{-1, 1\}$. Each also suffers the (non-convex) loss $\ell_{0\text{-}1}(x_{t,j}, y_t) = \mathbf{1}\{x_{t,j} y_t \leq 0\}$. By assigning a weight w_j to each expert $x_{t,j}$ subject to the constraint that $w \succeq 0$ and $\langle w, \mathbf{1} \rangle = 1$, then if we were to randomly choose to predict using expert j with probability w_j, we would suffer expected loss at time t of

$$\ell_t(w) = \sum_{j=1}^d w_j \ell_{0\text{-}1}(x_{t,j}, y_t) = \langle g_t, w \rangle,$$

where we have defined the vector $g_t = [\ell_{0\text{-}1}(x_{t,j}, y_t)]_{j=1}^d \in \{0, 1\}^d$. Notably, the expected zero-one loss is convex (even linear), so that its online minimization falls into the online convex programming framework. $\diamond$

As we see in the sequel, online convex programming approaches are often quite simple, and, in fact, are often provably optimal in a variety of scenarios *outside* of online convex optimization. This motivates our study, and we will see that online convex programming approaches have a number of similarities to our regret minimization approaches in previous chapters on universal coding, regret, and redundancy.

18.2 Online gradient and non-Euclidean gradient (mirror) descent

We now turn to an investigation of the single approach we will use to solve online convex optimization problems, which is known as *mirror descent*.[2] Before describing the algorithm in its full generality, however, we first demonstrate a special case (though our analysis will be for the general algorithm).

Roughly, the intuition for our procedures is as follows: after observing a loss ℓ_t, we make a small update to move our estimate w_t in a direction to improve the value of the losses we have seen. However, so that we do not make progress too quickly—or too aggressively follow spurious information—we attempt to keep new iterates close to previous iterates. With that in mind, we present *(projected) online gradient descent*, which requires only that we specify a sequence η_t of non-increasing stepsizes.

Input: Parameter space $\mathcal{W}$, stepsize sequence η_t.
Repeat: for each iteration t, predict $w_t \in \mathcal{W}$, receive function ℓ_t and suffer loss $\ell_t(w_t)$. Compute any $g_t \in \partial \ell_t(w_t)$, and perform subgradient update

$$w_{t+\frac{1}{2}} = w_t - \eta_t g_t, \quad w_{t+1} = \mathrm{Proj}_{\mathcal{W}}(w_{t+\frac{1}{2}}), \tag{18.2.1}$$

where $\mathrm{Proj}_{\mathcal{W}}$ denotes (Euclidean) projection onto $\mathcal{W}$.

Figure 18.1: Online projected gradient descent.

An equivalent formulation of the update (18.2.1) is to write it as the single step

$$w_{t+1} = \operatorname*{argmin}_{w \in \mathcal{W}} \left\{ \langle g_t, w \rangle + \frac{1}{2\eta_t} \|w - w_t\|_2^2 \right\}, \tag{18.2.2}$$

which makes clear that we trade between improving performance on ℓ_t via the linear approximation of $\ell_t(w) \approx \ell_t(w_t) + g_t^\top (w - w_t)$ and remaining close to w_t according to the Euclidean distance $\|\cdot\|_2$. In a variety of scenarios, however, it is quite advantageous to measure distances in a way more amenable to the problem structure, for example, if $\mathcal{W}$ is a probability simplex or we have prior information about the loss functions ℓ_t that nature may choose. With this in mind, we present a slightly more general algorithm, which requires us to give a few more definitions.

Given a convex differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$, we define the *Bregman divergence* associated with ψ by

$$D_\psi(w, v) = \psi(w) - \psi(v) - \langle \nabla \psi(v), w - v \rangle. \tag{18.2.3}$$

The Bregman divergence is always non-negative, as $D_\psi(w, v)$ is the gap between the true function value $\psi(w)$ and its linear approximation at the point v (see Figure 18.2). A few examples illustrate its properties.

Example 18.2.1 (Euclidean distance as Bregman divergence)**:** Take $\psi(w) = \frac{1}{2}\|w\|_2^2$ to obtain $D(w, v) = \frac{1}{2}\|w - v\|_2^2$. More generally, if for a matrix A we define $\|w\|_A^2 = w^\top A w$, then takin $\psi(w) = \frac{1}{2} w^\top A w$, we have

$$D_\psi(w, v) = \frac{1}{2}(w - v)^\top A (w - v) = \frac{1}{2}\|w - v\|_A^2 .$$

So Bregman divergences generalize (squared) Euclidean distance. $\diamond$

[2] The reasons for this name are somewhat convoluted, and we do not dwell on them.

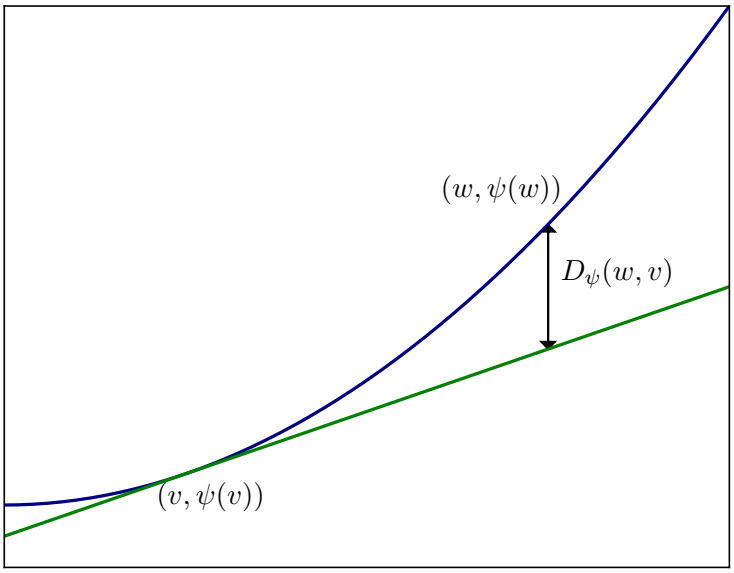

Figure 18.2: Illustration of Bregman divergence.

Example 18.2.2 (KL divergence as a Bregman divergence)**:** Take $\psi(w) = \sum_{j=1}^d w_j \log w_j$. Then ψ is convex over the positive orthant $\mathbb{R}_+^d$ (the second derivative of $w \log w$ is $1/w$), and for $w, v \in \Delta_d = \{u \in \mathbb{R}_+^d : \langle \mathbf{1}, u \rangle = 1\}$, we have

$$D_\psi(w, v) = \sum_j w_j \log w_j - \sum_j v_j \log v_j - \sum_j (1 + \log v_j)(w_j - v_j) = \sum_j w_j \log \frac{w_j}{v_j} = D_{\rm kl}\left(w \| v\right),$$

where in the final equality we treat w and v as probability distributions on $\{1, \ldots, d\}$. $\diamond$

With these examples in mind, we now present the mirror descent algorithm, which is the natural generalization of online gradient descent.

Input: proximal function ψ, parameter space $\mathcal{W}$, and non-increasing stepsize sequence $\eta_1, \eta_2, \ldots$.
Repeat: for each iteration t, predict $w_t \in \mathcal{W}$, receive function ℓ_t and suffer loss $\ell_t(w_t)$. Compute any $g_t \in \partial \ell_t(w_t)$, and perform non-Euclidean subgradient update

$$w_{t+1} = \operatorname*{argmin}_{w \in \mathcal{W}} \left\{ \langle g_t, w \rangle + \frac{1}{\eta_t} D_\psi(w, w_t) \right\}. \tag{18.2.4}$$

Figure 18.3: The online mirror descent algorithm

Before providing the analysis of Algorithm 18.3, we give a few examples of its implementation. First, by taking $\mathcal{W} = \mathbb{R}^d$ and $\psi(w) = \frac{1}{2}\|w\|_2^2$, we note that the mirror descent procedure simply corresponds to the gradient update $w_{t+1} = w_t - \eta_t g_t$. We can also recover the *exponentiated gradient* algorithm, also known as entropic mirror descent.

Example 18.2.3 (Exponentiated gradient algorithm)**:** Suppose that we have $\mathcal{W} = \Delta_d = \{w \in \mathbb{R}_+^d : \langle \mathbf{1}, w \rangle = 1\}$, the probability simplex in $\mathbb{R}^d$. Then a natural choice for ψ is the negative entropy, $\psi(w) = \sum_j w_j \log w_j$, which (as noted previously) gives $D_\psi(w, v) = \sum_j w_j \log \frac{w_j}{v_j}$. We now consider the update step (18.2.4). In this case, fixing $v = w_t$ for notational simplicity, we must solve

$$\text{minimize } \langle g, w \rangle + \frac{1}{\eta} \sum_j w_j \log \frac{w_j}{v_j} \text{ subject to } w \in \Delta_d$$

in w. Writing the Lagrangian for this problem after introducing multipliers $\tau \in \mathbb{R}$ for the contraint that $\langle \mathbf{1}, w \rangle = 1$ and $\lambda \in \mathbb{R}_+^d$ for $w \succeq 0$, we have

$$\mathcal{L}(w, \lambda, \tau) = \langle g, w \rangle + \frac{1}{\eta} \sum_{j=1}^d w_j \log \frac{w_j}{v_j} - \langle \lambda, w \rangle + \tau(\langle \mathbf{1}, w \rangle - 1),$$

which is minimized by taking

$$w_j = v_j \exp(-\eta g_j + \lambda_j \eta - \tau \eta - 1),$$

and as $w_j > 0$ certainly, the constraint $w \succeq 0$ is inactive and $\lambda_j = 0$. Thus, choosing τ to normalize the w_j, we obtain the *exponentiated gradient update*

$$w_{t+1,i} = \frac{w_{t,i} e^{-\eta_t g_{t,i}}}{\sum_j w_{t,j} e^{-\eta_t g_{t,j}}} \text{ for } i = 1, \ldots, d,$$

as the explicit calculation of the mirror descent update (18.2.4). $\diamond$

We now turn to an analysis of the mirror descent algorithm. Before presenting the analysis, we require two more definitions that allow us to relate Bregman divergences to various norms.

Definition 18.1. *Let $\|\cdot\|$ be a norm. The* dual norm $\|\cdot\|_*$ *associated with $\|\cdot\|$ is*

$$\|y\|_* := \sup_{x : \|x\| \le 1} x^\top y.$$

For example, a straightforward calculation shows that the dual to the ℓ_∞-norm is the ℓ_1-norm, and the Euclidean norm $\|\cdot\|_2$ is self-dual (by the Cauchy-Schwarz inequality). Lastly, we require a definition of functions of suitable curvature for use in mirror descent methods.

Definition 18.2. *A convex function $f : \mathbb{R}^d \to \mathbb{R}$ is* strongly convex with respect to the norm $\|\cdot\|$ over the set $\mathcal{W}$ *if for all $w, v \in \mathcal{W}$ and $g \in \partial f(w)$ we have*

$$f(v) \ge f(w) + \langle g, v - w \rangle + \frac{1}{2} \|w - v\|^2.$$

That is, the function f is strongly convex if it grows at least quadratically fast at every point in its domain. It is immediate from the definition of the Bregman divergence that ψ is strongly convex if and only if

$$D_\psi(w, v) \ge \frac{1}{2} \|w - v\|^2.$$

As two examples, we consider Euclidean distance and entropy. For the Euclidean distance, which uses $\psi(w) = \frac{1}{2} \|w\|_2^2$, we have $\nabla \psi(w) = w$, and

$$\frac{1}{2} \|v\|_2^2 = \frac{1}{2} \|w + v - w\|_2^2 = \frac{1}{2} \|w\|_2^2 + \langle w, v - w \rangle + \frac{1}{2} \|w - v\|_2^2$$

by a calculation, so that ψ is strongly convex with respect to the Euclidean norm. We also have the following observation.

Observation 18.2.4. *Let $\psi(w) = \sum_j w_j \log w_j$ be the negative entropy. Then ψ is strongly convex with respect to the ℓ_1-norm, that is,*

$$D_\psi(w, v) = D_{\rm kl}\left(w \| v\right) \geq \frac{1}{2} \left\|w - v\right\|_1^2.$$

Proof The result is an immediate consequence of Pinsker's inequality, Proposition 2.2.8. □

With these examples in place, we present the main theorem of this section.

Theorem 18.2.5 (Regret of mirror descent)**.** *Let ℓ_t be an arbitrary sequence of convex functions, and let w_t be generated according to the mirror descent algorithm 18.3. Assume that the proximal function ψ is strongly convex with respect to the norm $\|\cdot\|$, which has dual norm $\|\cdot\|_*$. Then*

(a) If $\eta_t = \eta$ for all t, then for any $w^\star \in \mathcal{W}$,

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \leq \frac{1}{\eta} D_\psi(w^\star, w_1) + \frac{\eta}{2} \sum_{t=1}^n \left\|g_t\right\|_*^2.$$

(b) If $\mathcal{W}$ is compact and $D_\psi(w^\star, w) \leq R^2$ for any $w \in \mathcal{W}$, then

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \leq \frac{1}{2\eta_n} R^2 + \sum_{t=1}^n \frac{\eta_t}{2} \left\|g_t\right\|_*^2.$$

Before proving the theorem, we provide a few comments to exhibit its power. First, we consider the Euclidean case, where $\psi(w) = \frac{1}{2}\left\|w\right\|_2^2$, and we assume that the loss functions ℓ_t are all L-Lipschitz, meaning that $|\ell_t(w) - \ell_t(v)| \leq L \left\|w - v\right\|_2$, which is equivalent to $\left\|g_t\right\|_2 \leq L$ for all $g_t \in \partial \ell_t(w)$. In this case, the two regret bounds above become

$$\frac{1}{2\eta} \left\|w^\star - w_1\right\|_2^2 + \frac{\eta}{2} n L^2 \quad \text{and} \quad \frac{1}{2\eta_n} R^2 + \sum_{t=1}^n \frac{\eta_t}{2} L^2,$$

respectively, where in the second case we assumed that $\left\|w^\star - w_t\right\|_2 \leq R$ for all t. In the former case, we take $\eta = \frac{R}{L\sqrt{n}}$, while in the second, we take $\eta_t = \frac{R}{L\sqrt{t}}$, which does not require knowledge of n ahead of time. Focusing on the latter case, we have the following corollary.

Corollary 18.2.6. *Assume that $\mathcal{W} \subset \{w \in \mathbb{R}^d : \left\|w\right\|_2 \leq R\}$ and that the loss functions ℓ_t are L-Lipschitz with respect to the Euclidean norm. Take $\eta_t = \frac{R}{L\sqrt{t}}$. Then for all $w^\star \in \mathcal{W}$,*

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \leq 3RL\sqrt{n}.$$

Proof For any $w, w^\star \in \mathcal{W}$, we have $\left\|w - w^\star\right\|_2 \leq 2R$, so that $D_\psi(w^\star, w) \leq 4R^2$. Using that

$$\sum_{t=1}^n t^{-\frac{1}{2}} \leq \int_0^n t^{-\frac{1}{2}} dt = 2\sqrt{n}$$

gives the result. □

Now that we have presented the Euclidean variant of online convex optimization, we turn to an example that achieves better performance in high dimensional settings, as long as the domain is the probability simplex. (Recall Example 18.1.3 for motivation.) In this case, we have the following corollary to Theorem 18.2.5.

Corollary 18.2.7. *Assume that $\mathcal{W} = \Delta_d = \{w \in \mathbb{R}^d_+ : \langle \mathbf{1}, w \rangle = 1\}$ and take the proximal function $\psi(w) = \sum_j w_j \log w_j$ to be the negative entropy in the mirror descent procedure 18.3. Then with the fixed stepsize η and initial point as the uniform distribution $w_1 = \mathbf{1}/d$, we have for any sequence of convex losses ℓ_t*

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \|g_t\|_\infty^2 .$$

Proof Using Pinsker's inequality in the form of Observation 18.2.4, we have that ψ is strongly convex with respect to $\|\cdot\|_1$. Consequently, taking the dual norm to be the ℓ_∞-norm, part (a) of Theorem 18.2.5 shows that

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{1}{\eta} \sum_{j=1}^d w_j^\star \log \frac{w_j^\star}{w_{1,j}} + \frac{\eta}{2} \sum_{t=1}^n \|g_t\|_\infty^2 .$$

Noting that with $w_1 = \mathbf{1}/d$, we have $D_\psi(w^\star, w_1) \le \log d$ for any $w^\star \in \mathcal{W}$ gives the result. □

Corollary 18.2.7 yields somewhat sharper results than Corollary 18.2.6, though in the restricted setting that $\mathcal{W}$ is the probability simplex in $\mathbb{R}^d$. Indeed, let us assume that the subgradients $g_t \in [-1, 1]^d$, the hypercube in $\mathbb{R}^d$. In this case, the tightest possible bound on their ℓ_2-norm is $\|g_t\|_2 \le \sqrt{d}$, while $\|g_t\|_\infty \le 1$ always. Similarly, if $\mathcal{W} = \Delta_d$, then while we are only guaranteed that $\|w^\star - w_1\|_2 \le 1$. Thus, the best regret guaranteed by the Euclidean case (Corollary 18.2.6) is

$$\frac{1}{2\eta} \|w^\star - w_1\|_2^2 + \frac{\eta}{2} nd \le \sqrt{nd} \quad \text{with the choice } \eta = \frac{1}{\sqrt{nd}},$$

while the entropic mirror descent procedure (Alg. 18.3 with $\psi(w) = \sum_j w_j \log w_j$) guarantees

$$\frac{\log d}{\eta} + \frac{\eta}{2} n \le \sqrt{2n \log d} \quad \text{with the choice } \eta = \frac{\sqrt{2 \log d}}{2\sqrt{n}}. \tag{18.2.5}$$

The latter guarantee is *exponentially* better in the dimension. Moreover, the key insight is that we essentially maintain a "prior," and then perform "Bayesian"-like updating of the posterior distribution w_t at each time step, exactly as in the setting of redundancy minimization.

18.2.1 Proof of Theorem 18.2.5

The proof of the theorem proceeds in three lemmas, which are essentially inductive applications of optimality conditions for convex optimization problems. The first is the explicit characterization of optimality for a convex optimization problem. (For a proof of this lemma, see, for example, the books of Hiriart-Urruty and Lemaréchal [104, 105], or Section 2.5 of Boyd et al. [36].)

Lemma 18.2.8. *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a convex function and $\mathcal{W}$ be a convex set. Then $w^\star$ minimizes $h(w)$ over $\mathcal{W}$ if and only if there exists $g \in \partial h(w^\star)$ such that*

$$\langle g, w - w^\star \rangle \ge 0 \quad \text{for all } w \in \mathcal{W}.$$

Lemma 18.2.9. *Let $\ell_t : \mathcal{W} \to \mathbb{R}$ be any sequence of convex loss functions and η_t be a non-increasing sequence, where $\eta_0 = \infty$. Then with the mirror descent strategy (18.2.4), for any $w^\star \in \mathcal{W}$ we have*

$$\sum_{t=1}^n \ell_t(w_t) - \ell_t(w^\star) \le \sum_{t=1}^n \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) D_\psi(w^\star, w_t) + \sum_{t=1}^n \left[-\frac{1}{\eta_t} D_\psi(w_{t+1}, w_t) + \langle g_t, w_t - w_{t+1}\rangle\right].$$

Proof Our proof follows by the application of a few key identities. First, we note that by convexity, we have for any $g_t \in \partial \ell_t(w_t)$ that

$$\ell_t(w_t) - \ell_t(w^\star) \le \langle g_t, w_t - w^\star\rangle. \tag{18.2.6}$$

Secondly, we have that because w_{t+1} minimizes

$$\langle g_t, w\rangle + \frac{1}{\eta_t} D_\psi(w, w_t)$$

over $w \in \mathcal{W}$, then Lemma 18.2.8 implies

$$\langle \eta_t g_t + \nabla\psi(w_{t+1}) - \nabla\psi(w_t), w - w_{t+1}\rangle \ge 0 \quad \text{for all } w \in \mathcal{W}. \tag{18.2.7}$$

Taking $w = w^\star$ in inequality (18.2.7) and making a substitution in inequality (18.2.6), we have

$$\begin{aligned}
\ell_t(w_t) - \ell_t(w^\star) &\le \langle g_t, w_t - w^\star\rangle = \langle g_t, w_{t+1} - w^\star\rangle + \langle g_t, w_t - w_{t+1}\rangle \\
&\le \frac{1}{\eta_t}\langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t), w^\star - w_{t+1}\rangle + \langle g_t, w_t - w_{t+1}\rangle \\
&= \frac{1}{\eta_t}\left[D_\psi(w^\star, w_t) - D_\psi(w^\star, w_{t+1}) - D_\psi(w_{t+1}, w_t)\right] + \langle g_t, w_t - w_{t+1}\rangle
\end{aligned} \tag{18.2.8}$$

where the final equality (18.2.8) follows from algebraic manipulations of $D_\psi(w, w')$. Summing inequality (18.2.8) gives

$$\begin{aligned}
\sum_{t=1}^n \ell_t(w_t) - \ell_t(w^\star) &\le \sum_{t=1}^n \frac{1}{\eta_t}\left[D_\psi(w^\star, w_t) - D_\psi(w^\star, w_{t+1}) - D_\psi(w_{t+1}, w_t)\right] + \sum_{t=1}^n \langle g_t, w_t - w_{t+1}\rangle \\
&= \sum_{t=2}^n \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) D_\psi(w^\star, w_t) + \frac{1}{\eta_1} D_\psi(w^\star, w_1) - \frac{1}{\eta_n} D_\psi(w^\star, w_{n+1}) \\
&\quad + \sum_{t=1}^n \left[-\frac{1}{\eta_t} D_\psi(w_{t+1}, w_t) + \langle g_t, w_t - w_{t+1}\rangle\right]
\end{aligned}$$

as desired. □

It remains to use the negative terms $-D_\psi(w_t, w_{t+1})$ to cancel the gradient terms $\langle g_t, w_t - w_{t+1}\rangle$. To that end, we recall Definition 18.1 of the dual norm $\|\cdot\|_*$ and the strong convexity assumption on ψ. Using the Fenchel-Young inequality, we have

$$\langle g_t, w_t - w_{t+1}\rangle \le \|g_t\|_* \|w_t - w_{t+1}\| \le \frac{\eta_t}{2}\|g_t\|_*^2 + \frac{1}{2\eta_t}\|w_t - w_{t+1}\|^2.$$

Now, we use the strong convexity condition, which gives

$$-\frac{1}{\eta_t} D_\psi(w_{t+1}, w_t) \le -\frac{1}{2\eta_t}\|w_t - w_{t+1}\|^2.$$

Combining the preceding two displays in Lemma 18.2.9 gives the result of Theorem 18.2.5.

18.3 Online to batch conversions

Martingales!

18.4 More refined convergence guarantees

It is sometimes possible to give more refined bounds than those we have so far provided. As motivation, let us revisit Example 18.1.3, but suppose that one of the experts has no loss—that is, it makes perfect predictions. We might expect—accurately!—that we should attain better convergence guarantees using exponentiated weights, as the points w_t be maintain should quickly eliminate non-optimal experts.

To that end, we present a refined regret bound for the mirror descent algorithm 18.3 with the entropic regularization $\psi(w) = \sum_j w_j \log w_j$.

Proposition 18.4.1. *Let $\psi(w) = \sum_j w_j \log w_j$, and assume that the losses ℓ_t are such that their subgradients have all non-negative entries, that is, $g_t \in \partial \ell_t(w)$ implies $g_t \succeq 0$. For any such sequence of loss functions ℓ_t and any $w^\star \in \mathcal{W} = \Delta_d$,*

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{j=1}^d w_{t,j} g_{t,j}^2.$$

While as stated, the bound of the proposition does not look substantially more powerful than Corollary 18.2.7, but a few remarks will exhibit its consequences. We prove the proposition in Section 18.4.1 to come.

First, we note that because $w_t \in \Delta_d$, we will *always* have $\sum_j w_{t,j} g_{t,j}^2 \le \|g_t\|_\infty^2$. So certainly the bound of Proposition 18.4.1 is never worse than that of Corollary 18.2.7. Sometimes this can be made tighter, however, as exhibited by the next corollary, which applies (for example) to the experts setting of Example 18.1.3. More specifically, we have d experts, each suffering losses in $[0, 1]$, and we seek to predict with the best of the d experts.

Corollary 18.4.2. *Consider the linear online convex optimization setting, that is, where $\ell_t(w_t) = \langle g_t, w_t \rangle$ for vectors g_t, and assume that $g_t \in \mathbb{R}^d_+$ with $\|g_t\|_\infty \le 1$. In addition, assume that we know an upper bound $L_n^\star$ on $\sum_{t=1}^n \ell_t(w^\star)$. Then taking the stepsize $\eta = \min\{1, \sqrt{\log d}/\sqrt{L_n^\star}\}$, we have*

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le 3 \max\left\{\log d, \sqrt{L_n^\star \log d}\right\}.$$

Note that when $\ell_t(w^\star) = 0$ for all $w^\star$, which corresponds to a perfect expert in Example 18.1.3, the upper bound becomes constant in n, yielding $3 \log d$ as a bound on the regret. Unfortunately, in our bound of Corollary 18.4.2, we had to assume that we *knew* ahead of time a bound on the loss of the best predictor $w^\star$, which is unrealistic in practice. There are a number of techniques for dealing with such issues, including a standard one in the online learning literature known as the *doubling* trick. We explore some in the exercises.

Proof First, we note that $\sum_j w_j g_{t,j}^2 \le \langle w, g_t \rangle$ for any nonnegative vector w, as $g_{t,j} \in [0, 1]$. Thus, Proposition 18.4.1 gives

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \langle w_t, g_t \rangle = \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \ell_t(w_t).$$

Rearranging via an algebraic manipulation, this is equivalent to

$$\left(1 - \frac{\eta}{2}\right) \sum_{t=1}^{n} [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \ell_t(w^\star).$$

Take $\eta = \min\{1, \sqrt{\log d / L_n^\star}\}$. Then if $\sqrt{\log d / L_n^\star} \le 1$, we have that the right hand side of the above inequality becomes $\sqrt{L_n^\star \log d} + \frac{1}{2}\sqrt{L_n^\star \log d}$. On the other hand, if $L_n^\star < \log d$, then the right hand side of the inequality becomes $\log d + \frac{1}{2} L_n^\star \le \frac{3}{2} \log d$. In either case, we obtain the desired result by noting that $1 - \frac{\eta}{2} \ge \frac{1}{2}$. □

18.4.1 Proof of Proposition 18.4.1

Our proof relies on a technical lemma, after which the derivation is a straightforward consequence of Lemma 18.2.9. We first state the technical lemma, which applies to the update that the exponentiated gradient procedure makes.

Lemma 18.4.3. *Let $\psi(x) = \sum_j x_j \log x_j$, and let $x, y \in \Delta_d$ be defined by*

$$y_i = \frac{x_i \exp(-\eta g_i)}{\sum_j x_j \exp(-\eta g_j)},$$

where $g \in \mathbb{R}_+^d$ is non-negative. Then

$$-\frac{1}{\eta} D_\psi(y, x) + \langle g, x - y \rangle \le \frac{\eta}{2} \sum_{i=1}^{d} g_i^2 x_i.$$

Deferring the proof of the lemma, we note that it precisely applies to the setting of Lemma 18.2.9. Indeed, with a fixed stepsize η, we have

$$\sum_{t=1}^{n} \ell_t(w_t) - \ell_t(w^\star) \le \frac{1}{\eta} D_\psi(w^\star, w_1) + \sum_{t=1}^{n} \left[-\frac{1}{\eta} D_\psi(w_{t+1}, w_t) + \langle g_t, w_t - w_{t+1} \rangle \right].$$

Earlier, we used the strong convexity of ψ to eliminate the gradient terms $\langle g_t, w_t - w_{t+1} \rangle$ using the bregman divergence D_ψ. This time, we use Lemma 18.2.9: setting $y = w_{t+1}$ and $x = w_t$ yields the bound

$$\sum_{t=1}^{n} \ell_t(w_t) - \ell_t(w^\star) \le \frac{1}{\eta} D_\psi(w^\star, w_1) + \sum_{t=1}^{n} \frac{\eta}{2} \sum_{i=1}^{d} g_{t,i}^2 w_{t,i}$$

as desired.

Proof of Lemma 18.4.3 We begin by noting that a direct calculation yields $D_\psi(y, x) = D_{\rm kl}(y \| x) = \sum_i y_i \log \frac{y_i}{x_i}$. Substituting the values for x and y into this expression, we have

$$\sum_i y_i \log \frac{y_i}{x_i} = \sum_i y_i \log \left(\frac{x_i \exp(-\eta g_i)}{x_i (\sum_j \exp(-\eta g_j) x_j)} \right) = -\eta \langle g, y \rangle - \sum_i y_i \log \left(\sum_j x_j e^{-\eta g_j} \right).$$

Now we use a Taylor expansion of the function $g \mapsto \log(\sum_j x_j e^{-\eta g_j})$ around the point 0. If we define the vector $p(g)$ by $p_i(g) = x_i e^{-\eta g_i}/(\sum_j x_j e^{-\eta g_j})$, then

$$\log\Big(\sum_j x_j e^{-\eta g_j}\Big) = \log(\langle \mathbf{1}, x\rangle) - \eta\langle p(0), g\rangle + \frac{\eta^2}{2} g^\top(\mathrm{diag}(p(\widetilde{g})) - p(\widetilde{g})p(\widetilde{g})^\top)g,$$

where $\widetilde{g} = \lambda g$ for some $\lambda \in [0,1]$. Noting that $p(0) = x$ and $\langle \mathbf{1}, x\rangle = \langle \mathbf{1}, y\rangle = 1$, we obtain

$$D_\psi(y,x) = -\eta\langle g, y\rangle + \log(1) + \eta\langle g, x\rangle - \frac{\eta^2}{2} g^\top(\mathrm{diag}(p(\widetilde{g})) - p(\widetilde{g})p(\widetilde{g})^\top)g,$$

whence

$$-\frac{1}{\eta} D_\psi(y,x) + \langle g, x - y\rangle \le \frac{\eta}{2}\sum_{i=1}^d g_i^2 p_i(\widetilde{g}). \tag{18.4.1}$$

Lastly, we claim that the function

$$s(\lambda) = \sum_{i=1}^d g_i^2 \frac{x_i e^{-\lambda g_i}}{\sum_j x_j e^{-\lambda g_j}}$$

is non-increasing on $\lambda \in [0,1]$. Indeed, we have

$$s'(\lambda) = \frac{(\sum_i g_i x_i e^{-\lambda g_i})(\sum_i g_i^2 x_i e^{-\lambda g_i})}{(\sum_i x_i e^{-\lambda g_i})^2} - \frac{\sum_i g_i^3 x_i e^{-\lambda g_i}}{\sum_i x_i e^{-\lambda g_i}} = \frac{\sum_{ij} g_i g_j^2 x_i x_j e^{-\lambda g_i - \lambda g_j} - \sum_{ij} g_i^3 x_i x_j e^{-\lambda g_i - \lambda g_j}}{(\sum_i x_i e^{-\lambda g_i})^2}.$$

Using the Fenchel-Young inequality, we have $ab \le \frac{1}{3}|a|^3 + \frac{2}{3}|b|^{3/2}$ for any a, b, so $g_i g_j^2 \le \frac{1}{3} g_i^3 + \frac{2}{3} g_j^3$. This implies that the numerator in our expression for $s'(\lambda)$ is non-positive. Thus we have $s(\lambda) \le s(0) = \sum_{i=1}^d g_i^2 x_i$, which gives the result when combined with inequality (18.4.1). $\square$

Chapter 19

Exploration, exploitation, and bandit problems

Consider the following problem: we have a possible treatment for a population with a disease, but we do not know whether the treatment will have a positive effect or not. We wish to evaluate the treatment to decide whether it is better to apply it or not, and we wish to optimally allocate our resources to attain the best outcome possible. There are challenges here, however, because for each patient, we may only observe the patient's behavior and disease status in one of two possible states—under treatment or under control—and we wish to allocate as few patients to the group with worse outcomes (be they control or treatment) as possible. This balancing act between exploration—observing the effects of treatment or non-treatment—and exploitation—giving treatment or not as we decide which has better palliative outcomes—underpins and is the paridigmatic aspect of the multi-armed bandit problem.[1]

Our main focus in this chapter is a fairly simple variant of the K-armed bandit problem, though we note that there is a substantial literature in statistics, operations research, economics, game theory, and computer science on variants of the problems we consider. In particular, we consider the following sequential decision making scenario. We assume that there are K distributions $P_1, \ldots, P_K$ on $\mathbb{R}$, which we identify (with no loss of generality) with K random variables $Y_1, \ldots, Y_K$. Each random variable Y_i has mean μ_i and is σ^2-sub-Gaussian, meaning that

$$\mathbb{E}\left[\exp\left(\lambda(Y_i - \mu_i)\right)\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right). \tag{19.0.1}$$

The goal is to find the index i with the maximal mean μ_i without evaluating sub-optimal "arms" (or random variables Y_i) too often. At each iteration t of the process, the player takes an action $A_t \in \{1, \ldots, K\}$, then, conditional on $i = A_t$, observes a reward $Y_i(t)$ drawn independently from the distribution P_i. Then the goal is to minimize the the regret after n steps, which is

$$\mathsf{Reg}_n := \sum_{t=1}^{n} \mu_{i^\star} - \mu_{A_t}, \tag{19.0.2}$$

[1]The problem is called the bandit problem in the literature because we imagine a player in a casino, choosing between K different slot machines (hence a K-armed bandit, as this is a casino and the player will surely lose eventually), each with a different unknown reward distribution. The player wishes to put as much of his money as possible into the machine with the greatest expected reward.

where $i^\star \in \operatorname{argmax}_i \mu_i$ so $\mu_{i^\star} = \max_i \mu_i$. The regret Reg_n as defined is a random quantity, so we generally seek to give bounds on its expectation or high-probability guarantees on its value. In this chapter, we generally focus for simplicity on the expected regret,

$$\overline{\mathsf{Reg}}_n := \mathbb{E}\left[\sum_{t=1}^n \mu_{i^\star} - \mu_{A_t}\right], \tag{19.0.3}$$

where the expectation is taken over any randomness in the player's actions A_t and in the repeated observations of the random variables $Y_1, \ldots, Y_K$.

19.1 Confidence-based algorithms

A natural first strategy to consider is one based on confidence intervals with slight optimism. Roughly, if we believe the true mean μ_i for an arm i lies within $[\widehat{\mu}_i - c_i, \widehat{\mu}_i + c_i]$, where c_i is some interval (whose length decreases with time t), then we optimistically "believe" that the value of arm i is $\widehat{\mu}_i + c_i$; then at iteration t, as our action A_t we choose the arm whose optimistic mean is the highest, thus hoping to maximize our received reward.

This strategy lies at the heart of the Upper Confidence Bound (UCB) family of algorithms, due to [12], a simple variant of which we describe here. Before continuing, we recall the standard result on sub-Gaussian random variables of Corollary 4.1.10 in our context, though we require a somewhat more careful calculation because of the sequential nature of our process. Let $T_i(t) = \operatorname{card}\{\tau \le t : A_\tau = i\}$ denote the number of times that arm i has been pulled by time t of the bandit process. Then if we define

$$\widehat{\mu}_i(t) := \frac{1}{T_i(t)} \sum_{\tau \le t, A_\tau = i} Y_i(\tau),$$

to be the running average of the rewards of arm i at time t (computed only on those instances in which arm i was selected), we claim that for all i and all t,

$$\mathbb{P}\left(\widehat{\mu}_i(t) \ge \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{T_i(t)}}\right) \vee \mathbb{P}\left(\widehat{\mu}_i(t) \le \mu_i - \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{T_i(t)}}\right) \le \delta. \tag{19.1.1}$$

That is, so long as we pull the arms sufficiently many times, we are unlikely to pull the wrong arm. We prove the claim (19.1.1) in the appendix to this chapter.

Here then is the UCB procedure:

Input: Sub-gaussian parameter σ^2 and sequence of deviation probabilities $\delta_1, \delta_2, \ldots$.
Initialization: Play each arm $i = 1, \ldots, K$ once
Repeat: for each iteration t, play the arm maximizing

$$\widehat{\mu}_i(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}}.$$

Figure 19.1: The Upper Confidence Bound (UCB) Algorithm

If we define

$$\Delta_i := \mu_{i^\star} - \mu_i$$

to be the gap in means between the optimal arm and any sub-optimal arm, we then obtain the following guarantee on the expected number of pulls of any sub-optimal arm i after n steps.

Proposition 19.1.1. *Assume that each of the K arms is σ^2-sub-Gaussian and let the sequence $\delta_1 \geq \delta_2 \geq \cdots$ be non-increasing and positive. Then for any n and any arm $i \neq i^\star$,*

$$\mathbb{E}[T_i(n)] \leq \left\lceil \frac{4\sigma^2 \log \frac{1}{\delta_n}}{\Delta_i^2} \right\rceil + 2 \sum_{t=2}^{n} \delta_t.$$

Proof Without loss of generality, we assume arm 1 satisfies $\mu_1 = \max_i \mu_i$, and let arm i be any sub-optimal arm. The key insight is to carefully consider what occurs if we play arm i in the UCB procedure of Figure 19.1. In particular, if we play arm i at time t, then we certainly have

$$\widehat{\mu}_i(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}} \geq \widehat{\mu}_1(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_1(t)}}.$$

For this to occur, at least one of the following three events must occur (we suppress the dependence on i for each of them):

$$\mathcal{E}_1(t) := \left\{ \widehat{\mu}_i(t) \geq \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}} \right\}, \quad \mathcal{E}_2(t) := \left\{ \widehat{\mu}_1(t) \leq \mu_1 - \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_1(t)}} \right\},$$

$$\mathcal{E}_3(t) := \left\{ \Delta_i \leq 2\sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}} \right\}.$$

Indeed, suppose that none of the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ occur at time t. Then we have

$$\widehat{\mu}_i(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}} < \mu_i + 2\sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}} < \mu_i + \Delta_i = \mu_1 < \widehat{\mu}_1(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_1(t)}},$$

the inequalities following by $\mathcal{E}_1$, $\mathcal{E}_3$, and $\mathcal{E}_2$, respectively.

Now, for any $l \in \{1, \ldots, n\}$, we see that

$$\begin{aligned}\mathbb{E}[T_i(n)] = \sum_{t=1}^{n} \mathbb{E}[\mathbf{1}\left\{A_t = i\right\}] &= \sum_{t=1}^{n} \mathbb{E}[\mathbf{1}\left\{A_t = i, T_i(t) > l\right\} + \mathbf{1}\left\{A_t = i, T_i(t) \leq l\right\}] \\ &\leq l + \sum_{t=l+1}^{n} \mathbb{P}(A_t = i, T_i(t) > l).\end{aligned}$$

Now, we use that δ_t is non-increasing, and see that if we set

$$l^\star = \left\lceil 4\frac{\sigma^2 \log \frac{1}{\delta_n}}{\Delta_i^2} \right\rceil,$$

then to have $T_i(t) > l^\star$ it must be the case that $\mathcal{E}_3(t)$ cannot occur—that is, we would have $2\sqrt{\sigma^2 \log \frac{1}{\delta_t}/T_i(t)} > 2\sqrt{\sigma^2 \log \frac{1}{\delta_t}/l} \geq \Delta_i$. Thus we have

$$\mathbb{E}[T_i(n)] = \sum_{t=1}^n \mathbb{E}[\mathbf{1}\left\{A_t = i\right\}] \leq l^\star + \sum_{t=l^\star+1}^n \mathbb{P}(A_t = i, \mathcal{E}_3(t) \text{ fails})$$
$$\leq l^\star + \sum_{t=l^\star+1}^n \mathbb{P}(\mathcal{E}_1(t) \text{ or } \mathcal{E}_2(t)) \leq l^\star + \sum_{t=l^\star+1}^n 2\delta_t.$$

This implies the desired result. □

Naturally, the number of times arm i is selected in the sequential game is related to the regret of a procedure; indeed, we have

$$\mathsf{Reg}_n = \sum_{t=1}^n (\mu_{i^\star} - \mu_{A_t}) = \sum_{i=1}^K (\mu_{i^\star} - \mu_i) T_i(n) = \sum_{i=1}^K \Delta_i T_i(n).$$

Using this identity, we immediately obtain two theorems on the (expected) regret of the UCB algorithm.

Theorem 19.1.2. *Let $\delta_t = \delta/t^2$ for all t. Then for any $n \in \mathbb{N}$ the UCB algorithm attains*

$$\overline{\mathsf{Reg}}_n \leq \sum_{i \neq i^\star} \frac{4\sigma^2[2\log n - \log \delta]}{\Delta_i} + \frac{\pi^2 - 2}{3}\left(\sum_{i=1}^K \Delta_i\right)\delta + \sum_{i=1}^K \Delta_i.$$

Proof First, we note that

$$\mathbb{E}[\Delta_i T_i(n)] \leq \Delta_i \left\lceil 4\sigma^2 \log \frac{1}{\delta_n}/\Delta_i^2 \right\rceil + 2\Delta_i \sum_{t=2}^n \frac{\delta}{t^2} \leq \frac{4\sigma^2 \log \frac{1}{\delta_n}}{\Delta_i} + \Delta_i + 2\Delta_i \sum_{t=2}^n \frac{\delta}{t^2}$$

by Proposition 19.1.1. Summing over $i \neq i^\star$ and noting that $\sum_{t \geq 2} t^{-2} = \pi^2/6 - 1$ gives the result. □

Let us unpack the bound of Theorem 19.1.2 slightly. First, we make the simplifying assumption that $\delta_t = 1/t^2$ for all t, and let $\Delta = \min_{i \neq i^\star} \Delta_i$. In this case, we have expected regret bounded by

$$\overline{\mathsf{Reg}}_n \leq 8\frac{K\sigma^2 \log n}{\Delta} + \frac{\pi^2 + 1}{3}\sum_{i=1}^K \Delta_i.$$

So we see that the asymptotic regret with this choice of δ scales as $(K\sigma^2/\Delta)\log n$, roughly linear in the classes, logarithmic in n, and inversely proportional to the gap in means. As a concrete example, if we know that the rewards for each arm Y_i belong to the interval $[0, 1]$, then Hoeffding's lemma (recall Example 4.1.6) states that we may take $\sigma^2 = 1/4$. Thus the mean regret becomes at most $\sum_{i:\Delta_i > 0} \frac{2\log n}{\Delta_i}(1 + o(1))$, where the $o(1)$ term tends to zero as $n \to \infty$.

If we knew a bit more about our problem, then by optimizing over δ and choosing $\delta = \sigma^2/\Delta$, we obtain the upper bound

$$\mathsf{R\overline{eg}}_n \le O(1) \left[\frac{K\sigma^2}{\Delta} \log \frac{n\Delta}{\sigma^2} + K \frac{\max_i \Delta_i}{\min_i \Delta_i} \right], \tag{19.1.2}$$

that is, the expected regret scales asymptotically as $(K\sigma^2/\Delta) \log(\frac{n\Delta}{\sigma^2})$—linearly in the number of classes, logarithmically in n, and inversely proportional to the gap between the largest and other means.

If any of the gaps $\Delta_i \to 0$ in the bound of Theorem 19.1.2, the bound becomes vacuous—it simply says that the regret is upper bounded by infinity. Intuitively, however, pulling a *slightly* sub-optimal arm should be insignificant for the regret. With that in mind, we present a slight variant of the above bounds, which has a worse scaling with n—the bound scales as $\sqrt{n}$ rather than $\log n$—but is independent of the gaps Δ_i.

Theorem 19.1.3. *If UCB is run with parameter $\delta_t = 1/t^2$, then*

$$\mathsf{R\overline{eg}}_n \le \sqrt{8K\sigma^2 n \log n} + 4 \sum_{i=1}^K \Delta_i.$$

Proof Fix any $\gamma > 0$. Then we may write the regret with the standard identity

$$\mathsf{Reg}_n = \sum_{i \neq i^\star} \Delta_i T_i(n) = \sum_{i:\Delta_i \ge \gamma} \Delta_i T_i(n) + \sum_{i:\Delta_i < \gamma} \Delta_i T_i(n) \le \sum_{i:\Delta_i \ge \gamma} \Delta_i T_i(n) + n\gamma,$$

where the final inequality uses that certainly $\sum_{i=1}^K T_i(n) \le n$. Taking expectations with our UCB procedure and $\delta = 1$, we have by Theorem 19.1.2 that

$$\mathsf{R\overline{eg}}_n \le \sum_{i:\Delta_i \ge \gamma} \Delta_i \frac{8\sigma^2 \log n}{\Delta_i^2} + \frac{\pi^2 + 1}{3} \sum_{i=1}^K \Delta_i + n\gamma \le K \frac{8\sigma^2 \log n}{\gamma} + n\gamma + \frac{\pi^2 + 1}{3} \sum_{i=1}^K \Delta_i,$$

Optimizing over γ by taking $\gamma = \frac{\sqrt{8K\sigma^2 \log n}}{\sqrt{n}}$ gives the result. □

Combining the above two theorems, we see that the UCB algorithm with parameters $\delta_t = 1/t^2$ automatically achieves the expected regret guarantee

$$\mathsf{R\overline{eg}}_n \le C \cdot \min \left\{ \sum_{i:\Delta_i > 0} \frac{\sigma^2 \log n}{\Delta_i}, \sqrt{K\sigma^2 n \log n} \right\}. \tag{19.1.3}$$

That is, UCB enjoys some adaptive behavior. It is not, however, optimal; there are algorithms, including Audibert and Bubeck's MOSS (Minimax Optimal in the Stochastic Case) bandit procedure [11], which achieve regret

$$\mathsf{R\overline{eg}}_n \le C \cdot \min \left\{ \sqrt{Kn}, \frac{K}{\Delta} \log \frac{n\Delta^2}{K} \right\},$$

which is essentially the bound specified by inequality (19.1.2) (which required knowledge of the Δ_is) and an improvement by $\log n$ over the analysis of Theorem 19.1.3. It is also possible to provie a high-probability guarantee for the UCB algorithms, which follows essentially immediately from the proof techniques of Proposition 19.1.1, but we leave this to the interested reader.

19.2 Bayesian approaches to bandits

The upper confidence bound procedure, while elegant and straightforward, has a variety of competitors, including online gradient descent approaches and a variety of Bayesian strategies. Bayesian strategies—because they (can) incorporate prior knowledge—have the advantage that they suggest policies for exploration and trading between regret and information; that is, they allow us to quantify a value for information. They often yield very simple procedures, allowing simpler implementations.

In this section, we thus consider the following specialized setting; there is substantially more possible here. We assume that there is a finite set of actions (arms) $\mathcal{A}$ as before, and we have a collection of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ parameterized by a set Θ (often, this is some subset of $\mathbb{R}^K$ when we look at K-armed bandit problems with $\mathrm{card}(\mathcal{A}) = K$, but we stay in this abstract setting temporarily). We also have a loss function $\ell : \mathcal{A} \times \Theta \to \mathbb{R}$ that measure the quality of an action $a \in \mathcal{A}$ for the parameter θ.

Example 19.2.1 (Classical Bernoulli bandit problem)**:** The classical bandit problem, as in the UCB case of the previous section, has actions (arms) $\mathcal{A} = \{1, \ldots, K\}$, and the parameter space $\Theta = [0,1]^K$, and we have that P_θ is a distribution on $Y \in \{0,1\}^K$, where Y has independent coordinates $1, \ldots, K$ with $P(Y_j = 1) = \theta_j$, that is, $Y_j \sim \mathsf{Bernoulli}(\theta_j)$. The goal is to find the arm with highest mean reward, that is, $\mathrm{argmax}_j \theta_j$, and thus possible loss functions include $\ell(a, \theta) = -\theta_a$ or, if we wish the loss to be positive, $\ell(a, \theta) = 1 - \theta_a \in [0,1]$. $\diamond$

Lastly, in this Bayesian setting, we require a prior distribution π on the space Θ, where $\pi(\Theta) = 1$. We then define the Bayesian regret as

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) = \mathbb{E}_\pi \left[\sum_{t=1}^n \ell(A_t, \theta) - \ell(A^\star, \theta) \right], \tag{19.2.1}$$

where $A^\star \in \mathrm{argmin}_{a \in \mathcal{A}} \ell(a, \theta)$ is the minimizer of the loss, and $A_t \in \mathcal{A}$ is the action the player takes at time t of the process. The expectation (19.2.1) is taken both over the randomness in θ according to the prior π and any randomness in the player's strategy for choosing the actions A_t at each time.

Our approaches in this section build off of those in Chapter 16, except that we no longer fully observe the desired observations Y—we may only observe $Y_{A_t}(t)$ at time t, which may provide less information. The broad algorithmic framework for this section is as follows. We now give several

Input: Prior distribution π on space Θ, family of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$

Repeat: for each iteration t, choose distribution π_t on space Θ (based on history $Y_{A_1}(1), \ldots, Y_{A_{t-1}}(t-1)$). Draw

$$\theta_t \sim \pi_t.$$

Play action $A_t \in \mathcal{A}$ minimizing

$$\ell(a, \theta_t)$$

over $a \in \mathcal{A}$, observe $Y_{A_t}(t)$.

Figure 19.2: The generic Bayesian algorithm

concrete instantiations of this broad procedure, as well as tools (both information-theoretic and otherwise) for its analysis.

19.2.1 Posterior (Thompson) sampling

The first strategy we consider is perhaps the simplest; in Algorithm 19.2, it corresponds to using π_t to be the posterior distribution on θ conditional on the history $Y_{A_1}(1), \ldots, Y_{A_{t-}}(t-1)$. That is, we let

$$\mathcal{H}_t := \{A_1, Y_{A_1}(1), A_2, Y_{A_2}(2), \ldots, A_t, Y_{A_t}(t)\}$$

denote the history (or the σ-field thereof) of the procedure and rewards up to time t. Then at iteration t, we use the posterior

$$\pi_t(\theta) = \pi(\theta \mid \mathcal{H}_{t-1}),$$

the distribution on θ conditional on $\mathcal{H}_{t-1}$. This procedure was originally proposed by Thompson [165] in 1933 in the first paper on bandit problems. There are several analyses of Thompson (and related Bayesian) procedures possible; our first analysis proceeds by using confidence bounds, while our later analyses give a more information theoretic analysis.

First, we provide a more concrete specification of Algorithm 19.2 for Thompson (posterior) sampling in the case of Bernoulli rewards.

Example 19.2.2 (Thompson sampling with Bernoulli penalities)**:** Let us suppose that the vector $\theta \in [0,1]^K$, and we draw $\theta_i \sim \mathsf{Beta}(1,1)$, which corresponds to the uniform distribution on $[0,1]^d$. The actions available are simply to select one of the coordinates, $a \in \mathcal{A} = \{1, \ldots, K\}$, and we observe $Y_a \sim \mathsf{Bernoulli}(\theta_a)$, that is, $\mathbb{P}(Y_a = 1 \mid \theta) = \theta_a$. That is, $\ell(a, \theta) = \theta_a$. Let $T_a^1(t) = \mathrm{card}\{\tau \le t : A_t = a, Y_a(\tau) = 1\}$ be the number of times arm a is pulled and results in a loss of 1 by time t, and similarly let $T_a^0(t) = \mathrm{card}\{\tau \le t : A_t = a, Y_a(\tau) = 0\}$. Then, recalling Example 16.4.2 on Beta-Bernoulli distributions, Thompson sampling proceeds as follows:

(1) For each arm $a \in \mathcal{A} = \{1, \ldots, K\}$, draw $\theta_a(t) \sim \mathsf{Beta}(1 + T_a^1(t), 1 + T_a^0(t))$.

(2) Play the action $A_t = \mathrm{argmin}_a \, \theta_a(t)$.

(3) Observe the loss $Y_{A_t}(t) \in \{0, 1\}$, and increment the appropriate count.

Thompson sampling is simple in this case, and it is implementable with just a few counters. $\diamond$

We may extend Example 19.2.2 to the case in which the losses come from any distribution with mean θ_i, so long as the distribution is supported on $[0,1]$. In particular, we have the following example.

Example 19.2.3 (Thompson sampling with bounded random losses)**:** Let us again consider the setting of Example 19.2.2, except that the observed losses $Y_a(t) \in [0,1]$ with $\mathbb{E}[Y_a \mid \theta] = \theta_a$. The following modification allows us to perform Thompson sampling in this case, even without knowing the distribution of $Y_a \mid \theta$: instead of observing a loss $Y_a \in \{0,1\}$, we construct a random observation $\widetilde{Y}_a \in \{0,1\}$ with the property that $\mathbb{P}(\widetilde{Y}_a = 1 \mid Y_a) = Y_a$. Then the losses $\ell(a, \theta) = \theta_a$ are identical, and the posterior distribution over θ is still a Beta distribution. We simply redefine

$$T_a^0(t) := \mathrm{card}\{\tau \le t : A_t = a, \widetilde{Y}_a(\tau) = 0\} \quad \text{and} \quad T_a^1(t) := \mathrm{card}\{\tau \le t : A_t = a, \widetilde{Y}_a(\tau) = 0\}.$$

The Thompson sampling procedure is otherwise identical. $\diamond$

Our first analysis shows that Thompson sampling can guarantee performance similar to (or in some cases, better than) confidence-based procedures, which we do by using a sequence of

(potential) lower and upper bounds on the losses of actions. (Recall we wish to minimize our losses, so that we would optimistically play those arms with the lowest estimated loss.) This analysis is based on that of Russo and Van Roy [155]. Let $L_t : \mathcal{A} \to \mathbb{R}$ and $U_t : \mathcal{A} \to \mathbb{R}$ be an arbitrary sequence of (random) functions that are measurable with respect to $\mathcal{H}_{t-1}$, that is, they are constructed based only on $\{A_1, Y_{A_1}(1), \ldots, A_{t-1}, Y_{A_{t-1}}(t-1)\}$. Then we can decompose the Bayesian regret (19.2.1) as

$$\begin{aligned}
\mathsf{Reg}_n(\mathcal{A}, \ell, \pi) &= \mathbb{E}_\pi\left[\sum_{t=1}^n \ell(A_t, \theta) - \ell(A^\star, \theta)\right] \\
&= \sum_{t=1}^n \mathbb{E}_\pi[U_t(A_t) - L_t(A_t)] + \sum_{t=1}^n \mathbb{E}_\pi[\ell(A_t, \theta) - U_t(A_t)] + \sum_{t=1}^n \mathbb{E}_\pi[L_t(A_t) - \ell(A^\star, \theta)] \\
&\stackrel{(i)}{=} \sum_{t=1}^n \mathbb{E}_\pi[U_t(A_t) - L_t(A_t)] + \sum_{t=1}^n \mathbb{E}_\pi[\ell(A_t, \theta) - U_t(A_t)] + \sum_{t=1}^n \mathbb{E}_\pi[L_t(A_t^\star) - \ell(A_t^\star, \theta)],
\end{aligned} \tag{19.2.2}$$

where in equality (i) we used that conditional on $\mathcal{H}_{t-1}$, A_t and $A_t^\star = A^\star$ have the same distribution, as we sample from the posterior $\pi(\theta \mid \mathcal{H}_{t-1})$, and L_t is a function of $\mathcal{H}_{t-1}$. With the decomposition (19.2.2) at hand, we may now provide an expected regret bound for Thompson (or posterior) sampling. We remark that the behavior of Thompson sampling is independent of these upper and lower bounds U_t, L_t we have chosen—they are simply an artifact to make analysis easier.

Theorem 19.2.4. *Suppose that conditional on the choice of action $A_t = a$, the received loss $Y_a(t)$ is σ^2-sub-Gaussian with mean $\ell(a, \theta)$, that is,*

$$\mathbb{E}\left[\exp\left(\lambda(Y_a(t) - \ell(a, \theta))\right) \mid \mathcal{H}_{t-1}\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \quad \text{for all } a \in \mathcal{A}.$$

Then for all $\delta \ge 0$ we have

$$\mathsf{Reg}_n(\mathcal{A}, \ell, \pi) \le 4\sqrt{2\sigma^2 \log \frac{1}{\delta}}\sqrt{|\mathcal{A}|n} + 3n\delta\sigma|\mathcal{A}|.$$

In particular, choosing $\delta = \frac{1}{n}$ gives

$$\mathsf{Reg}_n(\mathcal{A}, \ell, \pi) \le 6\sigma\sqrt{|\mathcal{A}| n \log n} + 3\sigma|\mathcal{A}|.$$

Proof We choose the upper and lower bound functions somewhat carefully so as to get a fairly sharp regret guarantee. In particular, we (as in our analysis of the UCB algorithm) let $\delta \in (0, 1)$ and define $T_a(t) := \mathrm{card}\{\tau \le t : A_t = a\}$ to be the number of times that action a has been chosen by iteration t. Then we define the mean loss for action a at time t by

$$\widehat{\ell}_a(t) := \frac{1}{T_a(t)} \sum_{\tau \le t, A_\tau = a} Y_a(\tau)$$

and our bounds for the analysis by

$$U_t(a) := \widehat{\ell}_a(t) + \sqrt{\frac{2\sigma^2 \log \frac{1}{\delta}}{T_a(t)}} \quad \text{and} \quad L_t(a) := \widehat{\ell}_a(t) - \sqrt{\frac{2\sigma^2 \log \frac{1}{\delta}}{T_a(t)}}.$$

With these choices, we see that by the extension of the sub-Gaussian concentration bound (19.1.1) and the equality (19.5.1) showing that the sum $\sum_{\tau \le t, A_\tau = a} Y_a(\tau)$ is equal in distribution to the sum $\sum_{\tau \le t, A_\tau = a} Y_a'(\tau)$, where $Y_a'(\tau)$ are independent and identically distributed copies of $Y_a(\tau)$, we have for any $\epsilon \ge 0$ that

$$\mathbb{P}(U_t(a) \le \ell(a, \theta) - \epsilon \mid T_a(t)) \le \exp\left(-\frac{T_a(t)}{2\sigma^2}\left(\sqrt{\frac{2\sigma^2 \log \frac{1}{\delta}}{T_a(t)}} + \epsilon\right)^2\right) \le \exp\left(-\log \frac{1}{\delta} - \frac{T_a(t)\epsilon^2}{2\sigma^2}\right), \tag{19.2.3}$$

where the final inequality uses that $(a+b)^2 \ge a^2 + b^2$ for $ab \ge 0$. We have an identical bound for $\mathbb{P}(L_t(a) \ge \ell(a, \theta) + \epsilon \mid T_a(t))$.

We may now bound the final two sums in the regret expansion (19.2.2) using inequality (19.2.3). First, however, we make the observation that for any nonnegative random variable Z, we have $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \ge \epsilon) d\epsilon$. Using this, we have

$$\begin{aligned}
\sum_{t=1}^n \mathbb{E}_\pi \left[\ell(A_t, \theta) - U_t(A_t)\right] &\le \sum_{t=1}^n \sum_{a \in \mathcal{A}} \mathbb{E}_\pi \left[\left[\ell(a, \theta) - U_t(a)\right]_+\right] \\
&= \sum_{t=1}^n \sum_{a \in \mathcal{A}} \mathbb{E}_\pi \left[\int_0^\infty \mathbb{P}(U_t(a) \ge \ell(a, \theta) + \epsilon \mid T_a(t)) d\epsilon\right] \\
&\overset{(i)}{\le} \sum_{t=1}^n \sum_{a \in \mathcal{A}} \delta \mathbb{E}_\pi \left[\int_0^\infty \exp\left(-\frac{T_a(t)\epsilon^2}{2\sigma^2}\right) d\epsilon\right] \overset{(ii)}{=} \sum_{t=1}^n \delta \sum_{a \in \mathcal{A}} \mathbb{E}_\pi \left[\sqrt{\frac{\pi\sigma^2}{2T_a(t)}}\right],
\end{aligned}$$

where inequality (i) uses the bound (19.2.3) and equality (ii) uses that this is the integral of half of a normal density. Substituting this bound, as well as the identical one for the terms involving $L_t(A_t^\star)$, into the decomposition (19.2.2) yields

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) \le \sum_{t=1}^n \mathbb{E}_\pi[U_t(A_t) - L_t(A_t)] + \sum_{t=1}^n \delta \sum_{a \in \mathcal{A}} \mathbb{E}_\pi \left[\sqrt{\frac{2\pi\sigma^2}{T_a(t)}}\right].$$

Using that $T_a(t) \ge 1$ for each action a, we have $\sum_{a \in \mathcal{A}} \mathbb{E}_\pi[\sqrt{2\pi\sigma^2 / T_a(t)}] < 3\sigma|\mathcal{A}|$. Lastly, we use that

$$U_t(A_t) - L_t(A_t) = 2\sqrt{\frac{2\sigma^2 \log \frac{1}{\delta}}{T_{A_t}(t)}}.$$

Thus we have

$$\sum_{t=1}^n \mathbb{E}_\pi[U_t(A_t) - L_t(A_t)] = 2\sqrt{2\sigma^2 \log \frac{1}{\delta}} \sum_{a \in \mathcal{A}} \mathbb{E}_\pi \left[\sum_{t : A_t = a} \frac{1}{\sqrt{T_a(t)}}\right].$$

Once we see that $\sum_{t=1}^T t^{-\frac{1}{2}} \le \int_0^T t^{-\frac{1}{2}} dt = 2\sqrt{T}$, we have the upper bound

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) \le 4\sqrt{2\sigma^2 \log \frac{1}{\delta}} \sum_{a \in \mathcal{A}} \mathbb{E}_\pi[\sqrt{T_a(n)}] + 3n\delta\sigma|\mathcal{A}|.$$

As $\sum_{a \in \mathcal{A}} T_a(n) = n$, the Cauchy-Scwharz inequality implies $\sum_{a \in \mathcal{A}} \sqrt{T_a(n)} \le \sqrt{|\mathcal{A}|n}$, which gives the result. $\square$

An immediate Corollary of Theorem 19.2.4 is the following result, which applies in the case of bounded losses Y_a as in Examples 19.2.2 and 19.2.3.

Corollary 19.2.5. *Let the losses $Y_a \in [0,1]$ with $\mathbb{E}[Y_a \mid \theta] = \theta_a$, where $\theta_i \stackrel{\text{iid}}{\sim} \mathsf{Beta}(1,1)$ for $i = 1, \ldots, K$. Then Thompson sampling satisfies*

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) \le 3\sqrt{Kn \log n} + \frac{3}{2}K.$$

19.2.2 An information-theoretic analysis

19.2.3 Information and exploration

19.3 Online gradient descent approaches

It is also possible to use online gradient descent approaches to minimize regret in the more standard multi-armed bandit setting. In this scenario, our goal is to minimize a sequentially (partially) observed loss, as in the previous section. In this case, as usual we have K arms with non-negative means $\mu_1, \ldots, \mu_K$, and we wish to find the arm with lowest mean loss. We build off of the online convex optimization procedures of Chapter 18 to achieve good regret guarantees. In particular, at each step of the bandit procedure, we play a distribution $w_t \in \Delta_K$ on the arms, and then we select one arm j at random, each with probability $w_{t,j}$. The *expected* loss we suffer is then $\ell_t(w_t) = \langle w_t, \mu \rangle$, though we observe only a random realization of the loss for the arm a that we play.

Because of its natural connections with estimation of probability distributions, we would like to use the exponentiated gradient algorithm, Example 18.2.3, to play this game. We face one main difficulty: we must estimate the gradient of the losses, $\nabla \ell_t(w_t) = \mu$, even though we only observe a random variable $Y_a(t) \in \mathbb{R}_+$, conditional on selecting action $A_t = a$ at time t, with the property that $\mathbb{E}[Y_a(t)] = \mu_a$. Happily, we can construct such an estimate without too much additional variance.

Lemma 19.3.1. *Let $Y \in \mathbb{R}^K$ be a random variable with $\mathbb{E}[Y] = \mu$ and $w \in \Delta_K$ be a probability vector. Choose a coordinate a with probability w_a and define the random vector*

$$\widetilde{Y}_j = \begin{cases} Y_j / w_j & \text{if } j = a \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathbb{E}[\widetilde{Y} \mid Y] = Y$.

Proof The proof is immediate: for each coordinate j of $\widetilde{Y}$, we have $\mathbb{E}[\widetilde{Y}_j \mid Y] = w_j Y_j / w_j = Y_j$. $\square$

Lemma 19.3.1 suggests the following procedure, which gives rise to (a variant of) Auer et al.'s EXP3 (Exponentiated gradient for Exploration and Exploitation) algorithm [13]. We can prove the following bound on the expected regret of the EXP3 Algorithm 19.3 by leveraging our refined analysis of exponentiated gradients in Proposition 18.4.1.

Proposition 19.3.2. *Assume that for each j, we have $\mathbb{E}[Y_j^2] \le \sigma^2$ and the observed loss $Y_j \ge 0$. Then Alg. 19.3 attains expected regret (we are minimizing)*

$$\overline{\mathsf{Reg}}_n = \sum_{t=1}^n \mathbb{E}[\mu_{A_t} - \mu_{i^\star}] \le \frac{\log K}{\eta} + \frac{\eta}{2}\sigma^2 K n.$$

Input: stepsize parameter η, initial vector $w_1 = [\frac{1}{K} \; \cdots \; \frac{1}{K}]^\top$
Repeat: for each iteration t, choose random action $A_t = a$ with probability $w_{t,a}$
Receive non-negative loss $Y_a(t)$, and define

$$g_{t,j} = \begin{cases} Y_j(t)/w_j & \text{if } A_t = j \\ 0 & \text{otherwise.} \end{cases}$$

Update for each $i = 1, \ldots, K$

$$w_{t+1,i} = \frac{w_{t,i} \exp(-\eta g_{t,i})}{\sum_j w_{t,j} \exp(-\eta g_{t,j})}.$$

Figure 19.3: Exponentiated gradient for bandit problems.

In particular, choosing $\eta = \sqrt{\log K/(K\sigma^2 n)}$ *gives*

$$\mathsf{R\overline{eg}}_n = \sum_{t=1}^n \mathbb{E}[\mu_{A_t} - \mu_{i^\star}] \le \frac{3}{2}\sigma\sqrt{Kn \log K}.$$

Proof With Lemma 19.3.1 in place, we recall the refined regret bound of Proposition 18.4.1. We have that for $w^\star \in \Delta_K$ and any sequence of vectors $g_1, g_2, \ldots$ with $g_t \in \mathbb{R}_+^K$, then exponentiated gradient descent achieves

$$\sum_{t=1}^n \langle g_t, w_t - w^\star \rangle \le \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{j=1}^k w_{t,j} g_{t,j}^2.$$

To transform this into a useful bound, we take expectations. Indeed, we have

$$\mathbb{E}[g_t \mid w_t] = \mathbb{E}[Y] = \mu$$

by construction, and we also have

$$\mathbb{E}\left[\sum_{j=1}^k w_{t,j} g_{t,j}^2 \mid w_t\right] = \sum_{j=1}^K w_{t,j}^2 \mathbb{E}[Y_j(t)^2/w_{t,j}^2 \mid w_t] = \sum_{j=1}^K \mathbb{E}[Y_j^2] = \mathbb{E}[\|Y\|_2^2].$$

This careful normalizing, allowed by Proposition 18.4.1, is essential to our analysis (and fails for more naive applications of online convex optimization bounds). In particular, we have

$$\mathsf{R\overline{eg}}_n = \sum_{t=1}^n \mathbb{E}[\langle \mu, w_t - w^\star \rangle] = \sum_{t=1}^n \mathbb{E}[\langle g_t, w_t - w^\star \rangle] \le \frac{\log K}{\eta} + \frac{\eta}{2} n \mathbb{E}[\|Y\|_2^2].$$

Taking expectations gives the result. □

When the random observed losses $Y_a(t)$ are bounded in $[0, 1]$, then we have the mean regret bound $\frac{3}{2}\sqrt{Kn \log K}$, which is as sharp as any of our other bounds.

19.4 Further notes and references

An extraordinarily abbreviated bibliography follows.

The golden oldies: Thompson [165], Robbins [152], and Lai and Robbins [125].

More recent work in machine learning (there are far too many references to list): the books Cesa-Bianchi and Lugosi [47] and Bubeck and Cesa-Bianchi [40] are good references. The papers of Auer et al. [13] and Auer et al. [12] introduced UCB and EXP3.

Our approach to Bayesian bandits follows Russo and Van Roy [155, 156, 157]. More advanced techniques allow Thompson sampling to apply even when the prior is unknown (e.g. Agrawal and Goyal [2]).

19.5 Technical proofs

19.5.1 Proof of Claim (19.1.1)

We let $Y_i'(\tau)$, for $\tau = 1, 2, \ldots$, be independent and identically distributed copies of the random variables $Y_i(\tau)$, so that $Y_i'(\tau)$ is also independent of $T_i(t)$ for all t and τ. We claim that the pairs

$$(\widehat{\mu}_i(t), T_i(t)) \stackrel{\text{dist}}{=} \left(\widehat{\mu}_i'(t), T_i(t)\right), \tag{19.5.1}$$

where $\widehat{\mu}_i'(t) = \frac{1}{T_i(t)} \sum_{\tau : A_\tau = i} Y_i'(\tau)$ is the empirical mean of the copies $Y_i'(\tau)$ for those steps when arm i is selected. To see this, we use the standard fact that the characteristic function of a random variable completely characterizes the random variable. Let $\varphi_{Y_i}(\lambda) = \mathbb{E}[e^{\iota \lambda Y_i}]$, where $\iota = \sqrt{-1}$ is the imaginary unit, denote the characteristic function of Y_i, noting that by construction we have $\varphi_{Y_i} = \varphi_{Y_i'}$. Then writing the joint characteristic function of $T_i(t)\widehat{\mu}_i(t)$ and $T_i(t)$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\iota\lambda_1 \sum_{\tau=1}^t \mathbf{1}\left\{A_\tau = i\right\} Y_i(\tau) + \iota\lambda_2 T_i(t)\right)\right] \\
&\stackrel{(i)}{=} \mathbb{E}\left[\prod_{\tau=1}^t \mathbb{E}\left[\exp\left(\iota\lambda_1 \mathbf{1}\left\{A_\tau = i\right\} Y_i(\tau) + \iota\lambda_2 \mathbf{1}\left\{A_\tau = i\right\}\right) \mid \mathcal{H}_{\tau-1}\right]\right] \\
&\stackrel{(ii)}{=} \mathbb{E}\left[\prod_{\tau=1}^t \left(\mathbf{1}\left\{A_\tau = i\right\} e^{\iota\lambda_2} \mathbb{E}\left[\exp(\iota\lambda_1 Y_i(\tau)) \mid \mathcal{H}_{\tau-1}\right] + \mathbf{1}\left\{A_\tau \neq i\right\}\right)\right] \\
&\stackrel{(iii)}{=} \mathbb{E}\left[\prod_{\tau=1}^t \left(\mathbf{1}\left\{A_\tau = i\right\} e^{\lambda_2\iota} \varphi_{Y_i}(\lambda_1) + \mathbf{1}\left\{A_\tau \neq i\right\}\right)\right] \\
&\stackrel{(iv)}{=} \mathbb{E}\left[\prod_{\tau=1}^t \left(\mathbf{1}\left\{A_\tau = i\right\} e^{\lambda_2\iota} \varphi_{Y_i'}(\lambda_1) + \mathbf{1}\left\{A_\tau \neq i\right\}\right)\right] \\
&= \mathbb{E}\left[\exp\left(\iota\lambda_1 \sum_{\tau=1}^t \mathbf{1}\left\{A_\tau = i\right\} Y_i'(\tau) + \iota\lambda_2 T_i(t)\right)\right],
\end{aligned}$$

where equality (i) is the usual tower property of conditional expectations, where $\mathcal{H}_{\tau-1}$ denotes the history to time $\tau - 1$, equality (ii) because $A_\tau \in \mathcal{H}_{\tau-1}$ (that is, it is a function of the history), equality (iii) follows because $Y_i(\tau)$ is independent of $\mathcal{H}_{\tau-1}$, and equality (iv) follows because Y_i' and Y_i have identical distributions. The final step is simply reversing the steps.

With the distributional equality (19.5.1) in place, we see that for any $\delta \in [0, 1]$, we have

$$\begin{aligned}
\mathbb{P}\left(\widehat{\mu}_i(t) \geq \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{T_i(t)}}\right) &= \mathbb{P}\left(\widehat{\mu}_i'(t) \geq \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{T_i(t)}}\right) = \mathbb{P}\left(\widehat{\mu}_i'(t) \geq \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{T_i(t)}}\right) \\
&= \sum_{s=1}^{t} \mathbb{P}\left(\widehat{\mu}_i'(t) \geq \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{s}} \mid T_i(t) = s\right) \mathbb{P}(T_i(t) = s) \\
&\leq \sum_{s=1}^{t} \delta \mathbb{P}(T_i(t) = s) = \delta.
\end{aligned}$$

The proof for the lower tail is similar.

Part V

Appendices

Appendix A

Miscellaneous mathematical results

This appendix collects several mathematical results and some of the more advanced mathematical treatment required for full proofs of the results in the book. It is not a core part of the book, but it does provide readers who wish to see the measure-theoretic rigor necessary for some of our results, or otherwise, to dot the appropriate I's and cross the appropriate T's.

A.1 The roots of a polynomial

A.2 Measure-theoretic development of divergence measures

Appendix B

Convex Analysis

In this appendix, we review several results in convex analysis that are useful for our purposes. We give only a cursory study here, identifying the basic results and those that will be of most use to us; the field of convex analysis as a whole is vast. The study of convex analysis and optimization has become very important practically in the last fourty to fifty years for a few reasons, the most important of which is probably that convex optimization problems—those optimization problems in which the objective and constraints are convex—are tractable, while many others are not. We do not focus on optimization ideas here, however, building only some analytic tools that we will find useful. We borrow most of our results from Hiriart-Urruty and Lemaréchal [104], focusing mostly on the finite-dimensional case (though we present results that apply in infinite dimensional cases with proofs that extend straightforwardly, and we do not specify the domains of our functions unless necessary), as we require no results from infinite-dimensional analysis.

In addition, we abuse notation and assume that the range of any function is the *extended real line*, meaning that if $f : C \to \mathbb{R}$ we mean that $f(x) \in \mathbb{R} \cup \{-\infty, +\infty\}$, where $-\infty$ and $+\infty$ are infinite and satisfy $a + \infty = +\infty$ and $a - \infty = -\infty$ for any $a \in \mathbb{R}$. However, we assume throughout and without further mention that our functions are *proper*, meaning that $f(x) > -\infty$ for all x, as this allows us to avoid annoying pathologies.

B.1 Convex sets

We begin with the simplest and most important object in convex analysis, a convex set.

Definition B.1. *A set C is* convex *if for all $\lambda \in [0, 1]$ and all $x, y \in C$, we have*

$$\lambda x + (1 - \lambda)y \in C.$$

An important restriction of convex sets is to *closed* convex sets, those convex sets that are, well, closed.

JCD Comment: Picture

We now consider two operations that extend sets, convexifying them in nice ways.

Definition B.2. *The* affine hull *of a set C is the smallest affine set containing C. That is,*

$$\operatorname{aff}(C) := \left\{ \sum_{i=1}^{k} \lambda_i x_i : k \in \mathbb{N}, x_i \in C, \lambda \in \mathbb{R}^k, \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

Associated with any set is also its convex hull:

Definition B.3. *The* convex hull *of a set $C \subset \mathbb{R}^d$, denoted* $\mathrm{Conv}(C)$*, is the intersection of all convex sets containing C.*

JCD Comment: picture

An almost immediate associated result is that the convex hull of a set is equal to the set of all convex combinations of points in the set.

Proposition B.1.1. *Let C be an arbitrary set. Then*

$$\mathrm{Conv}(C) = \left\{ \sum_{i=1}^k \lambda_i x_i : k \in \mathbb{N}, x_i \in C, \lambda \in \mathbb{R}_+^k, \sum_{i=1}^k \lambda_i = 1 \right\}.$$

Proof Call T the set on the right hand side of the equality in the proposition. Then $T \supset C$ is clear, as we may simply take $\lambda_1 = 1$ and vary $x \in C$. Moreover, the set $T \subset \mathrm{Conv}(C)$, as any convex set containing C must contain all convex combinations of its elements; similarly, any convex set $S \supset C$ must have $S \supset T$.

Thus if we show that T is convex, then we are done. Take any two points $x, y \in T$. Then $x = \sum_{i=1}^k \alpha_i x_i$ and $y = \sum_{i=1}^l \beta_i y_i$ for $x_i, y_i \in C$. Fix $\lambda \in [0, 1]$. Then $(1 - \lambda)\beta_i \geq 0$ and $\lambda \alpha_i \geq 0$ for all i,

$$\lambda \sum_{i=1}^k \alpha_i + (1 - \lambda) \sum_{i=1}^l \beta_i = \lambda + (1 - \lambda) = 1,$$

and $\lambda x + (1 - \lambda) y$ is a convex combination of the points x_i and y_i weighted by $\lambda \alpha_i$ and $(1 - \lambda)\beta_i$, respectively. So $\lambda x + (1 - \lambda) y \in T$ and T is convex. □

We also give one more definition, which is useful for dealing with some pathalogical cases in convex analysis, as it allows us to assume many sets are full-dimensional.

Definition B.4. *The* relative interior *of a set C is the interior of C relative to its affine hull, that is,*

$$\mathrm{relint}(C) := \{x \in C : B(x, \epsilon) \cap \mathrm{aff}(C) \subset C \text{ for some } \epsilon > 0\},$$

where $B(x, \epsilon) := \{y : \|y - x\| < \epsilon\}$ denotes the open ball of radius ϵ centered at x.

An example may make Definition B.4 clearer.

Example B.1.2 (Relative interior of a disc)**:** Consider the (convex) set

$$C = \left\{ x \in \mathbb{R}^d : x_1^2 + x_2^2 \leq 1, \; x_j = 0 \text{ for } j \in \{3, \ldots, d\} \right\}.$$

The affine hull $\mathrm{aff}(C) = \mathbb{R}^2 \times \{0\} = \{(x_1, x_2, 0, \ldots, 0) : x_1, x_2 \in \mathbb{R}\}$ is simply the (x_1, x_2)-plane in $\mathbb{R}^d$, while the relative interior $\mathrm{relint}(C) = \{x \in \mathbb{R}^d : x_1^2 + x_2^2 < 1\} \cap \mathrm{aff}(C)$ is the "interior" of the 2-dimensional disc in $\mathbb{R}^d$. ◇

In finite dimensions, we may actually restrict the definition of the convex hull of a set C to convex combinations of a bounded number (the dimension plus one) of the points in C, rather than arbitrary convex combinations as required by Proposition B.1.1. This result is known as *Carathéodory's theorem*.

Theorem B.1.3. *Let $C \subset \mathbb{R}^d$. Then $x \in \mathrm{Conv}(C)$ if and only if there exist points $x_1, \ldots, x_{d+1} \in C$ and $\lambda \in \mathbb{R}_+^{d+1}$ with $\sum_{i=1}^{d+1} \lambda_i = 1$ such that*

$$x = \sum_{i=1}^{d+1} \lambda_i x_i.$$

Proof It is clear that if x can be represented as such a sum, then $x \in \mathrm{Conv}(C)$. Conversely, Proposition B.1.1 implies that for any $x \in \mathrm{Conv}(C)$ we have

$$x = \sum_{i=1}^{k} \lambda_i x_i, \quad \lambda_i \geq 0, \; \sum_{i=1}^{k} \lambda_i = 1, \; x_i \in C$$

for some λ_i, x_i. Assume that $k > d+1$ and $\lambda_i > 0$ for each i, as otherwise, there is nothing to prove. Then we know that the points $x_i - x_1$ are certainly linearly dependent (as there are $k - 1 > d$ of them), and we can find (not identically zero) values $\alpha_2, \ldots, \alpha_k$ such that $\sum_{i=2}^{k} \alpha_i (x_i - x_1) = 0$. Let $\alpha_1 = -\sum_{i=2}^{k} \alpha_i$ to obtain that we have both

$$\sum_{i=1}^{k} \alpha_i x_i = 0 \;\; \text{and} \;\; \sum_{i=1}^{k} \alpha_i = 0. \tag{B.1.1}$$

Notably, the equalities (B.1.1) imply that at least one $\alpha_i > 0$, and if we define $\lambda^* = \min_{i:\alpha_i > 0} \frac{\lambda_i}{\alpha_i} > 0$, then setting $\lambda_i' = \lambda_i - \lambda^* \alpha_i$ we have

$$\lambda_i' \geq 0 \text{ for all } i, \quad \sum_{i=1}^{k} \lambda_i' = \sum_{i=1}^{k} \lambda_i - \lambda^* \sum_{i=1}^{k} \alpha_i = 1, \quad \text{and} \quad \sum_{i=1}^{k} \lambda_i' x_i = \sum_{i=1}^{k} \lambda_i x_i - \lambda^* \sum_{i=1}^{k} \alpha_i x_i = x.$$

But we know that at least one of the $\lambda_i' = 0$, so that we could write x as a convex combination of $k - 1$ elements. Repeating this strategy until $k = d + 1$ gives the theorem. □

B.1.1 Operations preserving convexity

We now touch on a few simple results about operations that preserve convexity of convex sets. First, we make the following simple observation.

Observation B.1.4. *Let C be a convex set. Then $C = \mathrm{Conv}(C)$.*

Observation B.1.4 is clear, as we have $C \subset \mathrm{Conv}(C)$, while any other convex $S \supset C$ clearly satisfies $S \supset \mathrm{Conv}(C)$. Secondly, we note that intersections preserve convexity.

Observation B.1.5. *Let $\{C_\alpha\}_{\alpha \in \mathcal{A}}$ be an arbitrary collection of convex sets. Then*

$$C = \bigcap_{\alpha \in \mathcal{A}} C_\alpha$$

is convex. Moreover, if C_α is closed for each α, then C is closed as well.

The convexity property follows because if $x_1 \in C$ and $x_2 \in C$, then clearly $x_1, x_2 \in C_\alpha$ for all $\alpha \in \mathcal{A}$, and moreover $\lambda x_1 + (1-\lambda)x_2 \in C_\alpha$ for all α and any $\lambda \in [0,1]$. The closure property is standard. In addition, we note that closing a convex set maintains convexity.

Observation B.1.6. *Let C be convex. Then $\mathrm{cl}(C)$ is convex.*

To see this, we note that if $x, y \in \mathrm{cl}(C)$ and $x_n \to x$ and $y_n \to y$ (where $x_n, y_n \in C$), then for any $\lambda \in [0,1]$, we have $\lambda x_n + (1-\lambda)y_n \in C$ and $\lambda x_n + (1-\lambda)y_n \to \lambda x + (1-\lambda)y$. Thus we have $\lambda x + (1-\lambda)y \in \mathrm{cl}(C)$ as desired.

Observation B.1.6 also implies the following result.

Observation B.1.7. *Let D be an arbitrary set. Then*

$$\bigcap \{C : C \supset D, \; C \text{ is convex}\} = \mathrm{cl}\,\mathrm{Conv}(D).$$

Proof Let T denote the leftmost set. It is clear that $T \subset \mathrm{cl}\,\mathrm{Conv}(D)$ as $\mathrm{cl}\,\mathrm{Conv}(D)$ is a closed convex set (by Observation B.1.6) containing D. On the other hand, if $C \supset D$ is a closed convex set, then $C \supset \mathrm{Conv}(D)$, while the closedness of C implies it also contains the closure of $\mathrm{Conv}(D)$. Thus $T \supset \mathrm{cl}\,\mathrm{Conv}(D)$ as well. □

JCD Comment: Picture

As our last consideration of operations that preserve convexity, we consider what is known as the perspective of a set. To define this set, we need to define the perspective function, which, given a point $(x,t) \in \mathbb{R}^d \times \mathbb{R}_{++}$ (here $\mathbb{R}_{++} = \{t : t > 0\}$ denotes strictly positive points), is defined as

$$\mathrm{pers}(x,t) = \frac{x}{t}.$$

We have the following definition.

Definition B.5. *Let $C \subset \mathbb{R}^d \times \mathbb{R}_+$ be a set. The* perspective transform *of C, denoted by* $\mathrm{pers}(C)$, *is*

$$\mathrm{pers}(C) := \left\{\frac{x}{t} : (x,t) \in C \text{ and } t > 0\right\}.$$

This corresponds to taking all the points $z \in C$, normalizing them so their last coordinate is 1, and then removing the last coordinate. (For more on perspective functions, see Boyd and Vandenberghe [35, Chapter 2.3.3].)

It is interesting to note that the perspective of a convex set is convex. First, we note the following.

Lemma B.1.8. *Let $C \subset \mathbb{R}^{d+1}$ be a compact line segment, meaning that $C = \{\lambda x + (1-\lambda)y : \lambda \in [0,1]\}$, where $x_{d+1} > 0$ and $y_{d+1} > 0$. Then $\mathrm{pers}(C) = \{\lambda\,\mathrm{pers}(x) + (1-\lambda)\,\mathrm{pers}(y) : \lambda \in [0,1]\}$.*

Proof Let $\lambda \in [0,1]$. Then

$$\begin{aligned}
\mathrm{pers}(\lambda x + (1-\lambda)y) &= \frac{\lambda x_{1:d} + (1-\lambda)y_{1:d}}{\lambda x_{d+1} + (1-\lambda)y_{d+1}} \\
&= \frac{\lambda x_{d+1}}{\lambda x_{d+1} + (1-\lambda)y_{d+1}} \frac{x_{1:d}}{x_{d+1}} + \frac{(1-\lambda)y_{d+1}}{\lambda x_{d+1} + (1-\lambda)y_{d+1}} \frac{y_{1:d}}{y_{d+1}} \\
&= \theta\,\mathrm{pers}(x) + (1-\theta)\,\mathrm{pers}(y),
\end{aligned}$$

where $x_{1:d}$ and $y_{1:d}$ denote the vectors of the first d components of x and y, respectively, and

$$\theta = \frac{\lambda x_{d+1}}{\lambda x_{d+1} + (1-\lambda) y_{d+1}} \in [0, 1].$$

Sweeping λ from 0 to 1 sweeps $\theta \in [0, 1]$, giving the result. □

Based on Lemma B.1.8, we immediately obtain the following proposition.

Proposition B.1.9. *Let $C \subset \mathbb{R}^d \times \mathbb{R}_{++}$ be a convex set. Then* $\operatorname{pers}(C)$ *is convex.*

Proof Let $x, y \in C$ and define $L = \{\lambda x + (1-\lambda)y : \lambda \in [0, 1]\}$ to be the line segment between them. By Lemma B.1.8, $\operatorname{pers}(L) = \{\lambda \operatorname{pers}(x) + (1-\lambda)\operatorname{pers}(y) : \lambda \in [0, 1]\}$ is also a (convex) line segment, and we have $\operatorname{pers}(L) \subset \operatorname{pers}(C)$ as necessary. □

B.1.2 Representation and separation of convex sets

JCD Comment: Put normal and tangent cones here

We now consider some properties of convex sets, showing that (1) they have nice separation properties—we can put hyperplanes between them—and (2) this allows several interesting representations of convex sets. We begin with the separation properties, developing them via the existence of projections. Interestingly, this existence of projections does not rely on any finite-dimensional structure, and can even be shown to hold in arbitrary Banach spaces (assuming the axiom of choice) [133]. We provide the results in a *Hilbert space*, meaning a complete vector space for which there exists an inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|$ given by $\|x\|^2 = \langle x, x \rangle$. We first note that projections exist.

Theorem B.1.10 (Projections)**.** *Let C be a closed convex set. Then for any x, there exists a unique point $\pi_C(x)$ minimizing $\|y - x\|$ over $y \in C$. Moreover, this point is characterized by the inequality*

$$\langle \pi_C(x) - x, y - \pi_C(x) \rangle \geq 0 \quad \text{for all } y \in C. \tag{B.1.2}$$

Proof The existence and uniqueness of the projection follows from the parallelogram identity, that is, that for any x, y we have $\|x - y\|^2 + \|x + y\|^2 = 2(\|x\|^2 + \|y\|^2)$, which follows by noting that $\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle$. Indeed, let $\{y_n\} \subset C$ be a sequence such that

$$\|y_n - x\| \to \inf_{y \in C} \|y - x\| =: p_\star$$

as $n \to \infty$, where $p_\star$ is the infimal value. We show that y_n is Cauchy, so that there exists a (unique) limit point of the sequence. Fix $\epsilon > 0$ and let N be such that $n \geq N$ implies $\|y_n - x\|^2 \leq p_\star^2 + \epsilon^2$. Let $m, n \geq N$. Then by the parallelogram identity,

$$\|y_n - y_m\|^2 = \|(x - y_n) - (x - y_m)\|^2 = 2\left[\|x - y_n\|^2 + \|x - y_m\|^2\right] - \|(x - y_n) + (x - y_m)\|^2.$$

Noting that

$$(x - y_n) + (x - y_m) = 2\left[x - \frac{y_n + y_m}{2}\right] \quad \text{and} \quad \frac{y_n + y_m}{2} \in C \text{ (by convexity of } C),$$

we have

$$\|x - y_n\|^2 \le p_\star^2 + \epsilon^2, \;\; \|x - y_m\|^2 \le p_\star^2 + \epsilon^2, \;\; \text{and} \;\; \|(x - y_n) + (x - y_m)\|^2 = 4 \left\|x - \frac{y_n + y_m}{2}\right\|^2 \ge 4p_\star^2.$$

In particular, we have

$$\|y_n - y_m\|^2 \le 2\left[p_\star^2 + \epsilon^2 + p_\star^2 + \epsilon^2\right] - 4p_\star^2 = 4\epsilon^2.$$

As $\epsilon > 0$ was arbitrary, this completes the proof of the first statement of the theorem.

To see the second result, assume that z is a point satisfying inequality (B.1.2), that is, such that

$$\langle z - x, y - z\rangle \ge 0 \;\; \text{for all } y \in C.$$

Then we have

$$\|z - x\|^2 = \langle z - x, z - x\rangle = \underbrace{\langle z - x, z - y\rangle}_{\le 0} + \langle z - x, y - x\rangle \le \|z - x\| \, \|y - x\|$$

by the Cauchy-Schwarz inequality. Dividing both sides by $\|z - x\|$ yields $\|z - x\| \le \|y - x\|$ for any $y \in C$, giving the result. Conversely, let $t \in [0, 1]$. Then for any $y \in C$,

$$\begin{aligned}\|\pi_C(x) - x\|^2 &\le \|(1 - t)\pi_C(x) + ty - x\|^2 = \|\pi_C(x) - x + t(y - \pi_C(x))\|^2 \\ &= \|\pi_C(x) - x\|^2 + 2t\langle \pi_C(x) - x, y - \pi_C(x)\rangle + t^2 \|y - \pi_C(x)\|^2 .\end{aligned}$$

Subtracting the projection value $\|\pi_C(x) - x\|^2$ from both sides and dividing by $t > 0$, we have

$$0 \le 2\langle \pi_C(x) - x, y - \pi_C(x)\rangle + t\|y - \pi_C(x)\|^2 .$$

Taking $t \to 0$ gives inequality (B.1.2). □

As an immediate consequence of Theorem B.1.10, we obtain several separation properties of convex sets, as well as a theorem stating that a closed convex set (not equal to the entire space in which it lies) can be represented as the intersection of all the half-spaces containing it.

Corollary B.1.11. *Let C be closed convex and $x \notin C$. Then there is a vector v strictly separating x from C, that is,*

$$\langle v, x\rangle > \sup_{y \in C} \langle v, y\rangle.$$

Moreover, we can take $v = x - \pi_C(x)$.

Proof By Theorem B.1.10, we know that taking $v = x - \pi_C(x)$ we have

$$0 \le \langle y - \pi_C(x), \pi_C(x) - x\rangle = \langle y - \pi_C(x), -v\rangle = \langle y - x + v, -v\rangle = -\langle y, v\rangle + \langle x, v\rangle - \|v\|^2 .$$

That is, we have $\langle v, y\rangle \le \langle v, x\rangle - \|v\|^2$ for all $y \in C$ and $v \ne 0$. □

In addition, we can show the existence of supporting hyperplanes, that is, hyperplanes "separating" the boundary of a convex set from itself.

Theorem B.1.12. *Let C be a convex set and $x \in \mathrm{bd}(C)$, where $\mathrm{bd}(C) = \mathrm{cl}(C) \setminus \mathrm{int}\, C$. Then there exists a non-zero vector v such that $\langle v, x\rangle \geq \sup_{y \in C} \langle v, y\rangle$.*

Proof Let $D = \mathrm{cl}(C)$ be the closure of C and let $x_n \notin D$ be a sequence of points such that $x_n \to x$. Let us define the sequence of separating vectors $s_n = x_n - \pi_D(x_n)$ and the normalized version $v_n = s_n / \|s_n\|$. Notably, we have $\langle v_n, x_n\rangle > \sup_{y \in C} \langle v_n, y\rangle$ for all n. Now, the sequence $\{v_n\} \subset \{v : \|v\| = 1\}$ belongs to a compact set.[1] Passing to a subsequence if necessary, let us assume w.l.o.g. that $v_n \to v$ with $\|v\| = 1$. Then by a standard limiting argument for the $x_n \to x$, we have

$$\langle v, x\rangle \geq \langle v, y\rangle \quad \text{for all } y \in C,$$

which was our desired result. □

JCD Comment: Picture of supporting hyperplanes and representations

Theorem B.1.12 gives us an important result. In particular, let D be an arbitrary set, and let $C = \mathrm{cl}\,\mathrm{Conv}(D)$ be the closure of the convex hull of D, which is the smallest closed convex set containing D. Then we can write C as the intersection of all the closed half-spaces containing D; this is, in some sense, the most useful "convexification" of D. Recall that a closed half-space H is defined with respect to a vector v and real $a \in \mathbb{R}$ as

$$H := \{x : \langle v, x\rangle \leq r\}.$$

Before stating the theorem, we remark that by Observation B.1.6, the intersection of all the closed convex sets containing a set D is equal to the closure of the convex hull of D.

Theorem B.1.13. *Let D be an arbitrary set. If $C = \mathrm{cl}\,\mathrm{Conv}(D)$, then*

$$C = \bigcap_{H \supset D} H, \tag{B.1.3}$$

where H denotes a closed half-space containing D. Moreover, for any closed convex set C,

$$C = \bigcap_{x \in \mathrm{bd}(C)} H_x, \tag{B.1.4}$$

where H_x denotes the intersection of halfspaces supporting C at x.

Proof We begin with the proof of the second result (B.1.4). Indeed, by Theorem B.1.12, we know that at each point x on the boundary of C, there exists a non-zero supporting hyperplane v, so that the half-space

$$H_{x,v} := \{y : \langle v, y\rangle \leq \langle v, x\rangle\} \supset C$$

is closed, convex, and contains C. We clearly have the containment $C \subset \cap_{x \in \mathrm{bd}(C)} H_x$. Now let $x_0 \notin C$; we show that $x_0 \notin \cap_{x \in \mathrm{bd}(C)} H_x$. As $x_0 \notin C$, the projection $\pi_C(x_0)$ of x_0 onto C satisfies $\langle x_0 - \pi_C(x_0), x_0\rangle > \sup_{y \in C} \langle x_0 - \pi_C(x_0), y\rangle$ by Corollary B.1.11. Moreover, letting $v = x_0 - \pi_C(x_0)$, the hyperplane

$$h_{x_0,v} := \{y : \langle y, v\rangle = \langle \pi_C(x_0), v\rangle\}$$

[1] In infinite dimensions, this may not be the case. But we can apply the Banach-Alaoglu theorem, which states that, as v_n are linear operators, the sequence is weak-* compact, so that there is a vector v with $\|v\| \leq 1$ and a subsequence $m(n) \subset \mathbb{N}$ such that $\langle v_{m(n)}, x\rangle \to \langle v, x\rangle$ for all x.

is clearly supporting to C at the point $\pi_C(x_0)$. The half-space $\{y : \langle y, v\rangle \le \langle \pi_C(x_0), v\rangle\}$ thus contains C and does not contain x_0, implying that $x_0 \notin \cap_{x\in \mathrm{bd}(C)} H_x$.

Now we show the first result (B.1.3). Let C be the closed convex hull of D and $T = \cap_{H\supset D} H$. By a trivial extension of the representation (B.1.4), we have that $C = \cap_{H\supset C} H$, where H denotes any halfspace containing C. As $C \supset D$, we have that $H \supset C$ implies $H \supset D$, so that

$$T = \bigcap_{H\supset D} H \subset \bigcap_{H\supset C} H = C.$$

On the other hand, as $C = \mathrm{cl}\,\mathrm{Conv}(D)$, Observation B.1.7 implies that any closed set containing D contains C. As a closed halfspace is convex and closed, we have that $H \supset D$ implies $H \supset C$, and thus $T = C$ as desired. □

B.2 Sublinear and support functions

A special case of convex functions will be sublinear functions, which form the basis of the transition between convex sets and convex functions. Accordingly, we give a special treatment here. Recall that f is convex if $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y)$ for all x, y.

Definition B.6. *A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is* sublinear *if it is convex and positively homogeneous, meaning*

$$f(tx) = tf(x) \quad \textit{for all } x \in \mathbb{R}^d \textit{ and } t > 0.$$

Such functions are important in that they give some of the first dualities between convex sets and convex functions. As we see in the section to come, they also allow us to describe various first-order smoothness properties of convex functions.

The main result we shall need on sublinear functions is that they can be defined by a dual construction.

Proposition B.2.1. *Let f be a closed sublinear function and define $S := \{s \mid \langle s, x\rangle \le f(x) \textit{ for all } x\}$. Then*

$$f(x) = \sup_{s\in S} \langle s, x\rangle.$$

Proof As f is closed convex, there exist affine functions minorizing f at each point in its domain (Theorem B.3.3). That is, for some pair $(s, t) \in \mathbb{R}^d \times \mathbb{R}$, we have $\langle s, x\rangle - t \le f(x)$ for all $x \in \mathbb{R}^d$. Because necessarily $f(0) = 0$ by sublinearity, we have $t \ge 0$, and by positive homogeneity, we have $\langle s, \alpha x\rangle - t \le f(\alpha x)$ for all $\alpha > 0$, that is, $\langle s, x\rangle - t/\alpha \le f(x)$ for all x. Taking $\alpha \uparrow \infty$ we find that

$$\langle s, x\rangle \le f(x) \quad \text{for all } x \in \mathbb{R}^d.$$

Because any closed convex function is the supremum of all affine functions minorizing it (Theorem B.3.7), we evidently have $f(x) = \sup_s\{\langle s, x\rangle \mid \langle s, \cdot\rangle \text{ minorizes } f\}$. □

To any set S we can associate a particular sublinear function, the *support function* of S, defining

$$\sigma_S(x) := \sup_{s\in S} \langle s, x\rangle. \tag{B.2.1}$$

This function is evidently a closed convex function—it is the supremum of linear functions—and is positively homogeneous, so that it is sublinear. We thus immediately have the duality

Corollary B.2.2. *Let f be a sublinear function. Then it is the support function of the closed convex set*

$$S_f := \{s \mid \langle s, x\rangle \le f(x) \text{ for all } x \in \mathbb{R}^d\},$$

and hence if C is closed convex, then

$$C = \{x \mid \langle s, x\rangle \le \sigma_C(s) \text{ for all } s \in \mathbb{R}^d\}.$$

A few other consequences of the definition are immediate. We see that σ_S has $\dom \sigma_S = \mathbb{R}^d$ if and only if S is bounded: whenever $\|s\| \le L$ for all $s \in S$, then $\sigma_S(x) \le L\|x\|$. Conversely, if $\dom \sigma_S = \mathbb{R}^d$ then it is locally Lipschitz (Theorem B.3.4) and (by positive homogeneity) thus globally Lipschitz, so we have $\langle s, x\rangle \le \sigma_S(x) \le L\|x\|$ for some $L < \infty$ and taking $x = s/\|s\|$ gives $\|s\| \le L$. As another consequence, we see that support functions of a set S are the support functions of the closed convex hull of S:

Proposition B.2.3. *Let $S \subset \mathbb{R}^d$. Then*

$$\sigma_S(x) = \sigma_{\cl \Conv S}(x).$$

Proof Let $C = \Conv S$, and let s_n be any sequence with $\langle s_n, x\rangle \to \sup_{s\in C}\langle s, x\rangle$. Then there exist $s_{n,i} \in S$, $i = 1, \ldots, k(n)$, such that $s_n = \sum_{i=1}^{k(n)} \lambda_i s_{n,i}$ for some $\lambda \succeq 0$, $\langle \lambda, \mathbf{1}\rangle = 1$, which may change with n. But of course, $\langle s_n, x\rangle \le \max_i \langle s_{n,i}, x\rangle$, and thus $\sigma_S(x) \ge \sigma_C(x)$. To see that $\sigma_C(x) = \sigma_{\cl C}(x)$, note that for each $\epsilon > 0$, for each $s \in \cl C$ there is $s' \in C$ with $\|s - s'\| < \epsilon$. Then $\langle s, x\rangle \le \langle s', x\rangle + \epsilon\|x\|$ and $\sigma_{\cl C}(x) \le \sigma_C(x) + \epsilon\|x\|$. Take $\epsilon \downarrow 0$. □

This proposition, coupled with Corollary B.2.2, shows that if sets S_1, S_2 have identical support functions, then they have identical closed convex hulls, and if they are closed convex, they are thus identical.

Corollary B.2.4. *Let $S_1, S_2 \subset \mathbb{R}^d$. If $\sigma_{S_1} = \sigma_{S_2}$, then* $\cl \Conv S_1 = \cl \Conv S_2$.

Proof By Proposition B.2.3, we have $\sigma_{S_i} = \sigma_{\cl \Conv S_i}$ for each i, and Corollary B.2.2 shows that if $\sigma_{C_1} = \sigma_{C_2}$ for closed convex sets C_1 and C_2, then $C_1 = C_2$. □

As another corollary, we have

Corollary B.2.5. *Let σ_1 and σ_2 be the support functions of the nonempty closed convex sets S_1 and S_2. Then if $t_1 > 0$ and $t_2 > 0$,*

$$t_1\sigma_1 + t_2\sigma_2 = \sigma_{\cl(t_1 S_1 + t_2 S_2)}.$$

If either of S_1 or S_2 is compact, then $t_1\sigma_1 + t_2\sigma_2 = \sigma_{t_1 S_1 + t_2 S_2}$.

Proof Let $S = t_1 S_1 + t_2 S_2$. In first statement, we have

$$\sigma_{\cl S}(x) \overset{(\star)}{=} \sigma_S(x) = \sup\left\{\langle t_1 s_1 + t_2 s_2, x\rangle \mid s_1 \in S_1, s_2 \in S_2\right\},$$

equality $(\star)$ following from Proposition B.2.3. As the suprema run independently through their respective sets S_1, S_2, the latter quantity is evidently

$$\sigma_S(x) = t_1 \sup_{s_1 \in S_1} \langle s_1, x\rangle + t_2 \sup_{s_2 \in S_2} \langle s_2, x\rangle = t_1\sigma_{S_1}(x) + t_2\sigma_{S_2}(x).$$

The final result is an immediate consequence of the result that if C is a compact convex set and S is closed convex, then $C + S$ is closed convex. That $C + S$ is convex is immediate. To see that it is closed, let $x_n \in C, y_n \in S$ satisfy $x_n + y_n \to z$. Then proceeding to a subsequence, we have $x_{n(m)} \to x_\infty$ for some $x_\infty \in C$, and thus $y_{n(m)} \to z - x_\infty$, which is then necessarily in S. As the subsequence $x_{n(m)} + y_{n(m)} \to x_\infty + (z - x_\infty) \in C + S$ and $x_{n(m)} + y_{n(m)} \to z$ as well, this gives the result. □

Linear transformations of support functions are also calculable. In the result, recall that for a matrix A and set S, the set $AS = \{As \mid s \in S\}$.

Proposition B.2.6. *Let $S \subset \mathbb{R}^d$ and $A \in \mathbb{R}^{m\times d}$. Then $\sigma_{\operatorname{cl} AS}(x) = \sigma_S(A^\top x)$.*

Proof We have $\sigma_{AS}(x) = \sup_{s\in S}\langle As, x\rangle = \sup_{s\in S}\langle s, A^T x\rangle$. The closure operation changes nothing (Proposition B.2.3). □

Lastly, we show how to use support functions to characterize whether sets have interiors. Recall that for a set $S \subset \mathbb{R}^d$, the affine hull $\operatorname{aff}(S)$ (Definition B.2) is the set of affine combinations of a point in S, and the relative interior of S is its interior relative to its affine hull (Definition B.4).

Proposition B.2.7. *Let $S \subset \mathbb{R}^d$ be non-empty a closed convex set. Then*

(i) *$s \in \operatorname{int} S$ if and only if $\langle s, x\rangle < \sigma_S(x)$ for all $x \neq 0$.*

(ii) *$s \in \operatorname{relint} S$ if and only if $\langle s, x\rangle < \sigma_S(x)$ for all x with $\sigma_S(x) + \sigma_S(-x) > 0$.*

(iii) *$\operatorname{int} S$ is non-empty if and only if $\sigma_S(x) + \sigma_S(-x) > 0$ for all $x \neq 0$.*

Proof

(i) Because σ_S is positively homogeneous, an equivalent statement is that $\sigma_S(x) > \langle s, x\rangle$ for all $x \in \mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$. If $s \in \operatorname{int} S$, we there exists $\epsilon > 0$ such that $s + \epsilon x \in S$ for all $x \in \mathbb{S}^{d-1}$, and so

$$\sigma_S(x) \geq \langle s + \epsilon x, x\rangle = \langle s, x\rangle + \epsilon,$$

so that $\langle s, x\rangle < \sigma_S(x)$.

Conversely, let s be any point satisfying $\sigma_S(x) - \langle s, x\rangle > 0$ for all $x \in \mathbb{S}^{d-1}$. Because σ_S is lower semicontinuous, the infimum $\inf_{x\in\mathbb{S}^{d-1}}\{\sigma_S(x) - \langle s, x\rangle\}$ is attained at some $x^\star \in \mathbb{S}^{d-1}$ (see Proposition C.0.1). Then there exists some $\epsilon > 0$ such that $\langle s, x\rangle + \epsilon \leq \sigma_S(x)$ for all $x \in \mathbb{S}^{d-1}$. Let u be any vector with $\|u\|_2 < \epsilon$. Then $\langle s + u, x\rangle = \langle s, x\rangle + \langle u, x\rangle \leq \langle s, x\rangle + \epsilon \leq \sigma_S(x)$, so Corollary B.2.2 implies $s + u \in S$ and $s \in \operatorname{int} S$.

(ii) We decompose $\mathbb{R}^d$ into subspaces $V \oplus U$, where $U = V^\perp$ and V is parallel to $\operatorname{aff}(S)$. Writing $x = x_U + x_V$, where $x_U \in U$ and $x_V \in V$, the function $\langle s, x_U\rangle$ is constant for $s \in S$. Repeat the argument for part (i) in the subspace V.

(iii) Suppose $\operatorname{int} S$ is non-empty. Then $s \in \operatorname{int} S$ implies $\langle s, x\rangle < \sigma_S(x)$ for all x with $\|x\| = 1$. Then $\sigma_S(x) + \sigma_S(-x) > \langle s, x - x\rangle = 0$. Conversely, if $\operatorname{int} S$ is empty, there exists a hyperplane containing S (by a dimension counting argument and that the relative interior of S is never empty [104, Theorem III.2.1.3]), which we may write as $S \subset \{s \mid v^T s = b\}$ for some $v \neq 0$. For this $\sigma_S(v) + \sigma_S(-v) = b - b = 0$.

□

B.3 Convex functions

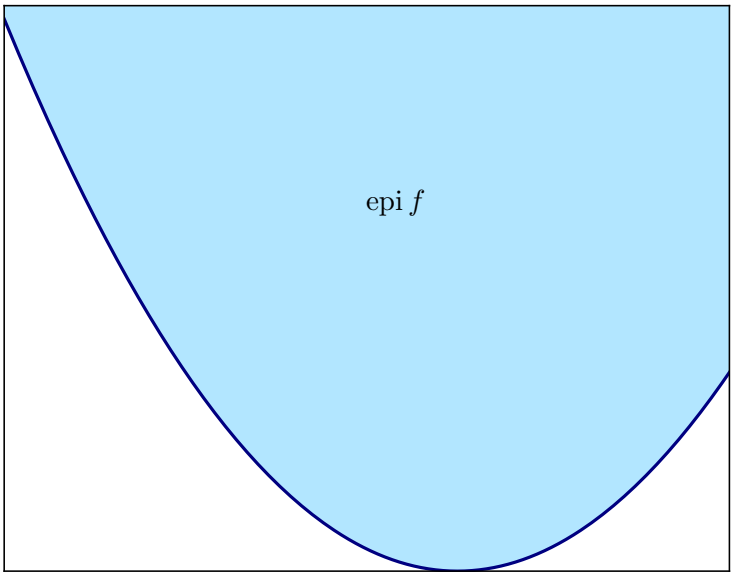

Figure B.1: The epigraph of a convex function.

We now build off of the definitions of convex sets to define convex functions. As we will see, convex functions have several nice properties that follow from the geometric (separation) properties of convex sets. First, we have

Definition B.7. *A function f is* convex *if for all $\lambda \in [0, 1]$ and $x, y \in \operatorname{dom} f$,*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y). \tag{B.3.1}$$

We define the domain $\operatorname{dom} f$ of a convex function to be those points x such that $f(x) < +\infty$. Note that Definition B.7 implies that the domain of f must be convex.

An equivalent definition of convexity follows by considering a natural convex set attached to the function f, known as its epigraph.

Definition B.8. *The epigraph* $\operatorname{epi} f$ *of a function is the set*

$$\operatorname{epi} f := \{(x, t) : t \in \mathbb{R}, f(x) \leq t\}.$$

That is, the epigraph of a function f is the set of points on or above the graph of the function itself, as depicted in Figure B.1. It is immediate from the definition of the epigraph that f is convex if and only if $\operatorname{epi} f$ is convex. Thus, we see that any convex set $C \subset \mathbb{R}^{d+1}$ that is unbounded "above," meaning that $C = C + \{0\} \times \mathbb{R}_+$, defines a convex function, and conversely, any convex function defines such a set C. This duality in the relationship between a convex function and its epigraph is central to many of the properties we exploit.

B.3.1 Equivalent definitions of convex functions

We begin our discussion of convex functions by enumerating a few standard properties that also characterize convexity. The simplest of these relate to properties of the derivatives and second derivatives of functions. We begin by elucidating one of the most basic properties of convexity: that the slopes of convex functions are increasing. Beginning with functions on $\mathbb{R}$, suppose that $f : \mathbb{R} \to \mathbb{R}$ is convex, and let $x \in \dom f$ and $v \in \mathbb{R}$ be otherwise arbitrary. Then define the quotient function

$$q(t) := \frac{f(x + tv) - f(x)}{t}, \quad t \geq 0, \tag{B.3.2}$$

which we claim is nondecreasing in $t \geq 0$ if and only if f is convex. Indeed, let $t \geq s > 0$ and define $\lambda = \frac{s}{t} \in [0, 1]$. Then

$$\begin{aligned} q(t) \geq q(s) \quad & \text{if and only if} \quad \lambda[f(x + tv) - f(x)] \geq f(x + \lambda tv) - f(x) \\ & \text{if and only if} \quad \lambda f(x + tv) + (1 - \lambda)f(x) \geq f((1 - \lambda)x + \lambda(x + tv)), \end{aligned}$$

the latter holding for all λ if and only if f is convex.

JCD Comment: Draw a picture of increasing quotient

Because the quotient function (B.3.2) is nondecreasing, we can relatively straightforwardly give first-order characterizations of convexity as well. Indeed, suppose that $f : \mathbb{R} \to \mathbb{R}$ is differentiable; then convexity is equivalent to the first-order inequality that for all $x, y \in \mathbb{R}$, we have

$$f(y) \geq f(x) + f'(x)(y - x). \tag{B.3.3}$$

To see that inequality (B.3.3) implies that f is convex follows from algebraic manipulations: let $\lambda \in [0, 1]$ and $z = \lambda x + (1 - \lambda)y$, so that $y - z = \lambda(y - x)$ and $x - z = (1 - \lambda)(x - y)$. Then

$$f(y) \geq f(z) + \lambda f'(z)(y - x) \quad \text{and} \quad f(x) \geq f(z) + (1 - \lambda)f'(z)(x - y),$$

and multiplying the former by $(1 - \lambda)$ and the latter by λ and adding the two inequalities yields

$$\lambda f(x)+(1-\lambda)f(y) \geq \lambda f(z)+(1-\lambda)f(z)+\lambda(1-\lambda)f'(z)(y-x)+\lambda(1-\lambda)f'(z)(x-y) = f(\lambda x+(1-\lambda)y),$$

as desired. Conversely, let $v = y - x$ in the quotient (B.3.2), so that $q(t) = \frac{f(x+tv)-f(x)}{t}$, which is non-decreasing. If f is differentiable, we see that $q(0) := \lim_{t \downarrow 0} q(t) = f'(x)(y - x)$, and so

$$q(1) = f(y) - f(x) \geq q(0) = f'(x)(y - x)$$

as desired.

We may also give the standard second order characterization: if $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable and $f''(x) \geq 0$ for all x, then f is convex. To see this, note that

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2}f''(tx + (1 - t)y)(x - y)^2$$

for some $t \in [0, 1]$ by Taylor's theorem, so that $f(y) \geq f(x) + f'(x)(y - x)$ for all x, y because $f''(tx + (1 - t)y) \geq 0$. As a consequence, we obtain inequality (B.3.3), which implies that f is convex.

As convexity is a property that depends only on properties of functions on lines—one dimensional projections—we can straightforwardly extend the preceding results to functions $f : \mathbb{R}^d \to \mathbb{R}$.

Indeed, noting that if $h(t) = f(x + ty)$ then $h'(0) = \langle \nabla f(x), y \rangle$ and $h''(0) = y^\top \nabla^2 f(x) y$, we have that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \quad \text{for all } x, y,$$

while a twice differentabile function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x.$$

Noting that nothing in the derivation that the quotient (B.3.2) was non-decreasing relied on f being a function on $\mathbb{R}$, we can see that a function $f : \mathbb{R}^d$ is convex if and only if it satisfies the *increasing slopes* criterion: for all $x \in \operatorname{dom} f$ and any vector v, the quotient

$$t \mapsto q(t) := \frac{f(x + tv) - f(x)}{t} \tag{B.3.4}$$

is nondecreasing in $t \geq 0$ (where we leave x, v implicit). An alternative version of the criterion (B.3.4) is that if $x \in \operatorname{dom} f$ and v is any vector, if we define the one-dimensional convex function $h(t) = f(x + tv)$ then for any $s < t$ and $\Delta > 0$, we have

$$\frac{h(t + \Delta) - h(t)}{\Delta} \geq \frac{h(t) - h(s)}{t - s} \geq \frac{h(t) - h(s - \Delta)}{t - (s - \Delta)}. \tag{B.3.5}$$

The proof that either of the inequalities (B.3.5) is equivalent to convexity we leave as an exercise (Q. C.1).

JCD Comment: Draw pictures of increasing slopes

We summarize each of these implications in a theorem for reference.

Proposition B.3.1 (Convexity)**.** *The following are all equivalent:*

(i) The function f is convex.

(ii) The function f satisfies the criterion of increasing slopes (B.3.4)*.*

If f is differentiable (respectively, twice differentiable), the following are also equivalent:

(iii) The function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \text{for all } x, y.$$

(iv) The function f has positive semidefinite Hessian: $\nabla^2 f(x) \succeq 0$ for all x.

JCD Comment: Draw a picture and of strict convexity

A condition slightly stronger than convexity is *strict convexity*, which makes each of the inequalities in Proposition B.3.1 strict. We begin with the classical definition: a function f is strictly convex if it is convex and

$$f(\lambda x + (1 - \lambda) y) < \lambda f(x) + (1 - \lambda) f(y)$$

whenever $\lambda \in (0, 1)$ and $x \neq y \in \operatorname{dom} f$. These are convex functions, but always have strictly increasing slopes—secants lie strictly above f. By tracing through the arguments leading to Proposition B.3.1 (replace appropriate non-strict inequalities with strict inequalities), one obtains the following corollary describing strictly convex functions.

Corollary B.3.2 (Strict convexity)**.** *The following are all equivalent:*

(i) The function f is strictly convex.

(ii) The function f has strictly increasing slopes (B.3.4)*.*

If f is differentiable (respectively, twice differentiable), the following are also equivalent:

(iii) The function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$f(y) > f(x) + \langle \nabla f(x), y - x \rangle \quad \text{for all } x \neq y.$$

(iv) The function f has positive definite Hessian: $\nabla^2 f(x) \succ 0$ for all x.

B.3.2 Continuity properties of convex functions

We now consider a few continuity properties of convex functions and a few basic relationships of the function f to its epigraph. First, we give a definition of the *subgradient* of a convex function.

Definition B.9. *A vector g is a* subgradient *of f at a point x_0 if for all x,*

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle. \tag{B.3.6}$$

The subdifferential *or* subgradient set *of f at x_0 is*

$$\partial f(x_0) := \{g \mid f(x) \geq f(x_0) + \langle g, x - x_0 \rangle \text{ for all } x\}.$$

See Figure B.2 for an illustration of the affine minorizing function given by the subgradient of a convex function at a particular point.

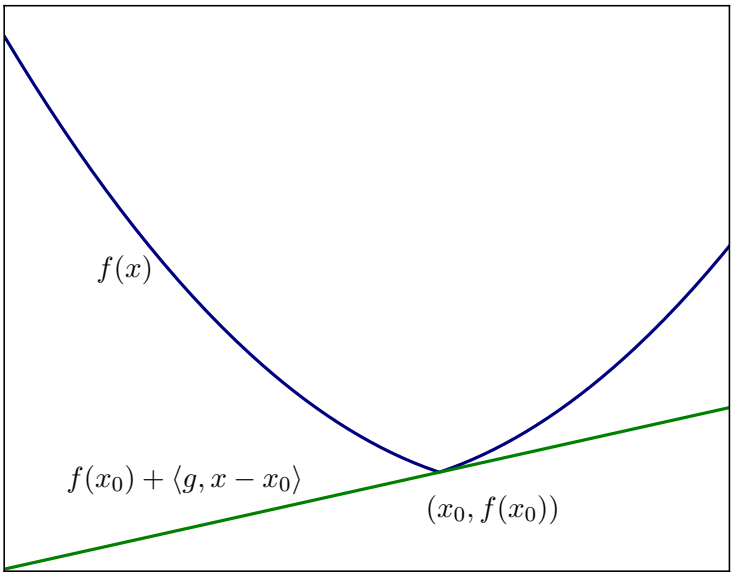

Figure B.2. The tangent (affine) function to the function f generated by a subgradient g at the point x_0.

Interestingly, convex functions have subgradients (at least, nearly everywhere). This is perhaps intuitively obvious by viewing a function in conjunction with its epigraph epi f and noting that epi f has supporting hyperplanes, but here we state a result that will have further use.

Theorem B.3.3. *Let f be convex. Then there is an affine function minorizing f. More precisely, for any $x_0 \in \operatorname{relint} \operatorname{dom} f$, there exists a vector g such that*

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle.$$

Proof If $\operatorname{relint} \operatorname{dom} f = \emptyset$, then it is clear that f is either identically $+\infty$ or its domain is a single point $\{x_0\}$, in which case the constant function $f(x_0)$ minorizes f. Now, we assume that $\operatorname{int} \operatorname{dom} f \neq \emptyset$, as we can simply always change basis to work in the affine hull of $\operatorname{dom} f$.

We use Theorem B.1.12 on the existence of supporting hyperplanes to construct a subgradient. Indeed, we note that $(x_0, f(x_0)) \in \operatorname{bd} \operatorname{epi} f$, as for any open set O we have that $(x_0, f(x_0)) + O$ contains points both inside and outside of $\operatorname{epi} f$. Thus, Theorem B.1.12 guarantees the existence of a vector v and $a \in \mathbb{R}$, not both simultaneously zero, such that

$$\langle v, x_0 \rangle + af(x_0) \leq \langle v, x \rangle + at \quad \text{for all } (x, t) \in \operatorname{epi} f. \tag{B.3.7}$$

Inequality (B.3.7) implies that $a \geq 0$, as for any x we may take $t \to +\infty$ while satisfying $(x, t) \in \operatorname{epi} f$. Now we argue that $a > 0$ strictly. To see this, note that for suitably small $\delta > 0$, we have $x = x_0 - \delta v \in \operatorname{dom} f$. Then we find by inequality (B.3.7) that

$$\langle v, x_0 \rangle + af(x_0) \leq \langle v, x_0 \rangle - \delta \|v\|^2 + af(x_0 - \delta v), \quad \text{or} \quad a\left[f(x_0) - f(x_0 - \delta v)\right] \leq -\delta \|v\|^2.$$

So if $v = 0$, then Theorem B.1.12 already guarantees $a \neq 0$, while if $v \neq 0$, then $\|v\|^2 > 0$ and we must have $a \neq 0$ and $f(x_0) \neq f(x_0 - \delta v)$. As we showed already that $a \geq 0$, we must have $a > 0$. Then by setting $t = f(x_0)$ and dividing both sides of inequality (B.3.7) by a, we obtain

$$\frac{1}{a}\langle v, x_0 - x \rangle + f(x_0) \leq f(x) \quad \text{for all } x \in \operatorname{dom} f.$$

Setting $g = -v/a$ gives the result of the theorem, as we have $f(x) = +\infty$ for $x \notin \operatorname{dom} f$. □

Convex functions generally have quite nice behavior. Indeed, they enjoy some quite remarkable continuity properties just by virtue of the defining convexity inequality (B.3.1). In particular, the following theorem shows that convex functions are continuous on the relative interiors of their domains. Even more, convex functions are Lipschitz continuous on any compact subsets contained in the (relative) interior of their domains. (See Figure B.3 for an illustration of this fact.)

Theorem B.3.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and $C \subset \operatorname{relint} \operatorname{dom} f$ be compact. Then there exists an $L = L(C) \geq 0$ such that*

$$|f(x) - f(x')| \leq L \left\|x - x'\right\|.$$

As an immediate consequence of Theorem B.3.4, we note that if $f : \mathbb{R}^d \to \mathbb{R}$ is convex and defined everywhere on $\mathbb{R}^d$, then it is continuous. Moreover, we also have that $f : \mathbb{R}^d \to \mathbb{R}$ is continuous everywhere on the (relative) interior of its domain: let any $x_0 \in \operatorname{relint} \operatorname{dom} f$. Then for small enough $\epsilon > 0$, the set $\operatorname{cl}(\{x_0 + \epsilon B\} \cap \operatorname{dom} f)$, where $B = \{x : \|x\|_2 \leq 1\}$, is a closed and bounded—and hence compact—set contained in the (relative) interior of $\operatorname{dom} f$. Thus f is Lipschitz on this set, which is a neighborhood of x_0. In addition, if $f : \mathbb{R} \to \mathbb{R}$, then f is continuous everywhere except (possibly) at the endpoints of its domain.

Proof of Theorem B.3.4 To prove the theorem, we require a technical lemma.

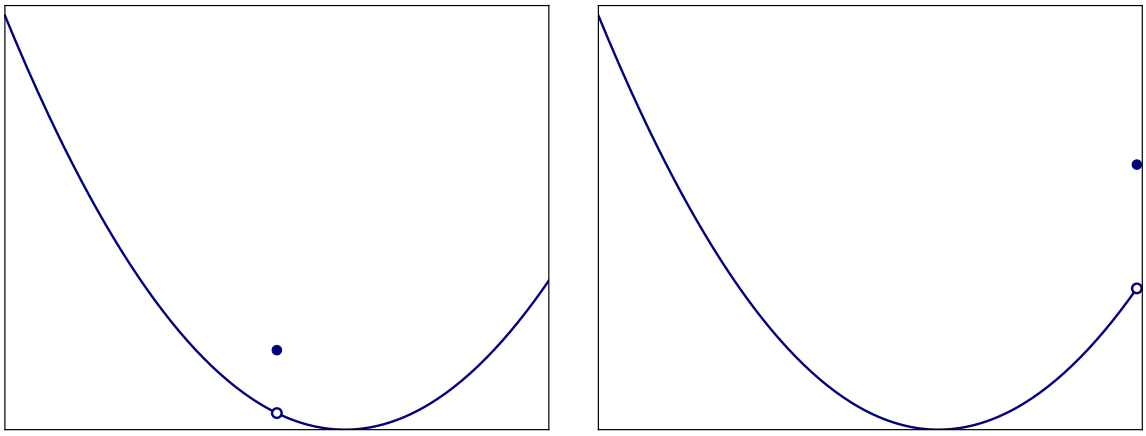

Figure B.3. Left: discontinuities in $\operatorname{int}\operatorname{dom} f$ are impossible while maintaining convexity (Theorem B.3.4). Right: At the edge of $\operatorname{dom} f$, there may be points of discontinuity.

Lemma B.3.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and suppose that there are x_0, $\delta > 0$, m, and M such that*

$$m \le f(x) \le M \quad \text{for } x \in B(x_0, 2\delta) := \{x : \|x - x_0\| < 2\delta\}.$$

Then f is Lipschitz on $B(x_0, \delta)$, and moreover,

$$|f(y) - f(y')| \le \frac{M - m}{\delta} \left\|y - y'\right\| \quad \text{for } y, y' \in B(x_0, \delta).$$

Proof Let $y, y' \in B(x_0, \delta)$, and define $y'' = y' + \delta(y' - y)/\left\|y' - y\right\| \in B(x_0, 2\delta)$. Then we can write y' as a convex combination of y and y'', specifically,

$$y' = \frac{\|y' - y\|}{\delta + \|y' - y\|} y'' + \frac{\delta}{\delta + \|y' - y\|} y.$$

Thus we obtain by convexity

$$\begin{aligned} f(y') - f(y) \le \frac{\|y' - y\|}{\delta + \|y' - y\|} f(y'') + \frac{\delta}{\delta + \|y' - y\|} f(y) - f(y) &= \frac{\|y - y'\|}{\delta + \|y - y'\|} [f(y'') - f(y)] \\ &\le \frac{M - m}{\delta + \|y - y'\|} \left\|y - y'\right\|. \end{aligned}$$

Here we have used the bounds on f assumed in the lemma. Swapping the assignments of y and y' gives the same lower bound, thus giving the desired Lipschitz continuity. □

With Lemma B.3.5 in place, we proceed to the proof proper. We assume without loss of generality that $\operatorname{dom} f$ has an interior; otherwise we prove the theorem restricting ourselves to the affine hull of $\operatorname{dom} f$. The proof follows a standard compactification argument. Suppose that for each $x \in C$, we could construct an open ball $B_x = B(x, \delta_x)$ with $\delta_x > 0$ such that

$$|f(y) - f(y')| \le L_x \left\|y - y'\right\| \quad \text{for } y, y' \in B_x. \tag{B.3.8}$$

As the B_x cover the compact set C, we can extract a finite number of them, call them $B_{x_1}, \ldots, B_{x_k}$, covering C, and then within each (overlapping) ball f is $\max_k L_{x_k}$ Lipschitz. As a consequence, we find that

$$|f(y) - f(y')| \le \max_k L_{x_k} \left\|y - y'\right\|$$

for any $y, y' \in C$.

We thus must derive inequality (B.3.8), for which we use the boundedness Lemma B.3.5. We must demonstrate that f is bounded in a neighborhood of each $x \in C$. To that end, fix $x \in \operatorname{int} \operatorname{dom} f$, and let the points $x_0, \ldots, x_d$ be affinely independent and such that

$$\Delta := \operatorname{Conv}\{x_0, \ldots, x_d\} \subset \operatorname{dom} f$$

and $x \in \operatorname{int} \Delta$; let $\delta > 0$ be such that $B(x, 2\delta) \subset \Delta$. Then by Carathéodory's theorem (Theorem B.1.3) we may write any point $y \in B(x, 2\delta)$ as $y = \sum_{i=0}^d \lambda_i x_i$ for $\sum_i \lambda_i = 1$ and $\lambda_i \ge 0$, and thus

$$f(y) \le \sum_{i=0}^d \lambda_i f(x_i) \le \max_{i \in \{0,\ldots,d\}} f(x_i) =: M.$$

Moreover, Theorem B.3.3 implies that there is some affine h function minorizing f; let $h(x) = a + \langle v, x \rangle$ denote this function. Then

$$m := \inf_{x \in C} f(x) \ge \inf_{x \in C} h(x) = a + \inf_{x \in C} \langle v, x \rangle > -\infty$$

exists and is finite, so that in the ball $B(x, 2\delta)$ constructed above, we have $f(y) \in [m, M]$ as required by Lemma B.3.5. This guarantees the existence of a ball B_x required by inequality (B.3.8). □

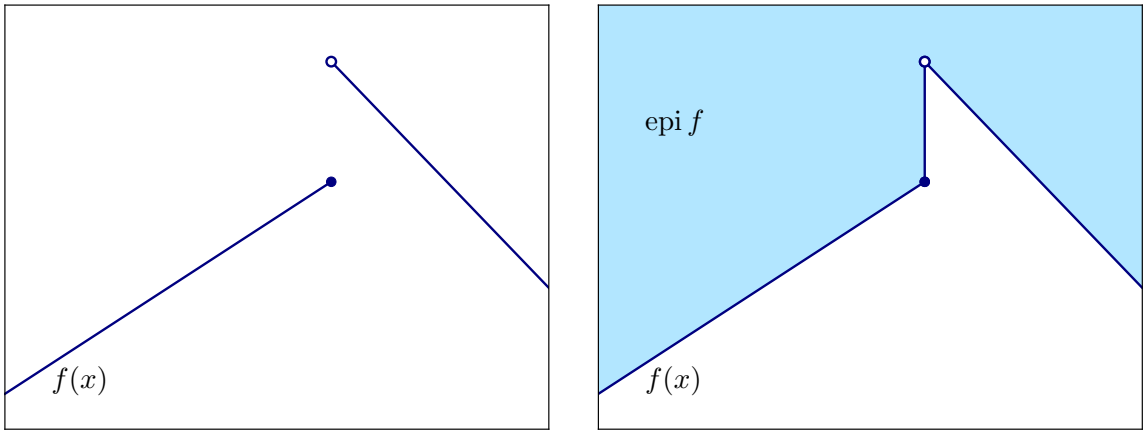

Figure B.4. A closed—equivalently, lower semi-continuous—function. On the right is shown the *closed* epigraph of the function.

Our final discussion of continuity properties of convex functions revolves around the most common and analytically convenient type of convex function, the so-called *closed-convex* functions.

Definition B.10. *A function f is* closed *if its epigraph,* $\operatorname{epi} f$*, is a closed set.*

Equivalently, a function is closed if it is lower semi-continuous, meaning that

$$\liminf_{x \to x_0} f(x) \geq f(x_0) \tag{B.3.9}$$

for all x_0 and any sequence of points tending toward x_0. See Figure B.4 for an example such function and its associated epigraph.

Interestingly, in the one-dimensional case, closed convexity implies continuity. Indeed, we have the following observation (compare Figures B.4 and B.3 previously):

Observation B.3.6. *Let $f : \mathbb{R} \to \mathbb{R}$ be a closed convex function. Then f is continuous on its domain, and for any $x_0 \in \operatorname{bd} \operatorname{dom} f$, $\lim_{x \to x_0} f(x) = f(x_0)$ whether or not $x_0 \in \operatorname{dom} f$.*

Proof By Theorem B.3.4, we need only consider the endpoints of the domain of f (the result is obvious by Theorem B.3.4 if $\operatorname{dom} f = \mathbb{R}$); let $x_0 \in \operatorname{bd} \operatorname{dom} f$. Let $y \in \operatorname{dom} f$ be an otherwise arbitrary point, and define $x = \lambda y + (1 - \lambda) x_0$. Then taking $\lambda \to 0$, we have

$$f(x) \leq \lambda f(y) + (1 - \lambda) f(x_0) \to f(x_0),$$

so that $\limsup_{x \to x_0} f(x) \leq f(x_0)$. By the closedness assumption (B.3.9), we have $\liminf_{x \to x_0} f(x) \geq f(x_0)$, and continuity follows. Note that in this argument, if $x_0 \notin \operatorname{dom} f$, then $f(x_0) = +\infty$ by convention; for $\operatorname{epi} f$ to be closed we require that for each $t < f(x_0) = \infty$, we may take a small enough open interval $U = (y, x_0)$ for which $f(x) > t$ for all $x \in U$. $\square$

In the full-dimensional case, we do not have quite the same continuity, though Theorem B.3.4 guarantees continuity on the (relative) interior of $\operatorname{dom} f$.

An important characterization of convex functions is as the supremum of all affine functionals (linear plus an offset) below them, which is one of the keys to duality relationships about functions to come.

Theorem B.3.7. *Let f be closed convex and let $\mathcal{A}$ be the collection of affine functions h satisfying $f(x) \geq h(x)$ for all x. Then $f(x) = \sup_{h \in \mathcal{A}} h(x)$.*

Proof By Theorem B.1.13 that any closed convex set is the intersection of all the halfspaces containing (even supporting) it, we can write $\operatorname{epi} f = \cap_{H \in \mathcal{H}} H$, where $\mathcal{H}$ is the collection of closed halfspaces $H \supset \operatorname{epi} f$. We may write any such halfspace as

$$H = \{(x, r) \in \mathbb{R}^d \times \mathbb{R} \mid \langle a, x \rangle + br \leq c\}$$

where $(a, b) \in \mathbb{R}^d \times \mathbb{R}$ is non-zero. As $H \supset \operatorname{epi} f$, the particular nature of epigraphs (that is, that if $(x, t) \in \operatorname{epi} f$ then $(x, t + \Delta) \in \operatorname{epi} f$ for all $\Delta > 0$) means that $b \leq 0$, and so for any $b < 0$ we may divide through by b to rewrite H as $H = \{(x, r) \mid \langle a/b, x \rangle + r \geq c/b\}$, while if $b = 0$ then $H = \{(x, r) \mid \langle a, x \rangle \leq c\}$. That is, it is no loss of generality to set

$$\begin{aligned} \mathcal{H}_1 &:= \{\text{Halfspaces } \{(x, r) \mid \langle a, x \rangle + r \geq c\} \text{ containing } \operatorname{epi} f\} \\ \mathcal{H}_0 &:= \{\text{Halfspaces } \{(x, r) \mid \langle a, x \rangle \geq c\} \text{ containing } \operatorname{epi} f\}, \end{aligned}$$

which (respectively) correspond to the non-vertical halfspaces containing $\operatorname{epi} f$ and the halfspaces containing $\operatorname{dom} f \subset \mathbb{R}^d$. We have $\operatorname{epi} f = \bigcap_{H \in \mathcal{H}_1} H \cap \bigcap_{H \in \mathcal{H}_0} H$.

Identify the halfspaces $H \in \mathcal{H}_0$ or $\mathcal{H}_1$ with the associated triple $(a, 0, c)$ or $(a, 1, c)$ and abuse notation to write $(a, i, c) \in \mathcal{H}_i$ for $i \in \{0, 1\}$. For any $(a, 1, c) \in \mathcal{H}_1$, the linear function

$$l(x) = c - \langle a, x\rangle = \inf\{r \mid \langle a, x\rangle + r \geq c\} \;\; \text{satisfies} \;\; \langle a, x\rangle + l(x) \geq c \text{ for all } x,$$

and so necessarily $l(x) \leq f(x)$ for all x, while for the function $h(x) = \sup_{(a,1,c)\in\mathcal{H}_1}\{c - \langle a, x\rangle\}$ we have

$$\operatorname{epi} h = \bigcap_{H\in\mathcal{H}_1} H.$$

Thus, if we can show that

$$\bigcap_{H\in\mathcal{H}_1} H \cap \bigcap_{H\in\mathcal{H}_0} H = \bigcap_{H\in\mathcal{H}_1} H \tag{B.3.10}$$

the proof will be complete.

To show the equality (B.3.10), take arbitrary vectors $v_0 = (a_0, 0, c_0) \in \mathcal{H}_0$ and $v_1 = (a_1, 1, c_1) \in \mathcal{H}_1$, and let $H_0 = \{(x, r) \mid \langle a_0, x\rangle \geq c_0\}$ and $H_1 = \{(x, r) \mid \langle a_1, x\rangle + r \geq c_1\}$ be the associated halfspaces. Consider the conic-like vector

$$v(t) := (a_1 + ta_0, 1, c_0 + tc_0) \quad \text{for } t \geq 0$$

and associated halfspace $H(t) := \{(x, r) \mid \langle a_1 + ta_0, x\rangle + r \geq c_1 + tc_0\}$. Then as $\langle a_0, x\rangle \geq c_0$ if and only if $t\langle a_0, x\rangle \geq tc_0$ for all $t \geq 0$, any point $(x, r) \in H_0 \cap H_1$ satisfies

$$\langle a_1 + ta_0, x\rangle + r \geq c_1 + tc_0 \;\; \text{for } t \geq 0,$$

that is, $H(t) \in \mathcal{H}_1$ and $(x, r) \in \cap_{t\geq 0} H(t)$. Additionally, taking $t = 0$ we see that $H(0) = H_1$ and so $\cap_{t\geq 0} H(t) \subset H_1$, while taking $t \uparrow \infty$ we obtain that each $(x, r) \in \cap_{t\geq 0} H(t)$ satisfies $\langle a_0, x\rangle \geq c_0$. That is, we have

$$\bigcap_{t\geq 0} H(t) = H_0 \cap H_1,$$

while $H(t) \in \mathcal{H}_1$ for all $t \geq 0$. This shows the equality (B.3.10). □

JCD Comment: Show a picture of the above argument

In spite of the continuity of closed convex functions on $\mathbb{R}$, closed convex functions on higher dimensional spaces need not be continuous. Indeed, it is immediate (see Proposition B.3.9 to follow) that $f(x) := \sup_{\alpha\in\mathcal{A}}\{f_\alpha(x)\}$ is closed convex whenever f_α are all closed convex for any index set $\mathcal{A}$. We have the following failure of continuity.

Example B.3.8 (A discontinuous closed convex function)**:** Define the function $f : \mathbb{R}^2 \to \overline{\mathbb{R}}$ by

$$f(x) := \sup\left\{\alpha x_1 + \beta x_2 \mid \frac{1}{2}\alpha^2 \leq \beta\right\}.$$

Then certainly $f(\mathbf{0}) = 0$ and f is closed convex. If the supremum is attained then $\beta = \frac{1}{2}\alpha^2$ and so $\beta \geq 0$ and

$$f(x) = \sup_\alpha\left\{\alpha x_1 + \frac{1}{2}\alpha^2 x_2\right\} = \begin{cases} 0 & \text{if } x = \mathbf{0} \\ -\frac{x_1^2}{2x_2} & \text{if } x_2 < 0 \\ +\infty & \text{otherwise.} \end{cases}$$

But then along the path $x_2 = -\frac{1}{2}x_1^2$, we always have $f(x) = 1$, while taking $x_1 \to 0$ gives $f(x) = 1 > 0 = f(\mathbf{0})$. ◇

B.3.3 Operations preserving convexity

We now turn to a description of a few simple operations on functions that preserve convexity. First, we extend the intersection properties of convex sets to operations on convex functions. (See Figure B.5 for an illustration of the proposition.)

Proposition B.3.9. *Let $\{f_\alpha\}_{\alpha\in\mathcal{A}}$ be an arbitrary collection of convex functions indexed by $\mathcal{A}$. Then*

$$f(x) := \sup_{\alpha\in\mathcal{A}} f_\alpha(x)$$

is convex. Moreover, if for each $\alpha \in \mathcal{A}$, the function f_α is closed convex, f is closed convex.

Proof The proof is immediate once we consider the epigraph $\operatorname{epi} f$. We have that

$$\operatorname{epi} f = \bigcap_{\alpha\in\mathcal{A}} \operatorname{epi} f_\alpha,$$

which is convex whenever $\operatorname{epi} f_\alpha$ is convex for all α and closed whenever $\operatorname{epi} f_\alpha$ is closed for all α (recall Observation B.1.5). □

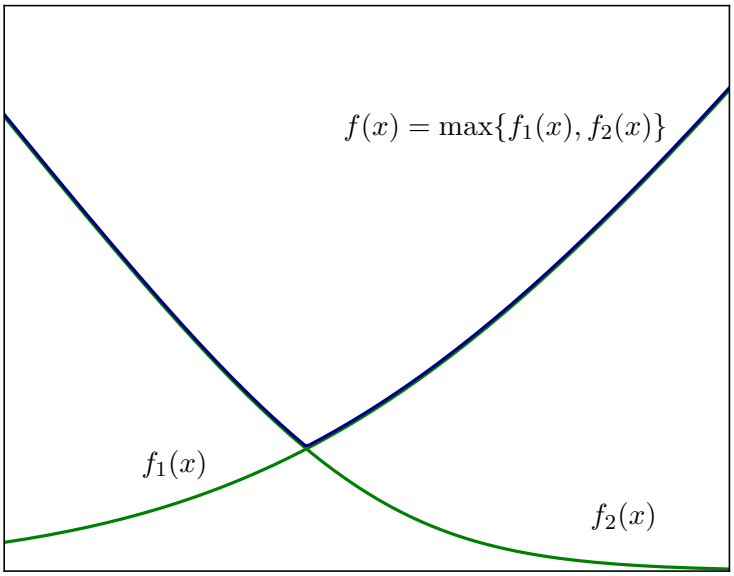

Figure B.5. The maximum of two convex functions is convex, as its epigraph is the intersection of the two epigraphs.

Another immediate result is that composition of a convex function with an affine transformation preserves convexity:

Proposition B.3.10. *Let $A \in \mathbb{R}^{d\times n}$ and $b \in \mathbb{R}^d$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then the function $g(y) = f(Ay + b)$ is convex.*

Partial minimization of convex functions and some related transformations preserve convexity as well.

Proposition B.3.11. *Let $A \in \mathbb{R}^{d\times n}$, $f : \mathbb{R}^n \to \mathbb{R}$ be convex, and $Y \subset \mathbb{R}^d$ be convex. Then $g(x) = \inf\{f(y) \mid Ay = x, y \in Y\}$ is convex. If Y is compact and f is closed convex, then g is closed convex.*

Proof Let $x_0, x_1 \in \mathbb{R}^n$. If $Ay = x_0$ has no solution in $y \in Y$, then $g(x_0) = +\infty$, and similarly if $Ay = x_1$ has no solutions then $g(x_1) = +\infty$, and we trivially have $g(\lambda x_0 + (1-\lambda)x_1) \le +\infty$ in either case for all $\lambda \in (0,1)$. Assuming that the sets $\{y \in Y \mid Ay = x_0\}$ and $\{y \in Y \mid Ay = x_1\}$ are non-empty, let $\epsilon > 0$ be arbitrary and y_0, y_1 satisfy $Ay_i = x_i$ and that $f(y_i) \le g(x_i) + \epsilon$. Then $y_\lambda = \lambda y_0 + (1-\lambda)y_1$ satisfies $Ay_\lambda = \lambda x_0 + (1-\lambda)x_1$, and so

$$g(\lambda x_0 + (1-\lambda)x_1) \le f(\lambda y_0 + (1-\lambda)y_1) \le \lambda f(y_0) + (1-\lambda)f(y_1) \le \lambda g(x_0) + (1-\lambda)g(x_1) + \epsilon$$

for all $\lambda \in [0,1]$. Take $\epsilon \to 0$.

For the lower semicontinuity (closed convexity) statement, let $x_n \to x$; we wish to show that $\liminf_n g(x_n) \ge g(x)$. If $g(x_n) = +\infty$ for all x_n, then we trivially have the result. Otherwise, assume $g(x_n) < \infty$ for all n, let $\epsilon > 0$ be arbitrary, and let $y_n \in Y$ satisfy $Ay_n = x_n$ and $f(y_n) \le g(x_n) + \epsilon$. Then as Y is compact, y_n has convergent subsequences; let y be any such limit. We have $Ay = x$, and $g(x) \le f(y) \le \liminf_n f(y_n) \le \liminf_n g(x_n) + \epsilon$. As $\epsilon > 0$ was arbitrary, we hav ethe result. □

From the proposition we immediately see that if $f(x,y)$ is jointly convex in x and y, then the partially minimized function $\inf_{y\in Y} f(x,y)$ is convex whenever Y is a convex set.

Lastly, we consider the functional analogue of the perspective transform. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, the *perspective transform* of f is defined as

$$\operatorname{pers}(f)(x,t) := \begin{cases} tf\left(\frac{x}{t}\right) & \text{if } t > 0 \text{ and } \frac{x}{t} \in \operatorname{dom} f \\ +\infty & \text{otherwise.} \end{cases} \tag{B.3.11}$$

In analogue with the perspective transform of a convex set, the perspective transform of a function is (jointly) convex.

Proposition B.3.12. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then $\operatorname{pers}(f) : \mathbb{R}^{d+1} \to \mathbb{R}$ is convex.*

Proof The result follows if we can show that $\operatorname{epi}\operatorname{pers}(f)$ is a convex set. With that in mind, note that

$$\mathbb{R}^d \times \mathbb{R}_{++} \times \mathbb{R} \ni (x,t,r) \in \operatorname{epi}\operatorname{pers}(f) \text{ if and only if } f\left(\frac{x}{t}\right) \le \frac{r}{t}.$$

Rewriting this, we have

$$\begin{aligned} \operatorname{epi}\operatorname{pers}(f) &= \left\{(x,t,r) \in \mathbb{R}^d \times \mathbb{R}_{++} \times \mathbb{R} : f\left(\frac{x}{t}\right) \le \frac{r}{t}\right\} \\ &= \left\{t(x',1,r') : x' \in \mathbb{R}^d, t \in \mathbb{R}_{++}, r' \in \mathbb{R}, f(x') \le r'\right\} \\ &= \{t(x,1,r) : t > 0, (x,r) \in \operatorname{epi} f\} = \mathbb{R}_{++} \times \{(x,1,r) : (x,r) \in \operatorname{epi} f\}. \end{aligned}$$

This is a convex cone. □

B.3.4 Smoothness properties, first-order developments for convex functions, and subdifferentiability

In addition to their continuity properties, convex functions typically enjoy strong differentiability properties. Some of these interact with the duality properties we present in the section C.2 to follow. Our main goal will be to show how there exist (roughly) derivative-like objects for convex functions, so that for some suitably nice object $D_f(x, v)$ we have

$$f(x + tv) = f(x) + D_f(x, v)t + o(t) \tag{B.3.12}$$

for t small and any v. In the case that f is differentiable, of course, this must coincide with the usual derivative, so that $D_f(x, v) = \langle \nabla f(x), v \rangle$. For convex functions, a directional derivative *always* exists (even if f is non-differentiable), meaning that we can make sense of the first-order development (B.3.12) in some generality.

As one prototypical result, we leverage Rademacher's theorem on almost everywhere differentiability of Lipschitz functions to show that convex functions are almost everywhere differentiable:

Theorem B.3.13 (Rademacher)**.** *Let $U \subset \mathbb{R}^d$ be open and $f : U \to \mathbb{R}^k$ be Lipschitz continuous. Then f is differentiable almost everywhere on U.*

Proofs of this result are standard in measure-theoretic analysis texts; see, e.g., [83, Section 3.5] or [171, Theorem 10.8(ii)]. As any convex function is locally Lipschitz on its domain (recall Theorem B.3.4), we thus have the following result (where we assume that dom f has an interior).

Corollary B.3.14. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then it is differentiable except on a set of Lebesgue measure zero on its domain.*

Other differentiability properties of convex functions are also of interest. We begin by considering directional differentiability properties, after which we expand to consider differentiability and continuous differentiability of (convex) functions. To begin, recall that the *directional derivative* of a function f in direction v at x is

$$f'(x; v) := \lim_{t \downarrow 0} \frac{f(x + tv) - f(x)}{t} \tag{B.3.13}$$

when this quantity exists. When $f'(x; v)$ exists for all directions v and is linear in v, we call the function *Gateaux differentiable*. A (stronger in infinite dimensions) notion of differentiability is *Fréchet differentiability*: f has Fréchet differential g at x if

$$f(y) = f(x) + \langle g, y - x \rangle + o(\|y - x\|) \tag{B.3.14}$$

as $y \to x$, which is then uniform in the distance $\|y - x\|$. It is immediate that if f is Fréchet differentiable with derivative g then it is Gateaux differentiable with $f'(x; v) = \langle g, v \rangle$. Conveniently, in finite dimensions, these notions coincide with the standard gradient, and $f'(x; v) = \langle \nabla f(x), v \rangle$, whenever f is locally Lipschitzian.

Proposition B.3.15. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be Gateaux differentiable at x, that is, its directional derivative $f'(x; v)$ is linear in v, and locally Lipschitz, so that there exists $L < \infty$ such that $|f(x) - f(y)| \le L \|x - y\|$ for y near x. Then f is Fréchet differentiable with Fréchet differential*

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_j}\right]_{j=1}^{d},$$

and $f'(x; v) = \langle \nabla f(x), v \rangle$ and $\|\nabla f(x)\| \le L$.

Proof If f is Fréchet differentiable at x with differential g, then we immediately have

$$\frac{f(x+tv)-f(x)}{t} = \frac{t\langle g, v\rangle + o(t)}{t} \to \langle g, v\rangle$$

as $t \to 0$, so that it is Gateaux differentiable.

Conversely, suppose that $f'(x; v) = \langle g, v\rangle$ for all $v \in \mathbb{R}^d$ for some $g \in \mathbb{R}^d$. Assume for the sake of contradiction that f is not Fréchet differentiable at x, so that

$$\limsup_{\|\Delta\|\downarrow 0} \frac{f(x+\Delta) - f(x) - \langle g, \Delta\rangle}{\|\Delta\|} = c > 0.$$

Take any sequence $\Delta_n \to 0$ achieving this limit supremum, and let $\Delta_n = \epsilon_n v_n$ for a sequence v_n on the sphere, that is, $\|v_n\| = 1$, so $\epsilon_n = \|\Delta_n\|$. Then by passing to a subsequence if necessary, we can assume w.l.o.g. that $v_n \to v$ with $\|v\| = 1$. Then

$$\begin{aligned}\frac{|f(x+\Delta_n) - f(x) - \langle g, \Delta_n\rangle|}{\epsilon_n} &= \frac{|f(x+\epsilon_n v + \epsilon_n(v_n - v)) - f(x) - \epsilon_n\langle g, v\rangle - \epsilon_n\langle g, v_n - v\rangle|}{\epsilon_n} \\ &\le \frac{|f(x+\epsilon_n v) - f(x) - \epsilon_n\langle g, v\rangle|}{\epsilon_n} + \frac{L\epsilon_n \|v_n - v\| + \epsilon_n \|g\| \|v_n - v\|}{\epsilon_n}.\end{aligned}$$

Both of these terms tend to zero, a contradiction, and so f is Fréchet differentiable at x, and its Fréchet derivative is g. That Fréchet differentiability implies differentiability follows by noting that the partial derivatives $f'(x; e_j) = \frac{\partial f(x)}{\partial x_j}$ for each coordinate j.

Finally, the Lipschitzian bound on $\|\nabla f(x)\|$ follows by noting that

$$L\|\Delta\| \ge |f(x+\Delta) - f(x)| = |\langle \nabla f(x), \Delta\rangle| + o(\|\Delta\|).$$

Taking $\Delta = tv$ and $t \downarrow 0$, this implies that $L\|v\| \ge \langle \nabla f(x), v\rangle$ for all v, which is equivalent to $\|\nabla f(x)\| \le L$. □

The main consequence of convexity that is important for us is that a convex function is directionally differentiable at every point in the interior of its domain, though the directional derivative need not be linear:

Proposition B.3.16. *Let f be convex and $x \in \operatorname{int}\operatorname{dom} f$. Then $f'(x; v)$ exists and the mapping $v \mapsto f'(x; v)$ is sublinear, convex, and globally Lipschitz.*

Proof If $x \in \operatorname{int}\operatorname{dom} f$, then the criterion (B.3.4) of increasing slopes guarantees that $f'(x; v) = \lim_{t\downarrow 0} \frac{f(x+tv)-f(x)}{t}$ exists for all $x \in \operatorname{int}\operatorname{dom} f$, as the quantity is monotone. To see that $f'(x; v)$ is convex and sublinear in v, note that positive homogeneity is immediate, as we have $\frac{1}{t}(f(x+\alpha tv) - f(x)) = \frac{\alpha}{\alpha t}(f(x+\alpha tv) - f(x))$ for all $\alpha > 0$, and $f'(x; 0) = 0$. That it is convex is straightforward as well: for any u, v we have

$$\frac{f(x+t(\lambda u + (1-\lambda)v)) - f(x)}{t} \le \lambda\frac{f(x+tu) - f(x)}{t} + (1-\lambda)\frac{f(x+tv) - f(x)}{t}$$

and take $t \downarrow 0$. For the global Lipschitz claim, note that f is already locally Lipschitz near $x \in \operatorname{int}\operatorname{dom} f$ (recall Theorem B.3.4), so that there exists some $L < \infty$ and $\epsilon > 0$ such that for all $\|v\| = 1$ and $0 \le t \le \epsilon$ $|f(x+tv) - f(x)| \le Lt$, whence $|f'(x; v)| \le L$ and by homogeneity

$|f'(x;v)| \le L\|v\|$ for all v. □

An inspection of the proof shows that the result extends even to all of $\operatorname{dom} f$ if we allow $f'(x;v) = +\infty$ whenever $x + tv \notin \operatorname{dom} f$ for all $t > 0$, though of course we lose that $f'(x;v)$ is finite-valued. Then we have the following corollary, showing that $f'(x;v)$ provides a valid first-order development of f in all directions from x (where we take $\infty \cdot t = \infty$ whenever $t > 0$).

Corollary B.3.17. *Let $x \in \operatorname{dom} f$. Then*

$$f(x + tv) = f(x) + f'(x;v)t + o(t)$$

as $t \downarrow 0$ and

$$f(x + tv) \ge f(x) + f'(x;v)t \quad \textit{for all } t \ge 0.$$

Proof The first part is immediate by definition of $f'(x;v) = \lim_{t\downarrow 0} \frac{f(x+tv)-f(x)}{t}$. The second is immediate from the criterion (B.3.4) of increasing slopes, as the limit in the directional derivative (B.3.13) becomes an infimum for convex functions: $f'(x;v) = \inf_{t>0} \frac{f(x+tv)-f(x)}{t}$. □

There are strong connections between subdifferentials and directional derivatives, and hence of the local developments (B.3.12). The following result makes this clear.

Proposition B.3.18. *Let f be convex and $x \in \operatorname{relint} \operatorname{dom} f$. Then*

$$\partial f(x) = \left\{s \mid \langle s, v\rangle \le f'(x;v) \textit{ for all } v\right\} \ne \varnothing.$$

Proof For shorthand let $S = \{s \mid \langle s, v\rangle \le f'(x;v) \text{ all } v\}$ be the set on the right. If $s \in S$, then the criterion (B.3.4) of increasing slopes guarantees that

$$\langle s, v\rangle \le \frac{f(x + tv) - f(x)}{t} \quad \text{for all } v \in \mathbb{R}^d, t > 0.$$

Recognizing that as v is allowed to vary over all of $\mathbb{R}^d$ and $t > 0$, then $x + tv$ similarly describes $\mathbb{R}^d$, we see that this condition is completely equivalent to the definition (B.3.6) of the subgradient.

That $\partial f(x) \ne \varnothing$ is Theorem B.3.3. □

We can also extend this to $x \in \operatorname{dom} f$—not necessarily the interior—where we see that there is no loss (even when f may be $+\infty$ valued) to defining

$$\partial f(x) := \left\{s \mid \langle s, v\rangle \le f'(x;v) \text{ for all } v\right\}. \tag{B.3.15}$$

Notably, the directional derivative function $v \mapsto f'(x;v)$ always exists for $x \in \operatorname{dom} f$ and is a sublinear convex function, and thus $\partial f(x)$ above is always a closed convex set whose support function (recall (B.2.1)) is the closure of $v \mapsto f'(x;v)$. While the subdifferential $\partial f(x)$ is always a compact convex set when $x \in \operatorname{int} \operatorname{dom} f$, even when it exists it may not be compact if x is on the boundary of $\operatorname{dom} f$. To see one important example of this, consider the indicator function

$$\mathbf{I}_C(x) := \begin{cases} +\infty & \text{if } x \notin C \\ 0 & \text{if } x \in C \end{cases}$$

of a closed convex set C. For simplicity, let $C = [a, b]$ be an interval. Then we have

$$\partial \mathbf{I}_C(x) = \begin{cases} [0, \infty] & \text{if } x = b \\ \{0\} & \text{if } a < x < b \\ [-\infty, 0] & \text{if } x = a. \end{cases}$$

Whether points $\pm\infty$ are included is a matter of convenience and whether we work with the extended real line.

JCD Comment: Draw a picture of this

These representations points to a certain closure property of subgradients, namely, that the subdifferential is closed under additions of the normal cone to the domain of f:

Lemma B.3.19. *Let $\mathcal{N}_{\operatorname{dom} f}(x)$ be the normal cone (Definition C.1) to $\operatorname{dom} f$ at the point x (where $\mathcal{N}_{\operatorname{dom} f}(x) = \{0\}$ for $x \in \operatorname{int} \operatorname{dom} f$ and $\mathcal{N}_{\operatorname{dom} f}(x) = \varnothing$ for $x \notin \operatorname{dom} f$). Then*

$$\partial f(x) = \partial f(x) + \mathcal{N}_{\operatorname{dom} f}(x).$$

In particular, if x is a boundary point $x \in \operatorname{bd} \operatorname{dom} f$ of the domain of f, then either $\partial f(x) = \varnothing$ or $\partial f(x)$ is unbounded.

Proof We only need concern ourselves with points $x \in \operatorname{bd} \operatorname{dom} f$, where the normal cone $\mathcal{N} = \mathcal{N}_{\operatorname{dom} f}(x)$ is non-trivial. If $\partial f(x)$ is empty, there is nothing to prove, so assume that $\partial f(x)$ is non-empty. Then the definition (B.3.15) of the subdifferential as $\partial f(x) = \{s \mid \langle s, u\rangle \le f'(x; u)\}$ allows us to prove the result. First, consider vectors u for which $f'(x; u) = +\infty$. Then certainly, for any $s \in \partial f(x)$, we have $\langle s + v, u\rangle \le f'(x; u)$ for all $v \in \mathcal{N}$. If $f'(x; u) < \infty$, then for small enough $t > 0$ we necessarily have $x + tu \in \operatorname{dom} f$. In particular, the definition of the normal cone gives that $v \in \mathcal{N}$ satisfies $0 \ge \langle v, x + tu - x\rangle = t\langle v, u\rangle$, or that $\langle v, u\rangle \le 0$. Thus $\langle s + v, u\rangle \le \langle s, u\rangle \le f'(x; u)$, and so $s + v \in \partial f(x)$ once again.

The claim about boundedness is immediate, because $\mathcal{N}_{\operatorname{dom} f}$ is a cone. □

A more compelling case for the importance of the subgradient set with respect to first-order developments and differentiability properties of convex functions is the following:

JCD Comment: Add a picture of this as well.

Proposition B.3.20. *Let f be convex and $x \in \operatorname{int} \operatorname{dom} f$. Then*

$$\begin{aligned} f(y) &= f(x) + \sup_{s \in \partial f(x)} \langle s, y - x\rangle + o(\|y - x\|) \\ &= f(x) + f'(x; y - x) + o(\|y - x\|). \end{aligned}$$

Proof That $\sup_{s \in \partial f(x)} \langle s, v\rangle = f'(x; v)$ is immediate by Theorem B.3.7 and Proposition B.2.1, because $f'(x; v)$ is sublinear and closed convex in v when $x \in \operatorname{int} \operatorname{dom} f$. Certainly the right hand sides are then equal.

We thus prove the equality $f(y) = f(x) + f'(x; y - x) + o(\|y - x\|)$, where the argument is similar to that for Proposition B.3.15. Let $y_n \to x$ be any sequence and let $\Delta_n = y_n - x$, so that $\|\Delta\| \to 0$; as $x \in \operatorname{int} \operatorname{dom} f$, there exists a (local) Lipschitz constant L such that $|f(x + \Delta) - f(x)| \le L \|\Delta\|$

for all small Δ. Similarly, because $v \mapsto f'(x; v)$ is convex (even positively homogeneous and thus sublinear), it has a Lipschitz constant, and we take this to be L as well. Now, write $\Delta_n = \epsilon_n v_n$ where $\|v_n\| = 1$ and $\epsilon_n \to 0$, and moving to a subsequence if necessary let $v_n \to v$. Then we have

$$\begin{aligned} f(x + \Delta_n) - f(x) - f'(x; \Delta_n) &= f(x + \epsilon_n v + \epsilon_n(v_n - v)) - f(x) - f'(x; \epsilon_n v + \epsilon_n(v_n - v)) \\ &= f(x + \epsilon_n v) - f(x) - f'(x; \epsilon_n v) \pm 2L\epsilon_n \|v_n - v\| \\ &= o(\epsilon_n) \end{aligned}$$

because $\|v_n - v\| \to 0$ and $f(x + \epsilon_n v) - f(x) = f'(x; \epsilon_n) + o(\epsilon_n)$ by definition of the directional derivative. □

Note that convexity only played the role of establishing the local Lipschitz property of f in the proof of Proposition B.3.20; any locally Lipschitz function with directional derivatives will enjoy a similar first-order expansion.

As our final result on smoothness properties of convex functions, we connect subdifferentials to differentiability properties of convex f. First, we give a lemma showing that the subdifferential set ∂f is outer semicontinuous.

Lemma B.3.21 (Closure of the graph of the subdifferential)**.** *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex. Then the graph $\{(x, s) \mid x \in \mathbb{R}^d, s \in \partial f(x)\}$ of its subdifferential is closed. Equivalently, whenever $x_n \to x$ with $s_n \in \partial f(x_n)$ and $s_n \to s$, f has non-empty subdifferential at x with $s \in \partial f(x)$.*

Proof We prove the second statement, whose equivalence to the first is definitional. Fix any $y \in \mathbb{R}^d$. Then $f(y) \geq f(x_n) + \langle s_n, y - x_n \rangle$, and because f is closed (i.e., lower semicontinuous), we have $\liminf f(x_n) \geq f(x)$. Let $\epsilon > 0$ be arbitrary. Then for all large enough n, we have $f(x_n) \geq f(x) - \epsilon$, and similarly, $\|s_n - s\| \leq \epsilon$, $\|x_n - x\| \leq \epsilon$, and $\|y - x_n\| \leq \|y - x\| + \epsilon$. Then

$$\begin{aligned} f(y) \geq f(x_n) + \langle s_n, y - x_n \rangle &\geq f(x) + \langle s, y - x_n \rangle - \epsilon - \|s - s_n\| \|y - x_n\| \\ &\geq f(x) + \langle s, y - x \rangle - \epsilon - \epsilon \|y - x_n\| - \|s\| \|x - x_n\| \\ &\geq f(x) + \langle s, y - x \rangle - \epsilon - \epsilon(1 + \epsilon) \|y - x\| - \|s\| \epsilon. \end{aligned}$$

As ϵ was arbitrary we have $f(y) \geq f(x) + \langle s, y - x \rangle$ as desired. □

Given the somewhat technical Lemma B.3.21, we can show that if f is convex and differentiable at a point, it is in fact continuously differentiable at the point.

Proposition B.3.22. *Let f be convex and $x \in \operatorname{int} \operatorname{dom} f$. Then $\partial f(x)$ is a singleton if and only if f is differentiable at x. If additionally f is differentiable on an open set U, then f is continuously differentiable on U.*

Proof Because $x \in \operatorname{int} \operatorname{dom} f$, there exists $L < \infty$ such that f is L-Lipschitz near x by Theorem B.3.4. Suppose that $\partial f(x) = \{s\}$. Then the directional derivative $f'(x; v) = \langle s, v \rangle$ for all v, and Proposition B.3.20 gives

$$f(y) = f(x) + \langle s, y - x \rangle + o(\|y - x\|)$$

as $y \to x$, that is, f is differentiable. Conversely, assume that f is differentiable at x. Then taking any vector v, we immediately have $f'(x; v) = \langle \nabla f(x), v \rangle$ and Proposition B.3.18 gives that $\partial f(x) = \{\nabla f(x)\}$.

To see that f is in fact continuously differentiable on U, let $x \in U$ and f be L-Lipschitz on a compact set $C \subset U$ containing x in its interior. Let $x_k \in C$ satisfy $x_k \to x$ and let $s_k = \nabla f(x_k) \in \partial f(x_k)$. Then $\|s_k\| \le L$, and each subsequence has a further convergent subsequence. Lemma B.3.21 implies that any convergent subsequence $s_{k(m)} \to s \in \partial f(x)$. But as $\partial f(x) = \{\nabla f(x)\}$, we have $\nabla f(x_{k(m)}) \to \nabla f(x)$ and so $\nabla f(x)$ is continuous in x. □

B.3.5 Calculus rules of subgradients

We close this section with a few calculus results on subdifferentials of of convex functions. These calculus rules show that we may take the subdifferential plays a similar role to the gradient for differentiable functions. Additionally, they allow us to take derivatives of various extremal functions.

Our first result shows that subdifferentials of sums are sums of subdifferentials, which relies on both the representation of sublinear functions as support functions for convex sets and the characterization of the subdifferential in terms of directional derivatives:

Proposition B.3.23. *Let f and g be closed convex functions, and let $x \in \operatorname{int}\operatorname{dom} f$ and g be subdifferentiable at x, meaning that $\partial g(x) \neq \varnothing$. Then*

$$\partial(f + g)(x) = \partial f(x) + \partial g(x).$$

Proof By Proposition B.3.18, the set $\partial f(x)$ is a compact convex set, and the general definition (B.3.15) of the subdifferential gives that $\partial g(x)$ is closed convex. Let $S_1 = \partial f(x)$ and $S_2 = \partial g(x)$. Then immediately $S_1 + S_2 \subset \partial(f + g)(x)$, so that

$$S := \partial(f + g)(x) = \left\{s \mid \langle s, v\rangle \le f'(x; v) + g'(x; v) \text{ for all } v \in \mathbb{R}^d\right\}$$

is non-empty. Because of the support function equality $f'(x; v) = \sigma_{S_1}(v)$ and $g'(x; v) = \sigma_{S_2}(v)$, Corollary B.2.5 gives

$$\sigma_S(v) = \sigma_{S_1}(v) + \sigma_{S_2}(v) = \sigma_{S_1+S_2}(v).$$

Thus (Corollary B.2.4) $S_1 + S_2 = S$. □

Other situations that arise frequently are composition with affine mappings and taking maxima or suprema of convex functions, so that finding a calculus for these is also important.

Corollary B.3.24. *Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex and for $A \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$, let $g(x) = f(Ax+b)$. Then*

$$\partial g(x) = A^T \partial f(Ax + b).$$

Proof Using the directional derivative, we have $g'(x; v) = f'(Ax + b; Av)$ for all $v \in \mathbb{R}^d$, and applying Proposition B.2.6 gives that the latter is the support function of the convex compact set $A^T \partial f(Ax + b)$. □

It is also useful to be able to compute subdifferentials of maxima and suprema (recall Proposition B.3.9). Consider a collection $\{f_\alpha\}_{\alpha\in\mathcal{A}}$ of convex functions, and define

$$f(x) := \sup_{\alpha\in\mathcal{A}} f_\alpha(x). \tag{B.3.16}$$

The function f is certainly convex. For a given x let

$$\mathcal{A}(x) := \{\alpha \in \mathcal{A} \mid f_\alpha(x) = f(x)\}$$

be the indices attaining the suprema, that is, the active index set (though this may be empty). Then there is an "easy" direction:

Lemma B.3.25. *With the notation above,*

$$\partial f(x) \supset \operatorname{cl} \operatorname{Conv}\left\{\bigcup \partial f_\alpha(x) \mid \alpha \in \mathcal{A}(x)\right\} = \operatorname{cl} \operatorname{Conv}\left\{g \mid g \in \partial f_\alpha(x) \text{ for some } \alpha \in \mathcal{A}(x)\right\}.$$

Proof Let $\alpha \in \mathcal{A}(x)$ and $g \in \partial f_\alpha(x)$. Then

$$f(y) \geq f_\alpha(y) \geq f_\alpha(x) + \langle g, y - x\rangle = f(x) + \langle g, y - x\rangle.$$

Thus $g \in \partial f(x)$, which as a closed convex set must thus include its closed convex hull. □

A much more challenging argument is to show that the active index set $\mathcal{A}(x)$ exactly characterizes the subdifferential of f at x; we simply state a typical result as a proposition.

Proposition B.3.26. *Let $\mathcal{A}$ be a compact set (for some metric) and assume that for each x, the mapping $\alpha \mapsto f_\alpha(x)$ is upper semi-continuous. Then*

$$\partial f(x) = \operatorname{Conv}\left\{\bigcup \partial f_\alpha(x) \mid \alpha \in \mathcal{A}(x)\right\} = \operatorname{Conv}\left\{g \mid g \in \partial f_\alpha(x) \text{ for some } \alpha \in \mathcal{A}(x)\right\}.$$

For a proof, see [104, Theorem 4.4.2].

JCD Comment: Draw a picture of this

Finally, we revisit the partial minimization operation in Proposition B.3.11. In this case, we require a bit more care when defining subdifferentials and subdifferentiability. For $A \in \mathbb{R}^{n\times m}$ with $m \geq n$, where A has rank n (so that $x \mapsto Ax$ is surjective) and $f : \mathbb{R}^m \to \overline{\mathbb{R}}$, define the function

$$f_A(x) = \inf\{f(y) \mid Ay = x\},$$

which is convex. Define the set $Y^\star(x) := \{y \mid Ay = x \text{ and } f_A(x) = f(y)\}$ to be the set of y attaining the infimum in the definition of f_A, which may be empty. When it is not, however, we can characterize the subdifferential of $f_A(x)$:

Proposition B.3.27. *Let $x \in \mathbb{R}^n$ be a point for which $Y^\star(x)$ is non-empty for the function f_A. Then*

$$\partial f_A(x) = \left\{s \mid A^T s \in \partial f(y)\right\}$$

for any $y \in Y^\star(x)$, and the set on the right is independent of the choice of y.

Proof A vector s is a subgradient of f at x if and only if

$$f_A(x') \geq f_A(x) + \langle s, x' - x\rangle \quad \text{for all } x' \in \mathbb{R}^n,$$

which (as $Ay = x$ for $y \in Y^\star(x)$) is equivalent to

$$f_A(x') \geq f(y) + \langle s, x' - Ay\rangle \quad \text{for all } x' \in \mathbb{R}^n.$$

Because A has full row rank, for any $x' \in \mathbb{R}^n$ there exists y' with $Ay' = x'$; by definition of f_A as the infimum, the preceding display is thus equivalent to

$$f(y') \geq f_A(Ay') \geq f(y) + \langle s, Ay' - Ay \rangle \text{ for all } y' \in \mathbb{R}^m.$$

This holds if and only if $A^T s$ is a subgradient of f at y. $\square$

Appendix C

Optimality, stability, and duality

The existence and continuity properties of minimizers of (convex) optimization problems play a central role in much of statistical theory. They are especially essential in our understanding of loss functions and the associated optimality properties. In our context, this is especially central for problems of classification calibration or surrogate risk consistency, as in Chapters 13 and 14. This appendix records several representative results along these lines, and also builds up the duality theory associated with convex conjugates, frequently identified as Fenchel-Young duality.

Broadly, throughout this appendix, we shall consider the generic optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array} \tag{C.0.1}$$

where C is a closed convex set (we have not yet assumed convexity of f), Throughout (as in the previous appendix) we assume that f is proper, so that $f(x) > -\infty$ for each x, and that $f(x) = +\infty$ if $x \notin \operatorname{dom} f$.

The most basic question we might ask is when minimizers even exist in the problem (C.0.1). The standard result in this vein is that if minimizers exist whenever C is compact and f is lower semicontinuous (B.3.9), that is, its epigraph is closed, i.e., $\liminf_{x \to x_0} f(x) \geq f(x_0)$.

Proposition C.0.1. *Let C be compact and $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be lower semi-continuous (B.3.9) over C. Then $\inf_{x \in C} f(x) > -\infty$ and the infimum is attained.*

Proof Let $f^\star = \inf_{x \in C} f(x)$, where for now we allow the possibility that $f^\star = -\infty$. Let $x_n \in C$ be a sequence of points satisfying $f(x_n) \to f^\star$. Proceeding to a subseqeunce if necessary, we can assume that $x_n \to x^\star \in C$ by the compactness of C. Then lower semi-continuity guarantees that $f^\star = \lim_n f(x_n) \geq f(x^\star) \geq f^\star$, and so $f(x^\star) = f^\star$ and so necessarily $f^\star > -\infty$. □

When the domain C is not compact but only closed, alternative conditions are necessary to guarantee the existence of minimizers. Perhaps the most frequent, and one especially useful with convexity (as we shall see), is that f is *coercive*, meaning that

$$f(x) \to \infty \quad \text{whenever} \quad \|x\| \to \infty.$$

Proposition C.0.2. *Let C be closed and $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be lower semi-continuous over C and coercive. Then $\inf_{x \in C} f(x) > -\infty$ and the infimum is attained.*

Proof Once again, let $f^\star = \inf_{x \in C} f(x)$ and let $x_n \in C$ satisfy $f(x_n) \to f^\star$. Certainly x_n must be a bounded sequence because f is coercive. Thus, it has a subsequent limit, and w.l.o.g. we assume that $x_n \to x^\star \in C$ by closedness. Lower semi-continuity guarantees that $f^\star \geq \liminf_n f(x_n) = f(x^\star) \geq f^\star$, giving the result. □

Finally, we make a small remark norms and dual norms, as these will be important for the more quantitative smoothness guarantees we provide. For a norm $\|\cdot\|$, the dual norm $\|\cdot\|_*$ has definition

$$\|y\|_* := \sup_x \{\langle x, y\rangle \mid \|x\| \leq 1\}.$$

This is a norm as it is positively homogeneous, $\|y\|_* = 0$ if and only if $y = 0$, and satisfies the triangle inequality. A few brief examples follow, which we leave as exercises to the reader.

(i) The ℓ_2-norm $\|x\|_2 = \sqrt{\langle x, x\rangle}$ is self-dual, so that its dual is $\|\cdot\|_2$.

(ii) The ℓ_1 and ℓ_∞ norms are dual, that is, $\|x\|_\infty = \sup_{\|y\|_1 \leq 1} \langle x, y\rangle$ and $\|y\|_1 = \sup_{\|x\|_\infty \leq 1} \langle x, y\rangle$.

(iii) For all $p \in [1, \infty]$, the dual to the ℓ_p norm $\|x\|_p = (\sum_{j=1}^d |x_j|^p)^{1/p}$ is the ℓ_q norm with $q = \frac{p}{p-1}$, that is, for the $q \geq 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$.

C.1 Optimality conditions and stability properties

With the basic results on existence of minimizers in place, we turn to convex optimization problems, where f is closed convex and C is a closed convex set, and we assume essentially without loss of generality that $\operatorname{dom} f \supset \operatorname{int} C$ (as otherwise, we may replace C with $C \cap \operatorname{cl} \operatorname{dom} f$). The benefits of convexity appear immediately: f has no local but non-global minimizers, and moreover, if f is strictly convex, then any minimizers (if they exist) are unique.

Proposition C.1.1. *Let f be convex. Then if x is a local minimizer of f over C, it is a global minimizer of f over C. If f is strictly convex, then x is unique.*

Proof To say that x is a local minimizer of f over C is to say that $f(x) \leq f(x')$ for all $x' \in C$ with $\|x' - x\| \leq \epsilon$ for some $\epsilon > 0$. Now, consider $y \in C$. By taking $\lambda > 0$ small enough, we have both $(1 - \lambda)x + \lambda y \in C$ and $\|(1 - \lambda)x + \lambda y - x\| \leq \epsilon$, and so

$$f(x) \leq f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y),$$

which rearranged yields $f(y) \geq f(x)$. If f is additionally strictly convex (recall Corollary B.3.2), then the preceding inequality is strict whenver $y \neq x$. □

C.1.1 Subgradient characterizations for optimality

First-order stationary conditions are sufficient for global optimality in convex problems. We can say more once we consider subgradients:

Observation C.1.2. *Let f be convex and subdifferentiable at x. Then x minimizes f if and only if $0 \in \partial f(x)$.*

Proof If $0 \in \partial f(x)$, then $f(y) \geq f(x) + \langle 0, y - x \rangle = f(x)$ for all y. Conversely, if x minimizes f, then we have $f(y) \geq f(x)$ for all y, and in particular, $0 \in \partial f(x)$. □

Things become a bit more complicated when we consider the constraints in the problem (C.0.1), so that the point x may be restricted. In this case, it is important and useful to consider the *normal cone* to the set C, which is (essentially) the collection of vectors pointing out of C.

Definition C.1. *Let C be a closed convex set. The* normal cone *to C at the point $x \in C$ is the collection of vectors*

$$\mathcal{N}_C(x) := \{v \mid \langle v, y - x \rangle \leq 0 \text{ for all } y \in C\}.$$

So $\mathcal{N}_C(x)$ is the collection of vectors making an obtuse angle with any direction into the set C from x. **JCD Comment:** Draw a picture, and also, put this earlier in the discussion of convex sets.

It is clear that $\mathcal{N}_C(x)$ is indeed a cone: if $v \in \mathcal{N}_C(x)$, then certainly $tv \in \mathcal{N}_C(x)$ for all $t \geq 0$. It is closed convex, being the intersection of halfspaces. Moreover, if $x \in \operatorname{int} C$, then we have $\mathcal{N}_C(x) = \{0\}$, and additionally, we can connect the supporting hyperplanes of C to its normal cones: Theorem B.1.12 gives the following corollary.

Corollary C.1.3. *Let C be closed convex. Then for any $x \in \operatorname{bd}(C)$, the normal cone $\mathcal{N}_C(x)$ is non-trivial and consists of the collection of supporting hyperplanes to C at x.*

By a bit of subgradient calculus, we can then write optimality conditions involving the normal cones to C. If C is a closed convex set, the convex indicator function $\mathbf{I}_C(x)$ has subdifferentials

$$\partial \mathbf{I}_C(x) = \begin{cases} \{0\} & \text{if } x \in \operatorname{int} C \\ \mathcal{N}_C(x) & \text{if } x \in \operatorname{bd}(C) \\ \varnothing & \text{otherwise.} \end{cases}$$

The only case requiring justification is the boundary case; for this, we note that $w \in \mathcal{N}_C(x)$ if and only if $\langle w, y - x \rangle \leq 0$ for all $y \in C$, which in turn occurs if and only if $\mathbf{I}_C(y) \geq \mathbf{I}_C(x) + \langle w, y - x \rangle$ for all y.

The subdifferential calculation for $\mathbf{I}_C(x)$ yields the following general optimality characterization for problem (C.0.1).

Proposition C.1.4. *In the problem (C.0.1), let $x \in \operatorname{int} \operatorname{dom} f$. Then x minimizes f over C if and only if*

$$0 \in \partial f(x) + \mathcal{N}_C(x). \tag{C.1.1}$$

Proof The minimimization problem (C.0.1) is equivalent to the problem

$$\operatorname*{minimize}_x \; f(x) + \mathbf{I}_C(x).$$

As $x \in \operatorname{int} \operatorname{dom} f$, f has nonempty compact convex subdifferential $\partial f(x)$, and so $\partial(f + \mathbf{I}_C)(x) = \partial f(x) + \partial \mathbf{I}_C(x) = \partial f(x) + \mathcal{N}_C(x)$ by Proposition B.3.23. Apply Observation C.1.2. □

Several equivalent versions of Proposition C.1.4 are possible. The first is that

$$-\partial f(x) \cap \mathcal{N}_C(x) \neq \varnothing,$$

that is, there is a subgradient vector $g \in \partial f(x)$ such that $-g \in \mathcal{N}_C(x)$, so that $-g$ points outside the set C. **JCD Comment:** Draw a picture

Another variant, frequently used, is to write Proposition C.1.4 as that x solves problem (C.0.1) if and only if there exists $g \in \partial f(x)$ such that

$$\langle g, y - x \rangle \geq 0 \quad \text{for all } y \in C. \tag{C.1.2}$$

Indeed, taking $g \in \partial f(x)$ to be the element satisfying $-g \in \mathcal{N}_C(x)$, we immediately see that $\langle -g, y - x \rangle \leq 0$ for all $y \in C$ by definition of the normal cone, which is (C.1.2). **JCD Comment:** Use picture above

C.1.2 Stability properties of minimizers

The characterizations (C.1.1) and (C.1.2) of optimality for convex optimization problems allow us to develop some stability properties of the minimizers of convex problems. These, in turn, will relate to smoothness properties of various dual functions (as we explore in the sequel), which again become important in the study of consistent losses. Here, we collect a few of the typical results. Typical results in this vein exhibit a few properties: that solutions are "stable," meaning that small tilts of the function f do not change solutions significantly, or that the function f exhibits various strong growth properties.

For our starting point, we begin by consider *strongly convex* functions, where a function is λ-*strongly convex* with respect to the norm $\|\cdot\|$ if for all $t \in [0, 1]$ and $x, y \in \operatorname{dom} f$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\lambda}{2} t(1-t) \|x - y\|^2 . \tag{C.1.3}$$

The definition (C.1.3) makes strict convexity quantitative in a fairly precise way, and has several equivalent characterizations.

Proposition C.1.5 (Equivalent characterizations of strong convexity)**.** *Let f be a convex function, subdifferentiable on its domain. Then the following are equivalent.*

(i) f is λ-strongly convex (Eq. (C.1.3)).

(ii) For all $y \in \mathbb{R}^d$ and $g \in \partial f(x)$,

$$f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\lambda}{2} \|x - y\|^2 .$$

(iii) For all $x, y \in \operatorname{dom} f$ and $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$,

$$\langle g_x - g_y, x - y \rangle \geq \lambda \|x - y\|^2 .$$

Proof Let us prove that (ii) if and only if (iii). Let $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$ and assume (ii) holds. Then

$$f(y) \geq f(x) + \langle g_x, y - x \rangle + \frac{\lambda}{2} \|y - x\|^2$$
$$f(x) \geq f(y) + \langle g_y, x - y \rangle + \frac{\lambda}{2} \|x - y\|^2$$

and adding the equations we obtain

$$0 \geq \langle g_x - g_y, y - x \rangle + \lambda \|x - y\|^2 .$$

Rearranging gives part (iii). Conversely, assume (iii), and for $t \in [0, 1]$ let $x_t = (1 - t)x + ty$ and define $h(t) = f(x(t))$. Then h is convex and hence almost everywhere differentiable (and locally Lipschitz), so that $h(1) = h(0) + \int_0^1 h'(t)dt$. Noting that

$$h'(t) = \langle g_t, y - x \rangle \text{ for some } g_t \in \partial f(x_t)$$

(recall the subgradient characterization of Proposition B.3.18), we have

$$h'(t) = \langle g_t, y - x \rangle = \langle g_t - g_x, y - x \rangle + \langle g_x, y - x \rangle = \frac{1}{t} \langle g_t - g_x, (1 - t)x + ty - x \rangle + \langle g_x, y - x \rangle$$

and so as $h(1) = f(y)$ and $h(0) = f(x)$,

$$\begin{aligned} f(y) &= h(0) + \int_0^1 \frac{\langle g_t - g_x, (1 - t)x + ty - x \rangle}{t} dt + \langle g_x, y - x \rangle \\ &\geq f(x) + \int_0^1 \frac{\lambda \|(1 - t)x + ty - x\|^2}{t} dt + \langle g_x, y - x \rangle \\ &= f(x) + \lambda \|x - y\|^2 \int_0^1 t dt + \langle g_x, y - x \rangle = f(x) + \langle g_x, y - x \rangle + \frac{\lambda}{2} \|y - x\|^2 . \end{aligned}$$

That (ii) implies (i) is relatively straightforward: we have

$$\begin{aligned} f(y) &\geq f(tx + (1 - t)y) + t\langle g_t, y - x \rangle + \frac{\lambda}{2} t^2 \|x - y\|^2 \\ f(x) &\geq f(tx + (1 - t)y) + (1 - t)\langle g_t, x - y \rangle + \frac{\lambda}{2} (1 - t)^2 \|x - y\|^2 \end{aligned}$$

for any $g_t \in \partial f(tx + (1 - t)y)$. Multiply the first inequality by $(1 - t)$ and the second by t, then add them to obtain

$$tf(x) + (1 - t)f(y) \geq f(tx + (1 - t)y) + \frac{\lambda}{2} \left[(1 - t)t^2 + t(1 - t)^2 \right] \|x - y\|^2 ,$$

and note that $(1 - t)t^2 + t(1 - t)^2 = t(1 - t)$. Finally, let (i) hold, and which is equivalent to the condition that

$$\frac{f((1 - t)x + ty) - f(x)}{t} + \frac{\lambda}{2} (1 - t) \|x - y\|^2 \leq f(y) - f(x)$$

for $t \in (0, 1)$. Taking $t \downarrow 0$ gives $f'(x; y - x) + \frac{\lambda}{2} \|x - y\|^2 \leq f(y) - f(x)$, and because $f'(x; y - x) = \sup_{s \in \partial f(x)} \langle s, y - x \rangle$ we obtain (ii). □

As a first example application of strong convexity, consider minimizers of the tilted functions

$$f_u(x) := f(x) - \langle u, x \rangle$$

as u varies. First, note that minimizers necessarily exist: the function $f_u(x) \to \infty$ whenever $\|x\| \to \infty$ by condition (ii) in Proposition C.1.5, and so we can restrict to minimizing f_u over

compacta. Moreover, the minimizers $x_u := \operatorname{argmin}_x f_u(x)$ are unique, as the functions f_u are strongly (and hence strictly) convex. However, we can say more. Indeed, let C be any closed convex set and let

$$x_u = \operatorname*{argmin}_{x \in C} f_u(x). \tag{C.1.4}$$

We claim the following:

Proposition C.1.6. *Let f be λ-strongly convex with respect to the norm $\|\cdot\|$ and subdifferentiable on C. Then the mapping $u \mapsto x_u$ is $\frac{1}{\lambda}$-Lipschitz continuous with respect to the dual norm $\|\cdot\|_*$, that is, $\|x_u - x_v\| \le \frac{1}{\lambda} \|u - v\|_*$.*

Proof We use the optimality condition (C.1.2). We have $\partial f_u(x) = \partial f(x) - u$, and thus for any u, v we have both

$$\langle g_u - u, y - x_u \rangle \ge 0 \quad \text{and} \quad \langle g_v - v, y - x_v \rangle \ge 0$$

for some $g_u \in \partial f(x_u)$ and $g_v \in \partial f(x_v)$ for all $y \in C$. Set $y = x_v$ in the first inequality and $y = x_u$ in the second and add them to obtain

$$\langle g_u - g_v + v - u, x_v - x_u \rangle \ge 0 \quad \text{or} \quad \langle v - u, x_v - x_u \rangle \ge \langle g_v - g_u, x_v - x_u \rangle.$$

By strong convexity the last term satisfies $\langle g_v - g_u, x_v - x_u \rangle \ge \lambda \|x_u - x_v\|^2$. By definition of the dual norm, $\|v - u\|_* \|x_v - x_u\| \ge \langle v - u, x_v - x_u \rangle$, so $\|u - v\|_* \|x_v - x_u\| \ge \lambda \|x_u - x_v\|^2$, which is the desired result. □

JCD Comment: Figure for the preceding lemma.

There are alternative versions of strong convexity, typically given the name *uniform convexity* in the convex analysis literature, which allow generalizations and similar quantitative stability properties. In analogy with the strong convexity condition (C.1.3), we say that f is (λ, κ)-uniformly convex, where $\kappa \ge 2$, over C if it is closed and for all $t \in [0, 1]$ and $x, y \in C$,

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) - \frac{\lambda}{2} t(1-t) \|x - y\|^\kappa \left[(1-t)^{\kappa-1} + t^{\kappa-1}\right]. \tag{C.1.5}$$

Notably the $\kappa = 2$ case is simply the familiar strong convexity.

JCD Comment: Give lemmas and propositions but leave as exercises, filling this out.

JCD Comment: Add some material on strict convexity implying a bit of growth around a neighborhood, and stability properties of strongly convex functions.

The weakest version of such strong convexity properties is strict convexity, for which a careful reading of the proof of Proposition C.1.5 (replace all λ with 0 and inequalities with strict inequalities) gives the following characterization of equivalent definitions of strict convexity (recall also Corollary B.3.2).

Corollary C.1.7. *Let f be a convex function subdifferentiable on C. The following are equivalent.*

(i) f is strictly convex on C.

(ii) For all $x \in C$, $y \ne x$, and $g \in \partial f(x)$,

$$f(y) > f(x) + \langle g, y - x \rangle.$$

(iii) For all $x, y \in C$ and $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$,

$$\langle g_x - g_y, x - y \rangle > 0.$$

Using Corollary C.1.7, we can then obtain certain smoothness properties of the tilted minimizers x_u of the minimization (C.1.4). We begin with a lemma that guarantees growth of convex functions over their first-order approximations.

Lemma C.1.8. *Let f be convex and subdifferentiable on the closed convex set C, and for any fixed $g \in \partial f(x_0)$ define the* Bregman divergence

$$D(x, x_0) := f(x) - f(x_0) - \langle g, x - x_0 \rangle.$$

Then for all $0 \le \epsilon \le \epsilon'$, $\delta(\epsilon) := \inf\{D(x, x_0) \mid x \in C, \|x - x_0\| \ge \epsilon\}$ is attained, nonnegative, and $\delta(\epsilon') \ge \frac{\epsilon'}{\epsilon}\delta(\epsilon)$.

Proof Fix $x \in C$. Letting $h(t) = D(x_0 + t(x - x_0), x_0)$, h is convex in $t \ge 0$, locally Lipschitz, and satisfies $h(0) = \inf_t h(t) = 0$, so we can write $h(t) = h(0) + \int_0^t h'(s; 1)ds$. Additionally, $s \mapsto h'(s; 1) \ge 0$ is nondecreasing by the increasing slopes criterion (B.3.4).

For all $\epsilon > 0$, then, we may restrict infimum in the definition of $\delta(\epsilon)$ to those $x \in C$ satisfying $\|x - x_0\| = \epsilon$, a compact set, so that the infimum is attained at some $x_\epsilon \in C$ with $\|x_\epsilon - x_0\| = \epsilon$. Now, let $\epsilon' > \epsilon$, and $x_{\epsilon'}$ achieve the infimum in $\delta(\epsilon)$. Then setting $x' = \frac{\epsilon}{\epsilon'}(x_{\epsilon'} - x_0) + x_0$ (implying $x_{\epsilon'} = \frac{\epsilon'}{\epsilon}(x' - x_0) + x_0$), we have $\|x' - x_0\| = \epsilon$ and so $D(x', x_0) \ge D(x_\epsilon, x_0) = \delta(\epsilon)$. Set $h(t) = D(x_0 + t(x' - x_0), x_0)$. Rewriting and using the first-order convexity condition,

$$\begin{aligned}\delta(\epsilon') = D(x_{\epsilon'}, x_0) = D\left(\frac{\epsilon'}{\epsilon}(x' - x_0) + x_0, x_0\right) = h\left(\frac{\epsilon'}{\epsilon}\right) &\ge h(1) + \left[\frac{\epsilon'}{\epsilon} - 1\right] h'(1; 1) \\ &= D(x', x_0) + \left[\frac{\epsilon'}{\epsilon} - 1\right] h'(1; 1).\end{aligned}$$

A minor variant of the criterion of increasing slopes (B.3.4) and that $h(0) = 0$ then gives $h'(1; 1) = \lim_{t \downarrow 0} \frac{h(1+t) - h(1)}{t} \ge \frac{h(1) - h(0)}{1} = h(1) = D(x', x_0)$, so we have

$$\delta(\epsilon') = D(x_{\epsilon'}, x_0) \ge D(x', x_0) + \left[\frac{\epsilon'}{\epsilon} - 1\right] D(x', x_0) = \frac{\epsilon'}{\epsilon} D(x', x_0) \ge \frac{\epsilon'}{\epsilon}\delta(\epsilon)$$

as desired. □

Whenever f is strictly convex, because the infimum in $\delta(\epsilon)$ is attained in Lemma C.1.8, we have the following guarantee.

Lemma C.1.9. *Let the conditions of Lemma C.1.8 hold and additionally let f be strictly convex. Then $\delta(\epsilon) > 0$ for all $\epsilon > 0$.*

Combining these results yields the following non-quantitative version of Proposition C.1.6:

Proposition C.1.10. *Let f be strictly convex and subdifferentiable on the closed convex set C, and assume that the minimum $x_0 = \operatorname{argmin}_{x \in C} f(x)$ is attained. Then the mapping $u \mapsto x_u$ is continuous in a neighborhood of $u = 0$.*

Proof We show first that x_u is continuous at $u = 0$. By Lemmas C.1.8 and C.1.9, we see that for $x \in C$ we have

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle + \delta(\|x - x_0\|) \geq f(x_0) + \delta(\|x - x_0\|),$$

where $g \in \partial f(x_0)$ satisfies $\langle g, x - x_0 \rangle \geq 0$ for all $x \in C$ by the optimality condition (C.1.2). Now, pick $\epsilon > 0$, so that if $\|x - x_0\| > \epsilon$ we have $\delta(\|x - x_0\|) \geq \|x - x_0\| \frac{\delta(\epsilon)}{\epsilon}$ by Lemma C.1.8. Then if u satisfies $\|u\| < \frac{\delta(\epsilon)}{\epsilon}$, we have

$$\begin{aligned}
f(x) - \langle u, x \rangle & = f(x) - \langle u, x - x_0 \rangle - \langle u, x_0 \rangle \\
& \geq f(x_0) - \langle u, x_0 \rangle + \delta(\|x - x_0\|) - \langle u, x - x_0 \rangle \\
& \geq f(x_0) - \langle u, x_0 \rangle + \delta(\|x - x_0\|) - \langle u, x - x_0 \rangle \\
& \geq f(x_0) - \langle u, x_0 \rangle + \delta(\|x - x_0\|) - \|u\| \|x - x_0\| \\
& > f(x_0) - \langle u, x_0 \rangle + \frac{\delta(\epsilon)}{\epsilon} \|x - x_0\| - \frac{\delta(\epsilon)}{\epsilon} \|x - x_0\| = f(x_0) - \langle u, x_0 \rangle.
\end{aligned}$$

Thus any minimizer x_u of $f(x) - \langle u, x \rangle$ over $x \in C$ must satisfy $\|x_u - x_0\| \leq \epsilon$, and strict convexity guarantees its uniqueness.

The argument that $u \mapsto x_u$ is continuous in a neighborhood of zero is completely similar once we recognize that for the divergence $D_f(x, x_0) := f(x) - \langle g, x - x_0 \rangle - f(x_0)$ (where $g \in \partial f(x_0)$ is fixed), we have $D_f = D_{f_u}$ for $f_u(x) = f(x) - \langle u, x \rangle$ and x_u is near x_0 for u small. □

C.2 Conjugacy and duality properties

Attached to any function is its *convex conjugate*, sometimes called the *Fenchel* or *Fenchel-Legendre* conjugate function, defined by

$$f^*(s) := \sup_x \left\{ \langle s, x \rangle - f(x) \right\}. \tag{C.2.1}$$

For any f, the conjugate f^* is a closed convex function, as it is the supremum of linear functions. This function helps to exibit a duality for convex functions similar to those for convex sets, which we can describe as the intersection of all halfspaces containing them (recall Theorem B.1.13 and the equalities (B.1.3)–(B.1.4)).

JCD Comment: Draw a picture of the conjugate

The conjugate function is the largest gap between the linear functional $x \mapsto \langle s, x \rangle$ and the function f itself. The remarkable property of such conjugates is that their biconjugates describe the function f itself, or at least the largest closed convex function below f. To make this a bit more precise, we state a theorem, and then connect to so-called *convex closures* of functions.

Theorem C.2.1. *Let f be closed convex and f^* be its conjugate (C.2.1). Then*

$$f^{**}(x) = f(x) \quad \textit{for all } x.$$

Proof By definition, we have

$$f^{**}(x) = \sup_s \left\{ \langle x, s \rangle - f^*(s) \right\},$$

and we always have $\langle x, s\rangle - f^*(s) \le f(x)$ by definition of $f^*(s) = \sup_x\{\langle s, x\rangle - f(x)\}$. So immediately we see that $f^{**}(x) \le f(x)$.

We essentially show that the linear functions $h_s(x) := \langle x, s\rangle - f^*(s)$ describe (enough) of the global linear underestimators of f so that $f(x) = \sup_s h_s(x)$, allowing us to apply Theorem B.3.7. Indeed, let $l(x) = \langle s, x\rangle + b$ be any global underestimator of f. Then we must have $b \le f(x) - \langle s, x\rangle$ for all x, that is, $b \le \inf_x\{f(x) - \langle s, x\rangle\} = -\sup_x\{\langle s, x\rangle - f(x)\} = -f^*(s)$, that is, $l(x) \le \langle s, x\rangle - f^*(s) = h_s(x)$. Apply Theorem B.3.7. □

We may visualize f^{**} as pulling a string up below a function f, yielding the largest closed convex underestimator of f. (While this is in fact a rigorous statement, we shall not prove it here.)

C.2.1 Gradient dualities and the Fenchel-Young inequality

It is immediate from the definition that for any pair s, x we have the *Fenchel-Young inequality*

$$\langle s, x\rangle \le f^*(s) + f(x). \tag{C.2.2}$$

Even more, combining Theorem C.2.1 with this observation, we can exhibit a duality between subgradients of f and f^* with this inequality.

Proposition C.2.2. *Let f be closed convex. Then*

$$\langle s, x\rangle = f^*(s) + f(x) \quad \textit{if and only if} \quad s \in \partial f(x) \quad \textit{if and only if} \quad x \in \partial f^*(s).$$

Proof If $\langle s, x\rangle = f^*(s) + f(x)$, then $-f(x) + \langle s, x\rangle = f^*(s) \ge \langle s, y\rangle - f(y)$ for all y, and rearranging, we have $f(y) \ge f(x) + \langle s, y - x\rangle$, that is, $s \in \partial f(x)$. Conversely, if $s \in \partial f(x)$ then $0 \in \partial f(x) - s$, so that x minimizes $f(x) - \langle s, x\rangle$, or equivalently, x maximizes $\langle s, x\rangle - f(x)$ and so $\langle s, x\rangle - f(x) = \sup_x\{\langle s, x\rangle - f(x)\}$ as desired. The final statement is immediate from a parallel argument and the duality in Theorem C.2.1. □

Writing Proposition C.2.2 differently, we see that ∂f and ∂f^* are inverses of one another. That is, as set-valued mappings, where

$$(\partial f)^{-1}(s) := \{x \mid s \in \partial f(x)\},$$

we have the following corollary.

Corollary C.2.3. *Let f and f^* be subdifferentiable. Then*

$$\partial f^* = (\partial f)^{-1} \quad \textit{and} \quad \partial f = (\partial f^*)^{-1}$$

and

$$\partial f^*(s) = \operatorname*{argmax}_x \left\{\langle s, x\rangle - f(x)\right\} \quad \textit{and} \quad \partial f(x) = \operatorname*{argmax}_s \left\{\langle s, x\rangle - f^*(s)\right\}.$$

Notably, if f and f^* are differentiable, then $\nabla f = (\nabla f^*)^{-1}$.

Additionally, we see that the domains and images of ∂f and ∂f^* are also related, which guarantees convexity properties of their images as well.

Corollary C.2.4. *Let f be closed convex. Then*

$$\operatorname{dom} \partial f = \operatorname{Im} \partial f^* \quad \textit{and} \quad \operatorname{dom} \partial f^* = \operatorname{Im} \partial f.$$

Proof Let $x \in \operatorname{dom} \partial f$, so that $\partial f(x)$ is non-empty. Then $s \in \partial f(x)$ implies that $\langle s, x\rangle = f(x) + f^*(s)$ and $x \in \partial f^*(s)$ by Proposition C.2.2. Similarly, if $x \in \operatorname{Im} \partial f^*$, then there is some s for which $x \in \partial f^*(s)$ and so $\langle s, x\rangle = f(x) + f^*(s)$ and $s \in \partial f(x)$. □

We can use the identification between the domains of ∂f and the images of ∂f^* to give a few additional characterizations of the domains of convex functions and their conjugates; the domain of f^* is intimately tied with the growth properties of f, and conversely by the relationship $f = f^{**}$ when f is closed convex. As one example of how we can make this identification, note that if f^* is defined everywhere, that is, $\operatorname{dom} f^* = \mathbb{R}^d$, then similarly $\operatorname{dom} \partial f^* = \mathbb{R}^d$, and so in particular the (sub)gradients of f must cover all of $\mathbb{R}^d$. Even more, as we shall see, this implies certain growth conditions on f.

To make this more rigorous, we require functions capturing the asymptotic growth of f. To that end, we present the following proposition, which has the benefit of defining the *recession function* (essentially, an asymptotic derivative) of f.

Proposition C.2.5. *Let f be a closed convex function and f^* is convex conjugate. Then for any $x \in \operatorname{dom} f$, we may define*

$$f'_\infty(v) := \sup_{t>0} \frac{f(x+tv) - f(x)}{t} = \lim_{t \to \infty} \frac{f(x+tv) - f(x)}{t} \tag{C.2.3}$$

independently of x, and moreover,

$$f'_\infty(v) = \sigma_{\operatorname{dom} f^*}(v)$$

where $\sigma_{\operatorname{dom} f^*}$ is the support function (B.2.1) of $\operatorname{dom} f^*$.

Proof That for any fixed $x \in \operatorname{dom} f$ the limit exists and is equal to the supremum follows because of the criterion of increasing slopes (B.3.4), making the equality with the supremum immediate. That $f'_\infty(v)$ is independent of x will follow once we show the second equality claimed in the proposition, to which we now turn.

Recall that

$$\operatorname{dom} f^* = \left\{ s \mid \sup_x \{\langle s, x\rangle - f(x)\} < \infty \right\} \quad \text{and} \quad f(x) = \sup_s \left\{\langle s, x\rangle - f^*(s)\right\}$$

by conjugate duality, as f is closed convex (Theorem C.2.1). Fix $x \in \operatorname{dom} f$. Then for any $s \in \operatorname{dom} f^*$, we evidently have

$$\frac{f(x+tv) - f(x)}{t} \geq \frac{\langle s, x+tv\rangle - f^*(s) - f(x)}{t} \to \langle s, v\rangle$$

as $t \uparrow \infty$. Taking a supremum over $s \in \operatorname{dom} f^*$ gives that $f'_\infty(v) \geq \sigma_{\operatorname{dom} f^*}(v)$. For the opposite direction, note that

$$\begin{aligned}
\frac{f(x+tv) - f(x)}{t} &= \frac{1}{t}\left[\sup_{s \in \operatorname{dom} f^*} \{\langle s, x+tv\rangle - f^*(s)\} - \sup_{s \in \operatorname{dom} f^*} \{\langle s, x\rangle - f^*(s)\}\right] \\
&\leq \frac{1}{t} \sup_{s \in \operatorname{dom} f^*} \{\langle s, x+tv\rangle - f^*(s) - (\langle s, x\rangle - f^*(s))\} = \frac{1}{t} \sup_{s \in \operatorname{dom} f^*} t\langle s, v\rangle.
\end{aligned}$$

Thus $f'_\infty(v) \le \sigma_{\dom f^*}(v)$, and we have the result. □

It is particularly interesting to understand the conditions under which $\dom f^* = \mathbb{R}^d$, that is, f^* is finite everywhere, and relatedly, under which the function $x \mapsto f(x) - \langle s, x\rangle$ has a minimizer. Recall that $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is *coercive* if $f(x) \to \infty$ whenever $\|x\| \to \infty$, so that if f is closed convex, then the tilted function $f(\cdot) - \langle s, \cdot\rangle$ has a minimizer if and only if it is coercive. We call f *super-coercive* if $f(x)/\|x\| \to \infty$ whenever $\|x\| \to \infty$, so that f grows more than linearly. These concepts are central to the existence of minimizers. A priori, any function with compact domain is super-coercive, because $f(x) = +\infty$ for $x \notin \dom f$. For convex functions, we can relate such coercivity ideas to the recession function f'_∞ associated with f as expression (C.2.3) defines. Particularly important are those f satisfying

$$f'_\infty(v) = +\infty \quad \text{for all } v \neq 0,$$

a class Rockafellar [153] calls *copositive* functions, as these exhibit superlinear growth on all rays toward infinity. We can relate this condition to the domains of f^* as well: using Proposition B.2.7, Proposition C.2.5 gives

Corollary C.2.6. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex. Then $s \in \dom f^*$ if and only if $\langle s, v\rangle \le f'_\infty(v)$ for all $v \neq 0$, and $s \in \operatorname{int}\dom f^*$ if and only if $\langle s, v\rangle < f'_\infty(v)$ for all $v \neq 0$. In particular,*

(i) If $f'_\infty(v) > 0$ for all $v \neq 0$, then $0 \in \operatorname{int}\dom f^*$ and f has a minimizer. A sufficient condition for this is that f be coercive.

(ii) We have $f'_\infty(v) = +\infty$ for all $v \neq 0$ if and only if

$$\dom f^* = \mathbb{R}^d.$$

A sufficient condition for this is that f be super-coercive.

Proof Combine Propositions B.2.7 and C.2.5: for part (i), note that if $f'_\infty(v) > 0$ for all $v \neq 0$, then $0 \in \operatorname{int}\dom f^*$, and so f^* has a non-trivial subdifferential $\partial f^*(0)$ at 0; letting $x \in \partial f^*(0)$ we have $x \in \operatorname{argmin} f$. To see that coercivity is sufficient, note that if $f(x) \to \infty$ whenever $\|x\| \to \infty$, the criterion of increasing slopes (B.3.4) gives $f'_\infty(v) > 0$ for all $v \neq 0$. Part (ii) is similarly immediate. □

C.2.2 Smoothness and strict convexity of conjugates

The dualities in derivative mappings extend to various smoothness dualities, which can be quite useful as well. These types of results build from the stability properties of solution mappings, as in those for tilted minimizers (C.1.4) in Propositions C.1.6 and C.1.10. They also relate different smoothness properties of f and f^*, as well as their domains of definition, to the existence and continuity of minimizers of $f(x) - \langle s, x\rangle$.

When we assume that f has quantitative strong convexity or smoothness properties, we can give similar quantitative guarantees for the smoothness and strong convexity of f^*:

Proposition C.2.7. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be λ-strongly convex with respect to the norm $\|\cdot\|$ (see Eq. (C.1.3)) on its domain. Then $\operatorname{dom} f^* = \mathbb{R}^d$ and ∇f^* is $\frac{1}{\lambda}$-Lipschitz continuous with respect to the dual norm $\|\cdot\|_*$, that is,*

$$\|\nabla f^*(u) - \nabla f^*(v)\| \le \frac{1}{\lambda} \|u - v\|_*$$

for all u, v. Conversely, let $f : \mathbb{R}^d \to \mathbb{R}$ be convex with L-Lipschitz gradient with respect to $\|\cdot\|$ on $\mathbb{R}^d$. Then f^* is $\frac{1}{L}$-strongly convex with respect to the dual norm $\|\cdot\|_*$ on convex subsets $C \subset \operatorname{dom} \partial f^*$, and in particular,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_*^2 . \tag{C.2.4}$$

Proof For the first claim, let $C = \operatorname{dom} f$. Then Proposition C.1.6 shows that if $x_1 = \operatorname{argmin}_x\{f(x) - \langle s_1, x\rangle\}$ and $x_2 = \operatorname{argmin}_x\{f(x) - \langle s_2, x\rangle\}$ (which exist and are necessarily unique), we have $\|x_1 - x_2\| \le \frac{1}{\lambda} \|s_1 - s_2\|_*$. Then Proposition C.2.2 shows that $x_i \in \partial f^*(s_i)$ for $i = 1, 2$, and hence $\partial f^*(s_i)$ is necessarily single-valued and $(1/\lambda)$-Lipschitz continuous.

The converse is a bit trickier. Let x and y be arbitrary and $s_x = \nabla f(x)$ and $s_y = \nabla f(y)$; we prove inequality (C.2.4), known as *co-coercivity*. By the L-Lipschitz continuity of ∇f, we have

$$\begin{aligned}
f(y) &= f(x) + \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\
&\le f(x) + \langle s_x, y - x \rangle + \int_0^1 Lt \|y - x\|^2 dt = f(x) + \langle s_x, y - x \rangle + \frac{L}{2} \|y - x\|^2 ,
\end{aligned}$$

which is valid for any x, y. Note that $f(x) - \langle s_x, x\rangle = -f^*(s_x)$, so that rearranging we have

$$\begin{aligned}
f^*(s_x) \le \langle s_x, y\rangle - f(y) + \frac{L}{2} \|y - x\|^2 &= \langle s, y\rangle - f(y) + \langle s_x - s, y\rangle + \frac{L}{2} \|y - x\|^2 \\
&\le f^*(s) + \langle s_x - s, y\rangle + \frac{L}{2} \|y - x\|^2 ,
\end{aligned}$$

valid for any vector s and any y. We may in particular take an infimum over y on the right hand side, where

$$\begin{aligned}
\inf_y \langle s_x - s, y\rangle + \frac{L}{2} \|y - x\|^2 &= \inf_y \langle s_x - s, y - x\rangle + \frac{L}{2} \|y - x\|^2 + \langle s_x - s, x\rangle \\
&\overset{(\star)}{=} \inf_t \left\{ t \|s_x - s\|_* + \frac{Lt^2}{2} \right\} + \langle s_x - s, x\rangle \\
&= -\frac{1}{2L} \|s_x - s\|_*^2 + \langle s_x - s, x\rangle,
\end{aligned}$$

where equality $(\star)$ follows by definition of the dual norm and we identify $t = \|y - x\|$. Thus

$$f^*(s_x) + \langle x, s - s_x\rangle + \frac{1}{2L} \|s - s_x\|_*^2 \le f^*(s)$$

for all s. As $x \in \partial f^*(s_x)$, Proposition C.1.5(ii) gives the strong convexity result. The rest is algebraic manipulations with $s_y = \nabla f(y)$ and an application of Proposition C.1.5(iii). □

There are more qualitative versions of Proposition C.2.7 that allow us to give a duality between strict convexity and continuous differentiability of f. Here, we give one typical result.

Proposition C.2.8. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be strictly convex and closed. Then* $\operatorname{int}\operatorname{dom} f^* \neq \varnothing$ *and f^* is continuously differentiable on* $\operatorname{int}\operatorname{dom} f^*$. *Conversely, let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be differentiable on $\Omega := \operatorname{int}\operatorname{dom} f$. Then f^* is strictly convex on each convex $C \subset \nabla f(\Omega)$.*

These results should be roughly expected becaues of the duality that $\nabla f = (\nabla f^*)^{-1}$ and that $\partial f^*(s) = \operatorname{argmin}_x\{\langle s, x\rangle - f(x)\}$, because strict convexity guarantees uniqueness of minimizers (Proposition C.1.1) so that ∂f^* should be a singleton.

Proof To see that $\operatorname{int}\operatorname{dom} f^*$ is non-empty, we use the identification $f'_\infty(v) = \sigma_{\operatorname{dom} f^*}(v)$ in Proposition C.2.5 and the interior identification in Proposition B.2.7. Because f is strictly convex, for any $x \in \operatorname{dom} f$ we have

$$0 < \frac{f(x - tv) - f(x)}{t} + \frac{f(x + td) - f(x)}{t} \quad \text{for } t > 0,$$

and taking $t \to \infty$ gives $0 < f'_\infty(-v) + f'_\infty(v)$. Proposition B.2.7 then shows that $\operatorname{int}\operatorname{dom} f^* \neq \varnothing$.

For the claim that f^* is continuously differentiable, take $s \in \operatorname{int}\operatorname{dom} f^*$, and suppose for the sake of contradiction that $\partial f^*(s)$ has distinct points x_1, x_2. Then Corollary C.2.3 gives that x_1 and x_2 both minimize $f(x) - \langle s, x\rangle$ over x. But Proposition C.1.1 guarantees $x_1 = x_2$, so that $\partial f^*(s) = \{\nabla f^*(s)\}$ is a singleton, and hence f^* is continuous differentiable at s (Proposition B.3.22).

For the converse claim, let C be a convex set as stated. Suppose for the sake of contradiction that f^* is not strictly convex on C, so that there are distinct points $s_1, s_2 \in C$ for which f^* is affine on the line segment $[s_1, s_2] = \{ts_1 + (1-t)s_2 \mid t \in [0, 1]\}$. As $C \subset \nabla f(\Omega)$ is convex, the midpoint $s = \frac{1}{2}(s_1 + s_2) \in C$ and there exists x satisfying $\nabla f(x) = s$, or $x \in \partial f^*(s)$. Then because f^* is assumed affine on $[s_1, s_2]$, we have $f^*(s) = \frac{1}{2}f^*(s_1) + \frac{1}{2}f^*(s_2)$ and $\langle s, x\rangle = \frac{1}{2}\langle s_1 + s_2, x\rangle$, so

$$\begin{aligned} 0 &= f(x) + f^*(s) - \langle s, x\rangle \\ &= \frac{1}{2}\left[(f(x) + f^*(s_1) - \langle s_1, x\rangle) + (f(x) + f^*(s_2) - \langle s_2, x\rangle)\right]. \end{aligned}$$

Each of the terms in parenthesis is 0 if and only if $s_i \in \partial f(x)$, but by assumption $\partial f(x) = \{\nabla f(x)\}$ is a singleton, and we must have $s_1 = s_2$. □

JCD Comment: Better transition

We close this section by investigating particularly nice classes of functions f, where f and its conjugate f^* are strictly convex and smooth. These results are central to the various conjugate linkage dualities we explore in Chapter 11.3. We therefore make the following definition:

Definition C.2. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex. Then f is of* Legendre type *if*

(i) $\operatorname{int}\operatorname{dom} f \neq \varnothing$

(ii) f is continuously differentiable on $\operatorname{int}\operatorname{dom} f$

(iii) f is strictly convex

(iv) f satisfies the gradient boundary conditions

$$\|\nabla f(x)\| \to \infty \quad \text{as} \quad x \to \operatorname{bd}\operatorname{dom} f \quad \text{or} \quad \|x\| \to \infty. \tag{C.2.5}$$

Thus, at the boundaries of their domains or as their argument tends off to infinity, functions of Legendre type have slopes tending to ∞. This does not guarantee that $f(x) \to \infty$ as $x \to \operatorname{bd}\operatorname{dom} f$, though it does provide guarantees of regularity that the next theorem highlights.

Theorem C.2.9. *Let f be a convex function of Legendre type (Def. C.2). Then f^* is strictly convex, continuously differentiable, and $\operatorname{dom} f^* = \mathbb{R}^d$.*

The theorem implies a number of results on continuity of minimizers and tilted minimizers (C.1.4), clarifying some of our earlier results. For example, we have the following corollary.

Corollary C.2.10. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a convex function of Legendre type. Then the tilted minimizer*

$$x_u := \operatorname{argmin}\{f(x) - \langle u, x\rangle\}$$

exists for all u, is continuous in u and unique, and $x_u \in \operatorname{int}\operatorname{dom} f$.

We turn to the proof of the theorem.

Proof of Theorem C.2.9 We an intermediate lemma, whose proof we defer.

Lemma C.2.11. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex and satisfy the gradient boundary condition that $\|s_n\| \to \infty$ for any sequence $x_n \to \operatorname{bd}\operatorname{dom} f$ and $s_n \in \partial f(x_n)$. Then*

$$f'_\infty(v) = \infty \text{ for all } v \neq 0$$

if and only if

$$\|s_n\| \to \infty \quad \text{whenever} \quad \|x_n\| \to \infty \text{ and } s_n \in \partial f(x_n).$$

The theorem follows straightforwardly from Lemma C.2.11. By the boundary conditions (C.2.5) associated with f, we have $f'_\infty(v) = \infty$ for all $v \neq 0$ (Lemma C.2.11). Because the support function of $\operatorname{dom} f^*$ satisfies $\sigma_{\operatorname{dom} f^*} = f'_\infty$ (Proposition C.2.5), we see that $\operatorname{dom} f^* = \mathbb{R}^d$ as $\operatorname{dom} f^* = \{s \mid \langle s, v\rangle \le \sigma_{\operatorname{dom} f^*}(v) \text{ for all } v\}$ (e.g., Proposition B.2.7 or Corollary B.2.2). With this, f^* is continuously differentiable and strictly convex on its domain (Proposition C.2.8). □

Proof of Lemma C.2.11 As $\operatorname{dom} f^* = \mathbb{R}^d$ if and only if $f'_\infty(v) = \infty$ for all $v \neq 0$ (Corollary C.2.6), it suffices to show the result that $\operatorname{int}\operatorname{dom} f^* \neq \mathbb{R}^d$ if and only if there exists an unbounded sequence x_n with $s_n \in \partial f(x_n)$ and for which s_n is convergent.

Let us begin with the unbounded sequence x_n for which $s_n \to s \in \mathbb{R}^d$; assume for the sake of contradiction that $s \in \operatorname{int}\operatorname{dom} f^*$. Because $s_n \in \partial f(x_n)$, we have $x_n \in \partial f^*(s_n)$ by Proposition C.2.2. The assumption that $s \in \operatorname{int}\operatorname{dom} f^*$ means that there exists an $\epsilon > 0$ such that $s + \epsilon\mathbb{B} \subset \operatorname{int}\operatorname{dom} f^*$ and f^* is Lipschitz on $s + \epsilon\mathbb{B}$ (Theorem B.3.4). But then $\partial f^*(s + \epsilon\mathbb{B})$ is bounded, and $x_n \in \partial f^*(s_n) \subset \partial f^*(s + \epsilon\mathbb{B})$ for large enough n, contradicting that $\|x_n\| \to \infty$, and so $s \notin \operatorname{int}\operatorname{dom} f^*$ and $\operatorname{int}\operatorname{dom} f^* \neq \mathbb{R}^d$.

Now let us assume that $\operatorname{int}\operatorname{dom} f^* \neq \mathbb{R}^d$. Let $s \in \operatorname{bd}\operatorname{dom} f^*$. Then either $\partial f^*(s) = \varnothing$ or $\partial f^*(s)$ is unbounded (Lemma B.3.19). If $\partial f^*(s) = \varnothing$, take $s_n \to s$ with $s_n \in \operatorname{relint}\operatorname{dom} f^*$, and let $x_n \in \partial f^*(s_n)$. We show that x_n must be unbounded. If x_n is bounded, then by passing to a subsequence if necessary we may assume $x_n \to x$, and the outer semicontinuity of the subdifferential (Lemma B.3.21) gives $x \in \partial f^*(s)$, contradicting that $\partial f^*(s) = \varnothing$. Thus we must have x_n unbounded, which is thus the desired unbounded sequence. On the other hand, if $\partial f^*(s)$ is unbounded, we can simply take $x_n \in \partial f^*(s)$ with $s \in \partial f(x_n)$ for each n, which is the desired

convergent sequence. □

As a last application of these ideas, in some cases we wish to allow constraints on the functions f to be minimized, returning to the original convex optimization problem (C.0.1) with f a function of Legendre type and C a closed convex set. We then have the following corollary.

Corollary C.2.12. *Let f be of Legendre type (Definition C.2) and $C \subset \mathbb{R}^d$ a closed convex set with $\operatorname{int}\operatorname{dom} f \cap C \neq \emptyset$. Define $f_C(x) = f(x) + \mathbf{I}_C(x)$. Then*

(i) f_C^* is continuously differentiable,

(ii) $\operatorname{dom} f_C^* = \mathbb{R}^d$, and

(iii) the constrained tilted minimizers

$$x_u = \operatorname*{argmin}_{x \in C} \{\langle u, x\rangle - f(x)\}$$

are unique, continuous in u, belong to $\operatorname{int}\operatorname{dom} f$, and satisfy

$$x_u = \nabla f^*(u - v) \quad \text{and} \quad \nabla f(x_u) = -v$$

for some vector $v \in \mathcal{N}_C(x_u)$.

Proof The function $f_C := f + \mathbf{I}_C$ is closed convex. To show that $\operatorname{dom} f_C^* = \mathbb{R}^d$, we can equivalently show that $(f_C)'_\infty(v) = \infty$ for all non-zero v. Because f is Legendre-type, Lemma C.2.11 guarantees that if $x \in \operatorname{dom} f \cap C$, then

$$(f_C)'_\infty(v) = \lim_{t \uparrow \infty} \frac{f(x + tv) + \mathbf{I}_C(x + tv) - f(x)}{t} \geq \lim_{t \uparrow \infty} \frac{f(x + tv) - f(x)}{t} = f'_\infty(v) = \infty.$$

So $\operatorname{dom} f_C^* = \mathbb{R}^d$, and thus $x_u \operatorname{argmin}_{x \in C}\{f(x) - \langle u, x\rangle\} = \nabla f_C^*(u)$ exists and is unique and continuous, as f is strictly convex.

By the standard subgradient conditions for optimality, the vector x_u is characterized by

$$0 \in \nabla f(x_u) - u + \mathcal{N}_C(x_u),$$

and so $x_u \in \operatorname{int}\operatorname{dom} f$ (as otherwise $\|\nabla f(x_u)\| = +\infty$ by Definition C.2) and

$$x_u = \nabla f^*(u - v)$$

for some vector $v \in \mathcal{N}_C(x_u)$. □

JCD Comment: Now do the particular case that we define $f_C = f + \mathbf{I}_C$ where C is an affine space. Then we should still have $\operatorname{dom} f^* = \mathbb{R}^d$, and ∇f_C^* exists, and should get *some* good dualities. Work it out!

JCD Comment: More smoothness dualities, and write an exercise? Perhaps uniform convexity versions and strict convexity versions.

a. Closed convex function as a supremum of affine functions minorizing it

b. Fenchel Conjugate functions f^*

c. Fenchel biconjugate

Further reading

There are a variety of references on the topic, beginning with the foundational book by Rockafellar [153], which initiated the study of convex functions and optimization in earnest. Since then, a variety of authors have written (perhaps more easily approachable) books on convex functions, optimization, and their related calculus. Hiriart-Urruty and Lemaréchal [104] have written two volumes explaining in great detail finite-dimensional convex analysis, and provide a treatment of some first-order algorithms for solving convex problems. Borwein and Lewis [33] and Luenberger [133] give general treatments that include infinite-dimensional convex analysis, and Bertsekas [27] gives a variety of theoretical results on duality and optimization theory.

There are, of course, books that combine theoretical treatment with questions of convex modeling and procedures for solving convex optimization problems (problems for which the objective and constraint sets are all convex). Boyd and Vandenberghe [35] gives a very readable treatment for those who wish to use convex optimization techniques and modeling, as well as the basic results in convex analytic background and duality theory. Ben-Tal and Nemirovski [23], as well as Nemirovski's various lecture notes, give a theory of the tractability of computing solutions to convex optimization problems as well as methods for solving them.

C.3 Exercises

Exercise C.1: Show that the alternative increasing slopes condition (B.3.5) is equivalent to convexity of f.

Exercise C.2: Do the uniform convexity version of Proposition C.1.5.

Exercise C.3: Do the uniform convexity version of Proposition C.1.6.

Bibliography

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66:671–687, 2003.

[2] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the Twenty Fifth Annual Conference on Computational Learning Theory*, 2012.

[3] N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[4] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

[5] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[6] A. Andoni, M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni. Locality-sensitive hashing using stable distributions. In T. Darrell, P. Indyk, and G. Shakhnarovich, editors, *Nearest Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press, 2006.

[7] E. Arias-Castro, E. Candés, and M. Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2013.

[8] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[9] S. Artstein, K. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004.

[10] P. Assouad. Deux remarques sur l'estimation. *Comptes Rendus des Séances de l'Académie des Sciences, Série I*, 296(23):1021–1024, 1983.

[11] J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. In *Journal of Machine Learning Research*, pages 2635–2686, 2010.

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[13] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[14] B. Balle, G. Barthe, and M. Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems 31*, pages 6277–6287, 2018.

[15] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[16] A. Barron. Entropy and the central limit theorem. *Annals of Probability*, 14(1):336–342, 1986.

[17] A. R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics*, pages 561–576. Kluwer Academic, 1991.

[18] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[19] P. L. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[20] R. Bassily, A. Smith, T. Steinke, and J. Ullman. More general queries and less generalization error in adaptive data analysis. *arXiv:1503.04843 [cs.LG]*, 2015.

[21] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the Forty-Eighth Annual ACM Symposium on the Theory of Computing*, pages 1046–1059, 2016.

[22] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[23] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization*. SIAM, 2001.

[24] D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18:1–7, 2018.

[25] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Annals of Statistics*, 41(2):802–837, 2013.

[26] J. M. Bernardo. Reference analysis. In D. Day and C. R. Rao, editors, *Bayesian Thinking, Modeling and Computation*, volume 25 of *Handbook of Statistics*, chapter 2, pages 17–90. Elsevier, 2005.

[27] D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[28] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwebte Gebiet*, 65:181–238, 1983.

[29] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Transactions on Information Theory*, 51(4):1611–1614, 2005.

[30] L. Birgé and P. Massart. Estimation of integral functionals of a density. *Annals of Statistics*, 23(1):11–29, 1995.

[31] J. Blasiok, P. Gopalan, L. Hu, and P. Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the Fifty-Fifth Annual ACM Symposium on the Theory of Computing*, 2023. URL https://arxiv.org/abs/2211.16886.

[32] J. Blasiok, P. Gopalan, L. Hu, and P. Nakkiran. When does optimizing a proper loss yield calibration? In *Advances in Neural Information Processing Systems 36*, 2023. URL https://arxiv.org/abs/2305.18764.

[33] J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, 2006.

[34] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[35] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[36] S. Boyd, J. Duchi, and L. Vandenberghe. Subgradients. Course notes for Stanford Course EE364b, 2015. URL http://web.stanford.edu/class/ee364b/lectures/subgradients_notes.pdf.

[37] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the Forty-Eighth Annual ACM Symposium on the Theory of Computing*, 2016. URL https://arxiv.org/abs/1506.07216.

[38] G. Brown, M. Bun, V. Feldman, A. Smith, and K. Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the Fifty-Third Annual ACM Symposium on the Theory of Computing*, pages 123–132, 2021.

[39] L. D. Brown. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, California, 1986.

[40] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[41] V. Buldygin and Y. Kozachenko. *Metric Characterization of Random Variables and Random Processes*, volume 188 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[42] T. Cai and M. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Annals of Statistics*, 39(2):1012–1041, 2011.

[43] E. J. Candès and M. A. Davenport. How well can we estimate a sparse vector. *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

[44] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes and Monographs*. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007. URL https://arxiv.org/abs/0712.0248.

[45] O. Catoni and I. Giulini. Dimension-free PAC-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv:1712.02747 [math.ST]*, 2017.

[46] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Annals of Statistics*, 27(6):1865–1895, 1999.

[47] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[48] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2–3):321–352, 2007.

[49] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

[50] J. E. Cohen, Y. Iwasa, G. Rautu, M. B. Ruskai, E. Seneta, and G. Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear Algebra and its Applications*, 179:211–235, 1993.

[51] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[52] J. Couzin. Whole-genome data not anonymous, challenging assumptions. *Science*, 321(5894): 1278, 2008.

[53] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[54] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientifica Mathematica Hungary*, 2:299–318, 1967.

[55] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, second edition, 2011.

[56] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15: 429–444, 1977.

[57] S. Dasgupta and A. Gupta. An elementray proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2002.

[58] L. D. Davisson. The prediction error of stationary gaussian time series of unknown covariance. *IEEE Transactions on Information Theory*, 11:527–532, 1965.

[59] A. Dawid and V. Vovk. Prequential probability: principles and properties. *Bernoulli*, 5: 125–162, 1999.

[60] M. H. DeGroot. *Optimal Statistical Decisions*. Mcgraw-Hill College, 1970.

[61] P. Del Moral, M. Ledoux, and L. Miclo. On contraction properties of Markov kernels. *Probability Theory and Related Fields*, 126:395–420, 2003.

[62] R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability and Its Applications*, 1(1):65–80, 1956.

[63] R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. II. *Theory of Probability and Its Applications*, 1(4):329–383, 1956.

[64] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence I. Technical Report 137, University of California, Berkeley, Department of Statistics, 1987.

[65] J. C. Duchi and R. Rogers. Lower bounds for locally private estimation via communication complexity. In *Proceedings of the Thirty Second Annual Conference on Computational Learning Theory*, 2019.

[66] J. C. Duchi and F. Ruan. A constrained risk inequality for general losses. In *Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics*, 2020.

[67] J. C. Duchi and M. J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. *arXiv:1311.2669 [cs.IT]*, 2013.

[68] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy, data processing inequalities, and minimax rates. *arXiv:1302.3203 [math.ST]*, 2013. URL http://arxiv.org/abs/1302.3203.

[69] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013.

[70] J. C. Duchi, K. Khosravi, and F. Ruan. Multiclass classification, information, divergence, and surrogate risk. *Annals of Statistics*, 46(6b):3246–3275, 2018.

[71] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

[72] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3 & 4):211–407, 2014.

[73] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology (EUROCRYPT 2006)*, 2006.

[74] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference*, pages 265–284, 2006.

[75] C. Dwork, G. N. Rothblum, and S. P. Vadhan. Boosting and differential privacy. In *51st Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010.

[76] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. *arXiv:1411.2664v2 [cs.LG]*, 2014.

[77] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM Symposium on the Theory of Computing*, 2015.

[78] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. The reusable holdout: Preserving statistical validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[79] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

[80] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, pages 211–222, 2003.

[81] K. Fan. Minimax theorems. *Proceedings of the National Academy of Sciences*, 39(1):42–47, 1953.

[82] V. Feldman and T. Steinke. Calibrating noise to variance in adaptive data analysis. In *Proceedings of the Thirty First Annual Conference on Computational Learning Theory*, 2018. URL http://arxiv.org/abs/1712.07196.

[83] G. Folland. *Real Analysis: Modern Techniques and their Applications*. Pure and Applied Mathematics. John Wiley & Sons, second edition, 1999.

[84] D. Foster and R. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

[85] D. P. Foster and S. Hart. "calibeating": Beating forecasters at their own game. *arXiv:2209.0489 [econ.TH]*, 2022.

[86] A. Franco, N. Malhotra, and G. Simonovits. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505, 2014.

[87] D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, Feb. 1975.

[88] R. Gallager. Source coding with side information and universal coding. Technical Report LIDS-P-937, MIT Laboratory for Information and Decision Systems, 1979.

[89] D. García-García and R. C. Williamson. Divergences and risks for multiclass experiments. In *Proceedings of the Twenty Fifth Annual Conference on Computational Learning Theory*, 2012.

[90] A. Garg, T. Ma, and H. L. Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems 27*, 2014.

[91] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Technical report, Columbia University, 2013.

[92] R. P. Gilbert. *Function Theoretic Methods in Partial Differential Equations*. Academic Press, 1969.

[93] T. Gneiting and A. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[94] R. M. Gray. *Entropy and Information Theory*. Springer, 1990.

[95] P. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[96] P. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32(4):1367–1433, 2004.

[97] A. Guntuboyina. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

[98] L. Györfi and T. Nemetz. f-dissimilarity: A generalization of the affinity of several distributions. *Annals of the Institute of Statistical Mathematics*, 30:105–113, 1978.

[99] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[100] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory of Probability and Applications*, 23:794–798, 1978.

[101] D. Haussler. A general minimax result for relative entropy. *IEEE Transactions on Information Theory*, 43(4):1276–1280, 1997.

[102] D. Haussler, J. Kivinen, and M. K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.

[103] E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2006.

[104] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

[105] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001.

[106] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, Mar. 1963.

[107] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):e1000167, 2008.

[108] P. J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

[109] K. Hung and W. Fithian. Statistical methods for replicability assessment. *Annals of Applied Statistics*, 14(3):1063–1087, 2020.

[110] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, 1981.

[111] P. Indyk. Nearest neighbors in high-dimensional spaces. In *Handbook of Discrete and Computational Geometry*. CRC Press, 2004.

[112] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing*, 1998.

[113] J. P. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8), 2005. doi: 10.1371/journal.pmed.0020124.

[114] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70 (9):939–952, Sept. 1982.

[115] T. S. Jayram. Hellinger strikes back: a note on the multi-party information complexity of AND. In *Proceedings of APPROX and RANDOM 2009*, volume 5687 of *Lecture Notes in Computer Science*, pages 562–573. Springer, 2009.

[116] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A: Mathematical and Physical Sciences*, 186:453–461, 1946.

[117] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[118] M. J. Kearns and L. Saul. Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 311–319, 1998.

[119] J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–64, Jan. 1997.

[120] J. Kivinen and M. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45(3):301–329, July 2001.

[121] A. Kolmogorov and V. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *Uspekhi Matematischeskikh Nauk*, 14(2):3–86, 1959.

[122] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer-Verlag, 2011.

[123] A. Kumar, P. Liang, and T. Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems 32*, 2019.

[124] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[125] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[126] J. Langford and R. Caruana. (not) bounding the true error. In *Advances in Neural Information Processing Systems 14*, 2001.

[127] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

[128] L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[129] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[130] E. L. Lehmann and G. Casella. *Theory of Point Estimation, Second Edition*. Springer, 1998.

[131] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[132] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[133] D. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

[134] M. Madiman and A. Barron. Generalized entropy power inequalities and monotonicity properties of information. *IEEE Transactions on Information Theory*, 53(7):2317–2329, 2007.

[135] D. A. McAllester. Some PAC-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998.

[136] D. A. McAllester. Simplified PAC-bayesian margin bounds. In *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory*, pages 203–215, 2003.

[137] D. A. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

[138] D. A. McAllester. A PAC-Bayestian tutorial with a dropout bound. *arXiv:1307.2118 [cs.LG]*, 2013.

[139] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *48th Annual Symposium on Foundations of Computer Science*, 2007.

[140] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[141] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236, 2005.

[142] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[143] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[144] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37(2):876–904, 2009.

[145] B. T. Polyak and J. Tsypkin. Robust identification. *Automatica*, 16:53–63, 1980. doi: 10.1016/0005-1098(80)90086-2. URL http://dx.doi.org/10.1016/0005-1098(80)90086-2.

[146] Y. Polyanskiy and Y. Wu. Strong data-processing inequalities for channels and Bayesian networks. In *Convexity and Concentration*, volume 141 of *The IMA Volumes in Mathematics and its Applications*, pages 211–249. Springer, 2017.

[147] M. Raginsky. Strong data processing inequalities and ϕ-Sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

[148] A. Rao and A. Yehudayoff. *Communication Complexity and Applications*. Cambridge University Press, 2020.

[149] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976—6994, 2011.

[150] M. Reid and R. Williamson. Information, divergence, and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

[151] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30:629–636, 1984.

[152] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, 55:527–535, 1952.

[153] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[154] H. Royden. *Real Analysis*. Pearson, third edition, 1988.

[155] D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, page To appear, 2014.

[156] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems 27*, 2014.

[157] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[158] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.

[159] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Programming Society, 2009.

[160] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[161] A. Slavkovic and F. Yu. Genomics and privacy. *Chance*, 28(2):37–39, 2015.

[162] J. Steinhardt and P. Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[163] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779, 2015.

[164] T. Tao. *An Epsilon of Room, I: Real Analysis (pages from year three of a mathematical blog)*, volume 117 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010.

[165] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[166] R. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514): 600–620, 2016.

[167] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[168] J. W. Tukey. *Exploratory Data Analysis*. Pearson, 1997.

[169] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[170] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

[171] C. Villani. *Optimal Transport: Old and New*. Springer, 2009.

[172] A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10(4):299–326, 1939.

[173] S. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[174] Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[175] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[176] A. C.-C. Yao. Some complexity questions related to distributive computing (preliminary report). In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing*, pages 209–213. ACM, 1979.

[177] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

[178] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed estimation with communication constraints. In *Advances in Neural Information Processing Systems 26*, 2013.