STATS 311: Information Theory and Statistics

Lecturer: Professor John Duchi

Notes by: Andrew Lin

Autumn 2025

Introduction

All of the information can be found on the course website – besides that, we'll have an Ed and Gradescope webpage for various logistics. The course will have four problem sets (weighted at 50%) and "either" a final project or a final exam. The material will generally follow the course textbook, which is getting ready for publication and available on the course webpage.

Some "convex combination" of Stats 300A, CS 229 (machine learning), or EE 276 or 364 will be good background for understanding this course overall.

Remark 1. *As a note, something like ChatGPT can solve many of the problems in this course, but we probably shouldn't do that since it'll make it harder for us to learn the material. We should treat the problem sets like crossword puzzles where "it's fun to try to solve" and it's good for us.*

The big goal of the course overall is to connect information theory and its tools with statistics and machine learning. These two fields tend to provide complementary perspectives, since the idea of information theory is that we design signals that we send so that they're easy to decode (or easy to send over noisy channel), and the idea of statistics is to extract information from what we are given (though we don't get as much choice in what we have to work with). But in both cases, we try to get optimal estimators.

For a general outline, we'll begin the first part of the course with **information, stability, and generalization**. That is, we'll take some sample $\{x_1, \cdots, x_n\}$ and want to understand whether it "looks like the population" p. The next parts will be about **fundamental limits and optimality** (getting lower bounds and showing that certain techniques are unimprovable) and then **losses and predictions** (that is, analyzing loss functions that measure the performance of procedures, and getting operational, engineering understanding of things like entropy). We'll then move on to **online learning and games**. The unifying thread of all of these is the big question animating statistics and information theory: "how do distributions get close together?".

1 September 23, 2025

We'll start with a "blitz of the basics of information theory," not spending much time but having some unified language to talk about things in the rest of the course. (For those of us who have taken information theory before, note that everything is base e instead of base 2.) The starting point is always entropy:

Definition 2

Let P be a distribution on a finite or countable set $\mathcal{X}$, and write $p(x) = \mathbb{P}(X = x)$. The **entropy** of X is given by

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

We clearly have $H(X) \geq 0$ (since $\log p$ is always negative) and in fact we'll see later that $H(X) \leq \log |X|$, with the maximum achieved for the uniform distribution $p(x) = \frac{1}{|\mathcal{X}|}$. For example for a binary random variable which takes value 1 with probability p, the entropy is maximized at $\frac{1}{2}$ and approaches 0 as $p \to 0$ or $p \to 1$.

Definition 3

Let X, Y be two random variables. The **conditional entropy** is first defined by letting

$$H(X|Y = y) = -\sum_x p(x|y) \log p(x|y),$$

and then summing over the values of Y:

$$H(X|Y) = \sum_y p(y) H(X|Y = y).$$

We can think of entropy as "uncertainty," and conditional entropy is then the "uncertainty left on average in X after observing Y."

The funny thing about entropy is that it's "not real," in that it's an object that has to assume that random variables are discrete. There's no entropy of a general probability distribution, so what we really should be talking about are divergences (which "actually exist" no matter what). For this, we'll take a small digression first – recall that a function $f : \mathbb{R}^k \to \mathbb{R}$ is **convex** if

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y)$$

for all $\lambda \in [0, 1]$, and it is **strictly convex** if this inequality is strict for all $\lambda \in (0, 1)$ and $x \neq y$. The picture we should have is basically that convex functions are "bowl-shaped," and all secant lines always stay above the function – something like x^2 is strictly convex, but something with linear pieces would only be convex.

Fact 4

For our purposes, we generally want to be slightly more careful and allow our functions to take on infinite values (specifically $+\infty$), so that for example $-\log$ is convex.

Proposition 5 (Jensen's inequality)

Let f be convex. Then for any random variable X,

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

and if f is strictly convex this inequality is strict unless X is constant.

This proposition allows us to define the following quantity – it turns out that for this next definition (and some of the following ones as well), the discreteness assumption is not necessary, but it'll make the presentation easier. (See Definition 14 below.)

Definition 6

The **Kullback-Leibler** or **KL-divergence** between two discrete distributions P, Q is given by

$$D_{\mathrm{KL}}(P||Q) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

By Jensen's inequality, we always have $D_{\mathrm{KL}}(P||Q) \geq 0$, and this inequality is strict unless $P = Q$. Indeed, the function $f(t) = -\log t$ is convex, and

$$D_{\mathrm{KL}}(P||Q) = \mathbb{E}_P\left[f\left(\frac{q}{p}\right)\right] \geq f\left(\mathbb{E}_P\left[\frac{q}{p}\right]\right) = f(1) = 0,$$

and the leftmost inequality is strict unless $\frac{p(x)}{q(x)}$ is constant, meaning $P = Q$.

Remark 7. *We wrote above that entropy satisfies $H(X) \leq \log |\mathcal{X}|$, and in fact KL-divergence lets us prove this with a "weird clever trick" (these types of "magic arguments" are common in information theory). Indeed, let Q be the uniform distribution on $\mathcal{X}$, meaning that $q(x) = \frac{1}{m}$ for all x (where $|\mathcal{X}| = m$). But then*

$$D_{KL}(P||Q) = \sum_x p(x) \log(mp(x)) = \sum_x p(x) \log m + \sum_x p(x) \log p(x),$$

and this right-hand side is exactly $\log m - H(X)$. Thus $0 \leq \log m - H(X)$ as desired, with strict inequality for a nontrivial distribution.

Definition 8

Let X, Y be discrete random variables. The **mutual information** or **Shannon information** between X and Y is

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

In other words, this quantity is the KL-divergence $D_{\mathrm{KL}}(P_{XY}||P_X \times P_Y)$ between the joint distribution and the independent marginals.

We can relate this quantity to the entropy of our random variables, since we can show via some algebraic manipulations that

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Indeed, expanding out the sum yields

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \log \frac{1}{p(x)} + \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)} \\ &= \sum_{x,y} p(x,y) \log \frac{1}{p(x)} + \sum_{x,y} p(x,y) \log p(x|y). \end{aligned}$$

Summing over y in the first term just yields $\sum_x p(x) \log \frac{1}{p(x)} = H(X)$, and the second term is exactly $-H(X|Y)$ by definition. And swapping the roles of x and y yields the other direction as well.

The way we can think about mutual information is as "uncertainty (in this case, entropy) in X that Y removes," or vice versa. If the two random variables are related or predictive, then the mutual information is quite large, but if

they are completely unrelated, then the mutual information is zero.

We'll need some tools for relating together these different quantities when we have a set of random variables $X_1, \cdots, X_n$ and some other set $Y_1, \cdots, Y_m$ and wanting to disentangle some facts between the two sets of samples. For this, conveniently all of these quantities we've defined (because of the appearance of log in the definitions) have some nice tensorization properties which make terms separate out.

Indeed, the **joint entropy** of a pair of random variables (X, Y) is (using that $p(x, y) = p(x)p(y|x)$)

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

Again marginalizing the first term over y yields $H(X)$, and the second term is the definition of the conditional entropy, so we find $H(X, Y) = H(X) + H(Y|X)$. More generally, we get the following:

Proposition 9 (Chain rule for entropy)

The joint entropy of the random variables $X_1, \cdots, X_n$ is given by

$$H(X_1^n) = \sum_{i=1}^{n} H(X_i|X_1^{i-1})$$

(here X_i^j will always denote the collection $(X_i, X_{i+1}, \cdots, X_j)$).

We'll see this come up a lot in controlling an "ugly object" with a bunch of individual entropy terms; being able to break apart complicated functions into smaller terms gives a lot of computational and representational power. Let's now get a similar chain rule for other settings (information and divergences) as well:

Definition 10

The **conditional mutual information** of X and Y given Z is defined by

$$I(X; Y|Z) = H(X|Z) - H(X|Y, Z) = H(Y|Z) - H(Y|X, Z).$$

Proposition 11 (Chain rule for mutual information)

We have

$$I(X_1^n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1^{i-1}).$$

Proof. We write the information as

$$I(X_1^n; Y) = H(X_1^n) - H(X_1^n|Y)$$

and then use the chain rule for each term individually: this simplifies to

$$\sum_{i=1}^{n} H(X_i|X_1^{i-1}) - H(X_i|Y, X_1^{i-1}),$$

and the ith term here is exactly the definition of $I(X_i; Y|X_1^{i-1})$. □

Finally, for KL-divergences, for distributions P, Q on $\mathcal{X} \times \mathcal{Y}$ (that is, we have probability mass functions $p(x, y)$ and $q(x, y)$), let P_{XY} and P_X be the distributions on $X \times Y$ and X for P, respectively (and similar for Q). We then

get (this time it's a bit weirder because KL-divergence is not symmetric)

$$\begin{aligned} D_{\mathrm{KL}}(P_{XY}||Q_{XY}) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \\ &= D_{\mathrm{KL}}(P_X||Q_X) + \sum_x p(x) D_{\mathrm{KL}}\left(P_{Y|X=x}||Q_{Y|X=x}\right). \end{aligned}$$

We'll now abuse notation and say that if $X \sim P$ and $Y \sim Q$, then $D_{\mathrm{KL}}(X||Y) = D_{\mathrm{KL}}(P||Q)$. Thus we can make the following definition:

Definition 12

The **conditional KL-divergence** between X and Y given Z is

$$D_{\mathrm{KL}}(X||Y|Z) = \sum_z p(z) D_{\mathrm{KL}}(P(\cdot|Z=z)||Q(\cdot|Z=z));$$

what this really is in measure-theoretic terms is $\int D_{\mathrm{KL}}(P(\cdot|z)||Q(\cdot|z))\, dP(z)$.

(We won't need measure theory for this course, so we can really just think of $dP(z)$ as $p(z)dz$ and we won't miss anything.) This allows us to write down our calculation above more nicely:

Proposition 13 (Chain rule for KL-divergence)

We have

$$D_{\mathrm{KL}}(X_1, X_2||Y_1, Y_2) = D_{\mathrm{KL}}(X_1||Y_1) + D_{\mathrm{KL}}(X_2||Y_2|X_1).$$

Extending this to general sequences of variables (again using that $p(x_1, \cdots, x_n) = \prod_{i=1}^n p(x_i|x_1^{i-1})$),

$$D_{\mathrm{KL}}(X_1^n||Y_1^n) = \sum_{i=1}^n D_{\mathrm{KL}}(X_i||Y_i|X_1^{i-1}).$$

Note in particular that **this conditioning is asymmetric** (we condition on the left terms), which makes sense because we're integrating the log factors against the probability mass functions for the left terms. Unfortunately, it's harder to interpret this as "summing up uncertainties" compared to something like the chain rule for entropy or mutual information – it kind of just comes out of the calculations.

We've written everything so far assuming that all random variables are discrete, but we can now write down a general definition (since the divergence measures are what form the basis of everything going forward). Given a set $\mathcal{X}$, we say that a collection of sets $\mathcal{A}_1, \cdots, \mathcal{A}_k$ **partitions** $\mathcal{X}$ if they are disjoint and their union is $\mathcal{X}$.

Definition 14

Given a partition $\mathcal{A} = \{\mathcal{A}_1, \cdots, \mathcal{A}_k\}$ of a not-necessarily discrete $\mathcal{X}$, define

$$D_{\mathrm{KL}}(P||Q|\mathcal{A}) = \sum_{i=1}^{k} P(\mathcal{A}_i) \log \frac{P(\mathcal{A}_i)}{Q(\mathcal{A}_i)}.$$

Alternatively, we can interpret $\mathcal{A}$ as being a **quantizer** $g : \mathcal{X} \to \{1, \cdots, k\}$ of a continuous space into just k different values, and then treating the random variables as only taking on those quantized values so that $D_{\mathrm{KL}}(P||Q|g) = \sum_{i=1}^{k} P(g(X) = i) \log \frac{P(g(X)=i)}{Q(g(X)=i)}$

We then define the **KL-divergence** between any general P and Q as

$$D_{\mathrm{KL}}(P||Q) = \sup_{\text{finite partitions } \mathcal{A}} D_{\mathrm{KL}}(P||Q|\mathcal{A}) = \sup_{\text{finite quantizer } g} D_{\mathrm{KL}}(P||Q|g).$$

We'll see some nice properties of this quantity next time! There is a theorem that if P, Q have a density with respect to some basis measure μ on $\mathcal{X}$, then the KL-divergence is actually just the integral

$$D_{\mathrm{KL}}(P||Q) = \int p(x) \log \frac{p(x)}{q(x)} d\mu(x)$$

(though this is a bit annoying to actually check). But the point is that our definition makes it a lot easier to assume without loss of generality that everything we're working with is finite and discrete, so that things like our chain rules do properly work.

2 September 25, 2025

We'll discuss data processing inequalities today, telling us that further processing can only reduce information. We'll then talk about f-divergences (a generalization of KL-divergences) and take a first look at optimal testing.

To recap, we've introduced the entropy $H(X)$ and conditional entropy $H(X|Y)$, the mutual information $I(X;Y)$, and the KL-divergence D_{KL}, and we established various chain rules to relate things with multiple random variables to those of a single random variable: for example, $I(X_1^n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1^{i-1})$, and $D_{\mathrm{KL}}(P^n||Q^n) = nD_{\mathrm{KL}}(P||Q)$.

Data processing inequalities (DPIs) will come up again and again in this class, and we'll make a first pass at them now. The way to think about them is that we wish to **understand degradation of signals or observations** in settings where we have some kind of Markov structure or Markov chain.

Proposition 15 (Data processing inequality)

Consider a Markov chain in the graphical notation

$$X \to Y \to Z,$$

meaning that we can factor the distribution as $p(x, y, z) = p(x)p(y|x)p(z|y)$. Then $I(X;Z) \le I(X;Y)$, and $I(X;Z) \le I(Y;Z)$.

Basically, further downstream processing of an observation can never increase the amount of information we have about our original signal.

Proof. Expanding the mutual information in different ways, we have

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y) = I(X;Y),$$

since given Y the variables X, Z are completely independent (we may write this as $X \perp\!\!\!\perp Z|Y$) and thus there is no mutual information. On the other hand,

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \geq I(X;Z),$$

so equating these expressions we see that $I(X;Z) \leq I(X;Y)$. A similar proof works for the other claim by swapping the roles of X and Z. □

It turns out that there are similar results for the KL-divergence, but they turn out to be special cases of stuff we'll develop later today. The basic idea is that if $Q(\cdot|x)$ is a Markov kernel $X \to Z$ and we define $Q \circ P$ via

$$Q \circ P = \int Q(\cdot|x)p(x)dx$$

(that is, the marginal distribution over Z when X is distributed as P), then

$$D_{\mathrm{KL}}(Q \circ P_1 || Q \circ P_2) \leq D_{\mathrm{KL}}(P_1||P_2).$$

(so passing variables through Markov kernels is non-expansive for KL). We'll prove this after we complete the following generalization of KL-divergence:

Definition 16

Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $f(1) = 0$, and let P, Q be discrete distributions. The ***f*-divergence** (also called **Ali-Silvey** or **Csiszár** or ***ϕ*-divergence**) is defined by

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

Exactly analogous to Definition 14, quantizers $g : \mathcal{X} \to [m] = \{1, 2, \cdots, m\}$ partition $\mathcal{X}$ into disjoint sets $A_i = g^{-1}(\{i\})$, so we can define (now for **general** P, Q)

$$D_f(P||Q|g) = \sum_i f\left(\frac{P(A_i)}{Q(A_i)}\right) Q(A_i),$$

and $D_f(P||Q)$ is the supremum of $D_f(P||Q|g)$ over all quantizers g.

This quantity is always nonnegative by Jensen's inequality, since

$$D_f(P||Q) = \mathbb{E}_Q[f(\frac{p}{q})] \geq f(\mathbb{E}_Q[\frac{p}{q}]) = f(\sum_x q(x)\frac{p(x)}{q(x)})f(1) = 0$$

and so the same holds under taking limits in general as well. And similar to what we said at the end of last lecture, if P and Q have densities with respect to same base measure μ (this is not losing any generality, since we can let $\mu = P + Q$), then

$$D_f(P||Q) = \int f\left(\frac{p(x)}{q(x)}\right) q(x)d\mu(x) = \int f\left(\frac{dP}{dQ}\right) dQ(x).$$

Example 17

If we let $f(t) = t \log t$ (which is indeed convex), then

$$D_f(P||Q) = \int \left(\frac{p}{q} \log \frac{p}{q}\right) q = \int p \log \frac{p}{q} = D_{\mathrm{KL}}(P||Q).$$

On the other hand, if we use $f(t) = -\log t$ instead, we actually get the KL-divergence the other way around:

$$D_f(P||Q) = D_{\mathrm{KL}}(Q||P).$$

Example 18 (Hellinger distance)

Let $f(t) = \frac{1}{2}(\sqrt{t} - 1)^2$ (which graphically looks pretty similar to $t \log t$). The corresponding f-divergence can be written as

$$D_f(P||Q) = \frac{1}{2} \int \left(\sqrt{\frac{p}{q}} - 1\right)^2 q = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2$$

(so this is like the ℓ^2 distance between the square root of the densities) – we write this as $d^2_{\mathrm{Hel}}(P, Q)$. We can also write this quantity as

$$d^2_{\mathrm{Hel}}(P, Q) = \frac{1}{2} \int (p + q - 2\sqrt{pq}) = 1 - \int \sqrt{pq}.$$

Example 19 (Total variation distance)

Let $f(t) = \frac{1}{2}|t - 1|$. Then

$$D_f(P||Q) = \frac{1}{2} \int |p - q| = \int (p - q)_+ = \int (q - p)_+ = \sup_A |P(A) - Q(A)| = ||P - Q||_{\mathrm{TV}}$$

(we can check some of these equalities as an exercise).

Example 20 (χ^2-divergence)

Let $f(t) = (t - 1)^2$. Then

$$D_f(P||Q) = \int \left(\frac{p}{q} - 1\right)^2 q = \int \frac{p^2}{q} - 2p + q = -1 + \int \frac{p^2}{q},$$

and this quantity is sometimes denoted $D_{\chi^2}(P||Q)$.

For a bit of context, total variation is intimately connected to optimality in testing. KL-divergence and Hellinger distance play nicely with product distributions, meaning repeated observations. And χ^2-divergence is a bit easier to work with for certain lower bounds because it has some quadratics, but we won't use it very much.

We'll now develop some properties that work for any f-divergence, including data processing inequalities. We won't prove them, but they will be homework problems that are worth working through.

Proposition 21

Let $K : \mathcal{X} \to \mathcal{Z}$ be a Markov kernel, meaning that for each $x \in \mathcal{X}$ we have a probability distribution $K(\cdot, x)$ (with interpretation $K(A, x) = \mathbb{P}(Z \in A | X = x)$. Define the "marginal distributions"

$$K_P = \int K(\cdot, x) p(x), \quad K_Q = \int K(\cdot, x) q(x).$$

Then $D_f(K_P || K_Q) \leq D_f(P||Q)$ for any f-divergence.

The picture to have in mind is that "our distributions always get closer" in some general sense whenever we add some noise K. (The proof is basically a clever careful application of Jensen's inequality, plus the convexity of the "perspective transform" of a function $f_{\text{per}}(u, t) = tf\left(\frac{u}{t}\right)$.)

Corollary 22

Write $g_1 \prec g_2$ if g_1 is a finer quantizer than g_2 (meaning that it just breaks the pieces of g_2 into smaller pieces). Then $D_f(P||Q|g_1) \geq D_f(P||Q|g_2)$.

The idea of the proof here is that the kernel associated to g_2 is a further processing of the kernel associated to g_1 (since knowing the value of g_2 at a point automatically tells us the value of g_1).

Next, we'll connect the different divergences to each other, which will be helpful for demonstrating optimality. The big-picture idea is that it's easier to manipulate divergences like KL or Hellinger than something like total variation, so we want to transform between them with some inequalities.

Proposition 23 (Pinsker)

For any distributions P, Q, we have

$$||P - Q||_{\text{TV}}^2 \leq \frac{1}{2} D_{\text{KL}}(P||Q).$$

Proposition 24 (Le Cam)

For any distributions P, Q, we have

$$d_{\text{Hel}}^2(P, Q) \leq ||P - Q||_{\text{TV}} \leq d_{\text{Hel}}(P, Q)\sqrt{2 - d_{\text{Hel}}^2(P, Q)} \leq 1.$$

From these inequalities, we see that Hellinger and total variation distance give the same topology on probability distributions (since one goes to zero if and only if the other does).

Proof of Proposition 23. We'll first prove this inequality for Bernoulli random variables and then show why that implies the general case. Define the "negative binary entropy"

$$h(p) = p \log p + (1 - p) \log(1 - p);$$

taking some derivatives shows that $h'(p) = 1 + \log p - 1 - \log(1 - p) = \log \frac{p}{1-p}$ and $h''(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$, so for all $p \in (0, 1)$ we have $h''(p) \geq 4$. Therefore the "Taylor expansion" (this is really called the **Bregman divergence** $D_h(q, p)$ associated with the negative entropy functional $h = t \log t$)

$$h(q) - h(p) - h'(p)(q - p) = p \log \frac{p}{q} + (1 - p) \log \frac{(1 - p)}{(1 - q)}$$

is exactly the KL-divergence $D_{\mathrm{KL}}(p||q)$. At the same time, Taylor's theorem also tells us that

$$\begin{aligned} h(q) &= h(p) + h'(p)(q-p) + \frac{1}{2}(p-q)^2 h''(x) \\ &\geq h(p) + h'(p)(q-p) + 2(p-q)^2, \end{aligned}$$

so rearranging yields that $2(p-q)^2 \leq D_{\mathrm{KL}}(p||q)$. But now for **any general distributions** P, Q, if we let $p = P(A)$ and $q = Q(A)$ for an arbitrary set A, we find that

$$2(P(A) - Q(A))^2 \leq D_{\mathrm{KL}}(\mathrm{Ber}(P(A))||\mathrm{Ber}(Q(A))) \leq D_{\mathrm{KL}}(P||Q)$$

by the data processing inequality. So taking supremum over A yields that

$$2 \sup_A |P(A) - Q(A)|^2 \leq D_{\mathrm{KL}}(P||Q),$$

which is exactly what we wished to prove. □

Example 25

Turning to optimality, we'll often be considering testing problems as our basic object. Our "canonical testing setting" is that nature chooses one of the distributions P_0 and P_1, and we observe $X \sim P_i$ (for nature's choice i). Our goal is to discover or detect which P_i we actually had.

Proposition 26 (Le Cam's inequality for binary testing)

Suppose that nature chooses P_0 or P_1 uniformly at random – let $V \in \{0, 1\}$ be the choice. Conditioned on $V = v$, we observe $X \sim P_v$. For any test $\psi : \mathcal{X} \to \{0, 1\}$ (encoding which choice we think was made), define the error

$$\mathbb{P}(\psi(X) \neq V) = \frac{1}{2}(P_0(\psi \neq 0) + P_1(\psi \neq 1)).$$

Then

$$\inf_\psi \left(P_0(\psi \neq 0) + P_1(\psi \neq 1)\right) = 1 - ||P_0 - P_1||_{\mathrm{TV}}.$$

Proof. Let $A = \{x : \psi(x) = 0\}$ be the acceptance region of the test. Then

$$P_0(\psi \neq 0) + P_1(\psi \neq 1) = P_0(A^c) + P_1(A) = 1 - (P_0(A) - P_1(A)).$$

Taking the infimum over all A,

$$\inf_\psi \left(P_0(\psi \neq 0) + P_1(\psi \neq 1)\right) = 1 - \sup_A (P_0(A) - P_1(A)) = 1 - ||P_0 - P_1||_{\mathrm{TV}},$$

as desired. □

Example 27

Suppose we want to test a Gaussian location $P_0 \sim N(0, \sigma^2)$ versus $P_1 \sim N(\delta, \sigma^2)$. To understand how far δ needs to be from zero to test, we can use Pinsker's inequality and bound

$$||P_0 - P_1||_{\text{TV}}^2 \leq \frac{1}{2} D_{\text{KL}}(P_0||P_1) = \frac{\delta^2}{4\sigma^2}$$

(this calculation is a nice property of Gaussians).

So in our canonical testing setting,

$$\begin{aligned} \inf_{\psi} 2\mathbb{P}(\psi \neq V) &= 1 - ||P_0 - P_1||_{\text{TV}} \\ &\geq 1 - \sqrt{\frac{1}{2} D_{\text{KL}}(P_0||P_1)} \\ &= 1 - \frac{|\delta|}{2\sigma}. \end{aligned}$$

So the "threshold" is that if $|\delta| \leq \sigma$, the probability that we make an error in the hypothesis test is at least $\frac{1}{4}$. So Pinsker's inequality is useful in these types of settings, and we'll often prove these "constant-probability-of-error-bounds" in this class.

3 September 30, 2025

We'll continue our discussion of "converse results" or lower bounds today, with examples of Le Cam's method and Fano's inequality. That'll complete our blitz through information theory, and we'll then move on to concentration inequalities.

Recall that Le Cam's inequality is a result bounding the probability of making a mistake when choosing between two distributions P_0, P_1 in a test. We worked out an example with Gaussians, and a big part of this was leveraging inequalities like Pinsker to get nicer expressions for total variation distance.

Example 28

Our next example will come up further in the rest of this course. Suppose P_0, P_1 are Bernoulli random variables, where P_0 is 1 with probability $\frac{1-\delta}{2}$ and 0 otherwise, while P_1 is 1 with probability $\frac{1+\delta}{2}$ and 0 otherwise, with $\delta > 0$ to be chosen later.

Our goal is to test between these: one way we can do so is by taking samples X_1^n iid from P_V (where V is the uniformly random choice given to us by nature). We then want to know the probability of finding a mistake; we know that for any test ψ

$$\inf_{\psi} \mathbb{P}(\psi(X_1^n) \neq V) = \frac{1}{2} - \frac{1}{2}||P_0^n - P_1^n||_{\text{TV}},$$

but unfortunately we have to sum over binary sequences and do some very messy calculation to evaluate $||P_0^n - P_1^n||_{\text{TV}}$. So instead we use the chain rule and write

$$||P_0^n - P_1^n||_{\text{TV}}^2 \leq \frac{1}{2} D_{\text{KL}}(P_0^n||P_1^n),$$

and this simplifies to $\frac{n}{2} D_{\text{KL}}(P_0||P_1)$ because we have a product distribution. (This inequality is exactly the reason why results like Pinsker's inequality matter!) So all we need to do know is work out the KL-divergence between two

Bernoulli random variables:

$$D_{\mathrm{KL}}(P_0||P_1) = \frac{1+\delta}{2}\log\frac{1+\delta}{1-\delta} + \frac{1-\delta}{2}\log\frac{1-\delta}{1+\delta} = \delta\log\frac{1+\delta}{1-\delta},$$

and for δ small this Taylor expands out to $\delta(\delta + O(\delta^2)) + \delta + O(\delta^2) = 2\delta^2 + O(\delta^3)$. The particular numbers don't matter so much, but it's definitely true that this is at most $4\delta^2$ for $|\delta| \le \frac{1}{2}$, and substituting back yields

$$\mathbb{P}(\text{Error}) \ge \frac{1}{2} - \frac{1}{2}\sqrt{\frac{n}{2}\cdot 4\delta^2} = \frac{1}{2} - \frac{1}{2}\sqrt{2n\delta^2}.$$

So taking $\delta = \frac{1}{\sqrt{8n}}$ guarantees a constant-size probability of error. So **at a separation in the means of $\delta \asymp \frac{1}{\sqrt{n}}$, testing always has a constant probability of error**.

The following is one of the main tools in information theory and statistics for proving lower bounds. It's particularly useful when we have $k > 2$ hypotheses that we wish to distinguish.

Proposition 29 (Fano's inequality)
Suppose we have a Markov chain $X \to Y \to \hat{X}$ (where $\hat{X}$ is our "guess" for X), where $X, \hat{X} \in \mathcal{X}$ for $|\mathcal{X}| = m$ finite. Let $h_2(p) = -p\log p - (1-p)\log(1-p)$ denote the binary entropy of a Ber(p) random variable. Then

$$h_2(\mathbb{P}(\hat{X} \ne X)) + \mathbb{P}(\hat{X} \ne X)\log(|\mathcal{X}| - 1) \ge H(X|\hat{X}).$$

Here's a "cleaner way" to write that fact (which is also often called Fano's inequality):

Corollary 30
With the notation above, suppose X is uniform from $\mathcal{X}$. Then

$$\mathbb{P}(\hat{X} \ne X) \ge 1 - \frac{I(X;Y) + \log 2}{\log|\mathcal{X}|}.$$

Proof. Let P_{err} denote the probability of a mistake. Proposition 29 above says that

$$\begin{aligned} P_{\text{err}}\log(|\mathcal{X}| - 1) &\ge H(X|\hat{X}) - h_2(P_{\text{err}}) \\ &\ge H(X) + (H(X|\hat{X}) - H(X)) - h_2(P_{\text{err}}) \\ &= \log|\mathcal{X}| - I(X;\hat{X}) - h_2(P_{\text{err}}) \\ &= \log|\mathcal{X}| - I(X;\hat{X}) - \log 2 \\ &= \log|\mathcal{X}| - I(X;Y) - \log 2, \end{aligned}$$

where the last step uses the data processing inequality. Dividing both sides by $\log(|\mathcal{X}| - 1)$ yields the result. □

The way to interpret this (using log base 2 instead of e here) is that if we think of $\log(|\mathcal{X}|)$ as "the number of bits needed to represent X," and $I(X;Y)$ is "the number of bits Y contains about X," then we can think of this as saying

$$\mathbb{P}(\text{error}) = 1 - \frac{1 + \#\text{ bits } Y \text{ carries about } X}{\#\text{ bits to describe } X}.$$

Proof of Proposition 29. We expand the entropy H in two ways. Let E be a random variable which is 1 if $X \ne \hat{X}$ and

0 otherwise. On the one hand, we have by the chain rule that

$$\begin{aligned} H(X, E|\hat{X}) &= H(X|E, \hat{X}) + H(E|\hat{X}) \\ &= \mathbb{P}(E=1)H(X|E=1, \hat{X}) + \mathbb{P}(E=0)H(X|E=0, \hat{X}) + H(E|\hat{X}). \end{aligned}$$

Now if we didn't make a mistake and we know $\hat{X}$, then we also know $X<$ so the second term here is just zero and this simplifies to $\mathbb{P}(E=1)H(X|E=1, \hat{X}) + H(E|\hat{X})$. On the other hand, we also have

$$\begin{aligned} H(X, E|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) \\ &= H(X|\hat{X}), \end{aligned}$$

since knowing $X, \hat{X}$ tells us whether or not we have an error. Therefore

$$\begin{aligned} H(X|\hat{X}) &= \mathbb{P}(E=1)H(X|E=1, \hat{X}) + H(E|\hat{X}) \\ &\leq \mathbb{P}(X \neq \hat{X}) \log(|\mathcal{X}| - 1) + H(E|\hat{X}) \\ &\leq \mathbb{P}(X \neq \hat{X}) \log(|\mathcal{X}| - 1) + H(E), \end{aligned}$$

since conditioning always reduces entropy. □

Example 31

Suppose we wanted to play the 20 questions game, meaning that we have an "asker" and an "answerer," the answerer choose $X \in \mathcal{X}$ from an m-element set, and the asker asks yes/no questions to identify it.

For a sufficient strategy, the asker can always "halve space" out of the possible choices $\mathcal{X}$ remaining, so it's always doable within $\lceil \log_2 m \rceil$ questions. And for the converse result (what's the best strategy for the asker), notice that if X is chosen uniformly at random, Fano's inequality states that (letting Y_i denote the responses to our questions)

$$\begin{aligned} \mathbb{P}(\text{error}) &\geq 1 - \frac{I(X; Y_1, \cdots, Y_n) + \log 2}{\log m} \\ &\geq 1 - \frac{\sum_{l=1}^{n} H(Y_i) + \log 2}{\log m} \\ &= 1 - \frac{n+1}{\log_2 m}. \end{aligned}$$

So we'll always need at least $\log_2 m - 1$ questions to guarantee no error.

These ideas and techniques will come up a lot as we jump into the rest of the course, so it's good to keep them in mind. We'll now transition to concentration inequalities – the big-picture idea is that we want to know when a random variable X "behaves like its population;" that is, when is a sample X_1^n "approximately the same" as its population.

Proposition 32 (Markov's inequality (first moment method))

Let X be a nonnegative random variable. Then for any $t > 0$,

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

Proof. We have

$$\mathbb{P}(X \geq t) = \mathbb{E}[1\{X \geq t\}] \leq \mathbb{E}\left[\frac{X}{t} 1\{X \geq t\}\right],$$

and now since X is nonnegative we can remove the conditioning and bound this from above by $\mathbb{E}\left[\frac{X}{t}\right]$, as desired. □

All other concentration inequalities follow from this in some way, and there's also a randomized version of this called "Markov's equality."

Corollary 33 (Chebyshev's inequality)

For any random variable X,

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

(Indeed, we apply Markov's inequality to the random variable $Z = (X - \mathbb{E}[X])^2$.)

As we increase the number of moments we have, we can get sharper and sharper inequalities. In particular, we have the following:

Corollary 34 (Chernoff bound technique)

Suppose X has exponential moments, meaning that the moment generating function $\phi_x(\lambda) = \mathbb{E}[e^{\lambda X}]$ is finite for some $\lambda \in \mathbb{R}$. Then

$$\mathbb{P}(X \geq t) \leq \inf_{\lambda \geq 0} \phi_x(\lambda) e^{-\lambda t} = \mathbb{E}[e^{\lambda X}] e^{-\lambda t},$$

and then we can optimize this over λ.

Proof. We have, for any $\lambda \geq 0$, that $\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t})$, so then we can use Markov on the random variable $e^{\lambda X}$. □

Example 35

Let X be a Gaussian distributed as $N(0, \sigma^2)$. Then $\mathbb{E}[e^{\lambda X}] = \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$. (Write the expectation as an integral $\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp(\lambda x - \frac{1}{2\sigma^2} x^2) dx$, and then complete the square to get $\frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(-\frac{1}{2\sigma^2}(x - \lambda\sigma^2)^2\right) \exp\left(\frac{\lambda^2\sigma^2}{2}\right) dx$. Then everything else except the last exponential factor is the integrated density of a shifted Gaussian, hence 1.) Thus we get the tail bound

$$\mathbb{P}(X \geq t) \leq \inf_{\lambda \geq 0} \exp\left(\frac{\lambda^2 \sigma^2}{2} - \lambda t\right),$$

and optimizing over the quadratic yields the exponential tail bounds $\boxed{\mathbb{P}(X \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)}$ (by setting $\lambda = \frac{t}{\sigma^2}$).

What we'll want to do now is construct convenient collections of random variables which play nice with Chernoff bounds, which motivates the following definition:

Definition 36

A random variable X is **σ^2-sub-Gaussian** if

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

for all $\lambda \in \mathbb{R}$.

The point is that these random variables are analytically convenient, and in particular we automatically have both of the tail bounds $\mathbb{P}(X - \mathbb{E}[X] \geq t), \mathbb{P}(X - \mathbb{E}[X] \leq -t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$ by plugging into our previous calculation. So deviations far above or below the mean are very unlikely for such random variables, and conveniently there are quite a few families which are sub-Gaussian. Obviously Gaussians are, but here are some others:

Example 37

Random signs (also called Rademacher variables) – that is, X equally likely to be $+1$ or -1 – are sub-Gaussian. Indeed,

$$\begin{aligned}\mathbb{E}[e^{\lambda X}] &= \frac{1}{2}(e^{\lambda} + e^{-\lambda}) \\ &= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!}\end{aligned}$$

since the odd terms in the Taylor expansion cancel. Now this last expression is bounded from above by $\sum_{k=0}^{\infty} \left(\frac{\lambda^2}{2}\right)^k \frac{1}{k!}$ (since $(2k)! \geq 2^k k!$), which is exactly $\exp\left(\frac{\lambda^2}{2}\right)$. Thus random signs are 1-sub-Gaussian.

Example 38

More generally, any bounded random variable is sub-Gaussian. This is sometimes known as Hoeffding's lemma: if $X \in [a, b]$, then

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

The standard "book proof" of this result exhibits a technique that we'll develop more later, so we won't bother with it here. But the point is that X is always $\frac{(b-a)^2}{4}$-sub-Gaussian.

The way these properties are most useful is when we actually have collections of such variables together, since we get much more powerful guarantees. The thing we should think about is (reminiscent of tensorization identities and the chain rule) is that **sums of independent sub-Gaussian random variables remain sub-Gaussian**:

Proposition 39

Suppose X_i are σ_i^2-sub-Gausssian and independent. Then $\sum_{i=1}^n X_i$ is $\left(\sum_{i=1}^n \sigma_i^2\right)$-sub-Gaussian.

Proof. By independence, the moment generating function of the sum is just

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(\lambda X_i\right)\right],$$

and then applying σ_i^2-sub-Gaussianity on each term yields the result. □

Corollary 40 (Hoeffding bound)

Suppose that X_i are σ_i^2-sub-Gaussian and independent. Then

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq t) \leq \exp\left(-\frac{nt^2}{\frac{2}{n}\sum_{i=1}^n \sigma_i^2}\right)$$

by applying our sub-Gaussian tail bounds to the random variable $\sum_{i=1}^n X_i$ with scalar nt.

Example 41

Suppose $X_i \in [a, b]$ for all i, and $\mathbb{E}[\overline{X}_n] = \mu$. Then

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

where we use that each X_i is $\frac{(b-a)^2}{4}$-sub-Gaussian.

So in particular, suppose we want to make this deviation smaller than some threshold δ. To understand the scale of fluctuations this gives us (or protects against), we have

$$\delta = 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right) \iff \log\frac{2}{\delta} = \frac{2nt^2}{(b-a)^2} \implies t = \frac{b-a}{\sqrt{2n}}\sqrt{\log\frac{2}{\delta}}.$$

That is, with probability at least $1-\delta$, we have

$$|\overline{X}_n - \mu| \leq \sqrt{\frac{(b-a)^2}{2n}\log\frac{2}{\delta}},$$

and again we see this threshold of fluctuations on the order $\frac{1}{\sqrt{n}}$.

The next class of random variables will also be useful and provides a bit more flexibility than sub-Gaussianity (that is, more nuanced control over complicated objects):

Definition 42

A random variable X is (τ^2, b)**-sub-Exponential** if we have the sub-Gaussian tail bound $\mathbb{E}\left[\exp(\lambda(X - \mathbb{E}[X])\right] \leq \exp\left(\frac{\lambda^2\tau^2}{2}\right)$ but **only for** $|\lambda| \leq \frac{1}{b}$.

Of course, any σ^2-sub-Gaussian random variable is $(\sigma^2, 0)$-sub-exponential, but here are some more interesting examples

Example 43

If $\mathbb{E}[X] = 0$ and $\mathbb{E}[e^{\lambda X}]$ is finite near $\lambda = 0$, then X is sub-Exponential with some parameters (ignoring the details here, since it's not so important for our techniques).

Example 44

Suppose $|X| \leq b$ (meaning that again we have a bounded random variable) and $\mathbb{E}[X] = 0$, $\mathrm{Var}(X) = \sigma^2$. We know that $\sigma^2 \leq b^2$ but that inequality can be very far from tight. We know that this is $\left(b^2, 0\right)$-sub-Exponential, but we can get better bounds: we claim that for $|\lambda| \leq \frac{1}{2b}$,

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{3}{5}\lambda^2\sigma^2\right),$$

meaning that we do look like a Gaussian of variance σ^2 in a neighborhood of 0, rather than just using the crude bound of b^2.

To prove that inequality, we'll use some techniques that are good to keep in mind: we write out the Taylor expansion

$$\begin{aligned}\mathbb{E}[\exp(\lambda X)] &= 1 + \lambda\mathbb{E}[X] + \frac{\lambda^2}{2}\mathbb{E}[X^2] + \sum_{k=3}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}[X^k] \\ &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}[X^k].\end{aligned}$$

Now by boundedness, we have $|\mathbb{E}[X^k]| \le \mathbb{E}[X^2 b^{k-2}] \le \sigma^2 b^{k-2}$ (this is sometimes called a Bernstein condition), so that

$$\begin{aligned}\mathbb{E}[\exp(\lambda X)] &\le 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\lambda^2\sigma^2\frac{\lambda^{k-2}b^{k-2}}{k!} \\ &= 1 + \frac{\lambda^2\sigma^2}{2} + \lambda^2\sigma^2\sum_{k=1}^{\infty}\frac{(\lambda b)^k}{(k+2)!}.\end{aligned}$$

Now if $|\lambda| \le \frac{1}{2b}$, then this sum is at most $\frac{1}{10}$, so that

$$\mathbb{E}[\exp(\lambda X)] \le 1 + \frac{\lambda^2\sigma^2}{2} + \frac{1}{10}\lambda^2\sigma^2 = 1 + \frac{3}{5}\lambda^2\sigma^2,$$

and therefore the moment generating function is bounded by $\exp\left(\frac{3}{5}\lambda^2\sigma^2\right)$, as desired. So the point is that we can replace all of the higher-order terms with "variance-type terms," and we've gotten much finer control near 0 by specifying a small neighborhood of control.

4 October 2, 2025

We'll continue our discussion of exponential random variables, looking at some Bernstein-type bounds, and then introduce some martingale methods and concentration results, moving into bounded differences, uniform laws of large numbers, and so on.

Recall that X is sub-Exponential if we have the sub-Gaussian property only in a neighborhood of 0. We gave an example with variances last time, showing that if X is bounded by b and the variance is bounded by σ^2, then X is $\left(\frac{6}{5}\sigma^2, 2b\right)$-sub-Exponential. We'll show another example now with an interesting technique:

Example 45

Squares of σ^2-sub-Gaussians are $(O(1)\sigma^4, O(1)\sigma^2)$-sub-Exponential.

We'll sketch why this is the case – the point is to use a technique called **pseudo-maximization** to bound the moment generating function. The key trick is that for any nonnegative λ,

$$\mathbb{E}\left[e^{\lambda X^2}\right] = \mathbb{E}\left[\exp(\sqrt{2\lambda}ZX)\right], \quad Z \sim N(0,1) \text{ independent of } X.$$

(To explain this, $\mathbb{E}[e^{sZ}] = \exp\left(\frac{s^2}{2}\right)$ for any scalar, and so we pick an appropriate random $s = \sqrt{2\lambda}X$ in this case.) This is somehow "almost maximizing" our quantity if Z and X are large at the same time, and then by sub-Gaussianity of X we can bound this by $\mathbb{E}[\exp(\lambda\sigma^2 Z^2)]$. But then we can compute this moment generating function by hand: if

$Z \sim N(0, \sigma^2)$, then

$$\begin{aligned}\mathbb{E}[e^{\lambda Z^2}] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(-\frac{1}{2\sigma^2} z^2 + \lambda z^2\right) dz \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int \left(-\frac{(1-2\lambda\sigma^2)_+}{2\sigma^2} z^2\right) dz,\end{aligned}$$

and now that's just a Gaussian integral, meaning this simplifies to

$$\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \sqrt{\frac{2\pi\sigma^2}{(1-2\lambda\sigma^2)_+}} = \frac{1}{\sqrt{(1-2\lambda\sigma^2)_+}}.$$

(The positive part in this calculation is saying that if λ is too big, then the expectation is just $+\infty$.) And so plugging everything back in, the moment generating function of our variable X of interest is bounded as

$$\mathbb{E}[e^{\lambda X^2}] \le \exp\left(-\frac{1}{2}\log(1-2\lambda\sigma^2)\right) \text{ for } \lambda \le \frac{1}{2\sigma^2}.$$

This gets us better concentration guarantees than what we can get otherwise:

Proposition 46 (Bernstein-type concentration)
Let X be (σ^2, b)-sub-Exponential. Then

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

So there's two kinds of behavior for these functions: for smaller deviations we get Gaussian-type behavior for the tails, but for larger ones we get exponential-type behavior.

Proof. We use a Chernoff bound. Assume without loss of generality that $\mathbb{E}[X] = 0$ by translation. Then we know that for all $\lambda \ge 0$

$$\mathbb{P}(X \ge t) \le \mathbb{E}[e^{\lambda X}] e^{-\lambda t},$$

and then if $0 \le \lambda \le \frac{1}{b}$ the sub-Exponential property tells us that this is at most $\exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda t\right)$. So we optimize this quadratic over λ except we're only allowed to use a smaller interval: we set $\lambda = \frac{t}{\sigma^2}$ if we're allowed to (meaning $t \le \sigma^2 b$), and otherwise (if t is too large for that) we set it to be $\frac{1}{b}$. That is, if $t \le \sigma^2 b$ then $\mathbb{P}(X \ge t) \le \exp\left(-\frac{t^2}{2\sigma^2}\right)$, and if $t > \sigma^2 b$ then $\mathbb{P}(X \ge t) \le \exp\left(\frac{\sigma^2}{2b^2} - \frac{t}{b}\right) \le \exp\left(-\frac{t}{2b}\right)$. So indeed taking the minimum of the two terms inside the exponential gets us a universal bound. □

We now want to develop tensorization-type inequalities for sums of such variables, since we want to also have a nice calculus under operations of that type.

Proposition 47
Let X_i be independent (σ_i^2, b_i)-sub-Exponential random variables. Let $b_* = \max_{i \le n} b_i$ be the "tightest window" of any of those variables. Then $\sum_{i=1}^n X_i$ is $(\sum_{i=1}^n \sigma_i^2, b^*)$-sub-Exponential.

Proof. We write out the moment generating functions: by independence we have

$$\mathbb{E}\left[\exp(\lambda \sum_{i=1}^n X_i)\right] = \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}],$$

and then we apply the exponential bound to each term as long as $|\lambda|$ is smallest than all $\frac{1}{b_i}$, or in other words smaller than $\frac{1}{b_*}$. $\square$

We can see a cool consequence now which is much better than the naive Hoeffding-type bound:

Proposition 48 (Bernstein inequality)
If $|X_i| \leq b$ and $\mathrm{Var}(X_i) \leq \sigma^2$, then (letting $\overline{X}_n$ denote the average of $X_1, \cdots, X_n$)

$$\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{nt^2}{\sigma^2}, \frac{nt}{b}\right)\right)$$

for some constant $c > 0$.

This is a consequence of the following more general tensorization inequality:

Proposition 49
Let X_i be (σ_i^2, b_i)-sub-Exponential and mean zero. Then

$$\mathbb{P}\left(\overline{X}_n \geq t\right), \mathbb{P}\left(\overline{X}_n \leq t\right) \leq \exp\left(\min\left(\frac{nt^2}{\frac{2}{n}\sum_{l=1}^n \sigma_i^2}, \frac{nt}{2b_*}\right)\right)$$

This follows directly by applying our Bernstein-type concentration to the random variable $n\overline{X}_n = \sum_{i=1}^n X_i$ to $(\sum_{i=1}^n \sigma_i^2, b_*)$. And so Proposition 48 follows by plugging everything in, using that each X_i is $\left(\frac{6}{5}\sigma^2, 2b\right)$-sub-Exponential.

To understand this relative to the Hoeffding-type bounds, we can provide some additional perspective. If we set $t = \frac{1}{c}\max\left(\sqrt{\frac{\sigma^2 \log\frac{2}{\delta}}{n}}, \frac{b\log\frac{2}{\delta}}{n}\right)$ in the Bernstein inequality, we find that

$$\boxed{\left|\overline{X}_n - \mathbb{E}[X]\right| \leq O(1)\left(\sqrt{\frac{\sigma^2\log\frac{2}{\delta}}{n}} + \frac{b\log\frac{2}{\delta}}{n}\right) \text{ with probability at least } 1-\delta}.$$

That is, we have some Gaussian / central-limit type term, together with some other sub-exponential part. On the other hand, if we were to apply the Hoeffding bounds directly for bounded random variables $|X_i| \leq b$, the (pure) sub-Gaussian constant is b^2, and the best concentration inequality we can get is that

$$\mathbb{P}\left(\overline{X}_n - \mathbb{E}[X] \geq t\right) \leq \exp\left(-\frac{nt^2}{2b^2}\right).$$

IF we then set $t = b\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$, then we find that

$$\boxed{\left|\overline{X}_n - \mathbb{E}[X]\right| \leq O(1)\cdot b\sqrt{\frac{\log\frac{1}{\delta}}{n}} \text{ with probability at least } 1-\delta}.$$

Since σ^2 is always bounded by b^2 (and can in fact be much smaller), the first boxed fact generally gets us much better concentration.

We'll now build up a second set of techniques, useful for showing that random variables concentrate. The game is that often we want to control general functions $f(X_1^n)$ that are more complicated than just $\sum_{i=1}^n X_i$; intuitively, if f is insensitive to changes in any one individual variable, then random versions of f should be quite close to their expectation.

Definition 50

A sequence of random variables $M_1, M_2, \cdots$ is a **martingale** adapted to a set of random variables $\{Z_i\}$, if $Z_1, Z_2, \cdots$ are any random variables so that each M_n is some deterministic function of $Z_1, \cdots, Z_n$, and

$$\mathbb{E}[M_n | Z_1^{n-1}] = M_{n-1}.$$

We say that $D_1, D_2, \cdots$ form a sequence of **martingale differences (MGD)** if $M_n = \sum_{i=1}^n D_i$ is a martingale; completely equivalently, we have that D_n is a function of Z_1^n and $\mathbb{E}[D_n | Z_1^{n-1}] = 0$.

Example 51

Let D_i be iid random signs, and let M_n be the random walk $\sum_{i=1}^n D_i$. Then the random walk is adapted to $\{D_n\}$, since clearly $\mathbb{E}[D_i | D_{i-1}] = 0$ by independence and each D_i is a function of itself.

This "sequential mean-zero structure" means that we still get concentration even with lots of random variables.

Definition 52

A sequence of random variables $\{D_i\}$ are **σ_i^2-sub-Gaussian MGDs** if we have the sub-Gaussian condition conditioned on the past; that is,

$$\mathbb{E}[e^{\lambda D_i} | Z_1^{i-1}] \le \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right)$$

Theorem 53 (Azuma-Hoeffding)

Suppose D_i are σ_i^2-sub-Gaussian MGDs. Then $\sum_{i=1}^n D_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian.

Proof. We have

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n D_i\right)\right] = \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} D_i\right)\right] \mathbb{E}\left[e^{\lambda D_n} | Z_1^{n-1}\right],$$

and now the last term is at most $\exp\left(\frac{\lambda^2 \sigma_n^2}{2}\right)$ by definition and then we can iterate backwards to get the result. □

The point is that given any function $f : \mathcal{X}^n \to \mathbb{R}$, we can actually define a martingale associated to it called the **Doob martingale**, where

$$D_i = \mathbb{E}\left[f(X_1^n) | X_1^i\right] - \mathbb{E}\left[f(X_1^n) | X_1^{i-1}\right].$$

We can think of this as "altering the ith coordinate of our function." Clearly D_i is adapted to $\{X_i\}$, and by the tower property of expectations (we take the "least conditioning")

$$\mathbb{E}[D_i | X_1^{i-1}] = \mathbb{E}\left[f(X_1^n) | X_1^{i-1}\right] - \mathbb{E}\left[f(X_1^n) | X_1^{i-1}\right] = 0.$$

Furthermore, we have the telescoping sum

$$\sum_{i=1}^n D_i = \mathbb{E}[f(X_1^n) | X_1^n] - \mathbb{E}[f(X_1^n)] = f(X_1^n) - \mathbb{E}[f(X_1^n)].$$

So we can control deviation of random functions from their mean using martingales, if we can control each D_i (that is, the effect of each coordinate individually on the function).

Definition 54

Say that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies $\{c_i\}$-**bounded differences** if $|f(x) - f(x')| \le c_i$ whenever $x = (x_1^{i-1}, x_i, x_{i+1}^n)$ and $x' = (x_1^{i-1}, x_i', x_{i+1}^n)$ differ only in the ith coordinate.

Proposition 55

If f has this bounded differences property **and the X_i are independent**, then the associated Doob martingale is $\sigma_i^2 = \frac{c_i^2}{4}$-sub-Gaussian.

Proof. We sequentially define the lower and upper bounds

$$L_i = \inf_x \mathbb{E}\left[f(X_1^n)|X_1^{i-1}, X_i = x\right] - \mathbb{E}[f(X_1^n)|X_1^{i-1}],$$

$$U_i = \sup_x \mathbb{E}\left[f(X_1^n)|X_1^{i-1}, X_i = x\right] - \mathbb{E}[f(X_1^n)|X_1^{i-1}].$$

We have $L_i \le D_i \le U_i$, and now we can combine the terms together and even take the worst case over all earlier x_is, yielding

$$U_i - L_i \le \sup_{x_1^{i-1}} \sup_{x,x'} \mathbb{E}\left[f(x_1^{i-1}, x, X_{i+1}^n)\right] - \mathbb{E}\left[f(x_1^{i-1}, x', X_{i+1}^n)\right].$$

So we see that $U_i - L_i \le c_i$, and thus by the Hoeffding lemma for bounded random variables we have the guarantee that

$$\mathbb{E}\left[e^{\lambda D_i}|X_1^{i-1}\right] \le \exp\left(\frac{\lambda^2 c_i^2}{8}\right),$$

which proves the desired claim. □

(Of course the X_is must be independent; otherwise if we made all of them exactly the same random variable we might run into some problems.)

Corollary 56

If f satisfies bounded differences and X_i are independent, then $f(X_1^n)$ is $\frac{1}{4}\sum_{i=1}^n c_i^2$-sub-Gaussian. Thus

$$\mathbb{P}\left(f(X_1^n) \ge \mathbb{E}[f(X_1^n)] + t\right), \mathbb{P}\left(f(X_1^n) \ge \mathbb{E}[f(X_1^n)] - t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Example 57

The fact above actually allows us to get "concentration for infinite-dimensional vectors" in some sense. Suppose we have independent random vectors X_i with $||X_i|| \le b$ for all i (for any norm $||\cdot||$ in any dimension). Consider the function $f(x_1, \cdots, x_n) = ||\overline{x}_n|| = \left|\left|\frac{1}{n}\sum_{i=1}^n x_i\right|\right|$.

Then f satisfies bounded differences, since if $x = (x_1, \cdots, x_n)$ and $x' = (x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)$, then

$$|f(x) - f(x')| = |||\overline{x}_n|| - ||\overline{x}_n'||| \le ||\overline{x}_n - \overline{x}_n'||$$

by the reverse triangle inequality, and this is exactly $\frac{1}{n}||x_i - x_i'|| \le \frac{2b}{n}$. So our corollary tells us that

$$\mathbb{P}\left(||\overline{X}_n|| \ge \mathbb{E}[||\overline{X}_n||] + t\right) \le \exp\left(-\frac{nt^2}{2b^2}\right)$$

for all $t \geq 0$. In particular, the norm of the mean is typically not much larger than the expected norm of the mean. Unfortunately, it can be quite annoying to control $\mathbb{E}[||\overline{X}_n||]$ in some arbitrary space – thus it's nice to specialize to a nice case. If our random vectors are all mean zero and our norm is Euclidean, meaning that $||x||^2 = \langle x, x \rangle$ for some inner product, then

$$\begin{aligned}\mathbb{E}[||\overline{X}_n||] &\leq \mathbb{E}[||\overline{X}_n||^2]^{1/2} \\ &\leq \left(\frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}[||X_i||^2] \right)^{1/2}\end{aligned}$$

by expanding out the norm and using independence of our random variables X_n. Since we know our norms are bounded, this can thus be bounded by $\left(\frac{1}{n^2} \cdot nb^2\right)^{1/2}$, and therefore what we find is that

$$\mathbb{P}\left(\left|\left| \frac{1}{n} \sum_{i=1}^{n} X_i \right|\right| \geq \frac{b}{\sqrt{n}} + t \right) \leq \exp\left(-\frac{nt^2}{2b^2} \right).$$

So in a Euclidean space, with probability at least $1 - \delta$ we have norm of a sample mean at most $\left(1 + \sqrt{\log \frac{1}{\delta}}\right) \frac{b}{\sqrt{n}}$.

5 October 7, 2025

We'll focus today on uniform laws, metric entropy, and M-estimation, defining the relevant objects as a motivation for developing fundamental limits (so that we see there are two faces of the estimation coin).

Last class, we showed some concentration inequalities, in particular using that sub-Exponential random variables have bounds on moment generating functions to get Bernstein-type inequalities (trading between variance and a uniform bound) and then using martingale arguments to reason about bounded-difference functions (if changing one coordinate of our function only changes the function by c_i, then the function is exponentially concentrated around its mean if its arguments are independent). We'll now extend this to larger classes, since we are often excited about more complicated functions and want "uniform versions" of these inequalities that hold simultaneously for many functions at once.

Thus we'll develop the idea of **uniform laws** (of large numbers), sometimes called ULLNs for short. (But we'll really be interested in the finite-sample versions.) We set the notation

$$P_n = \frac{1}{n} \sum_{i=1}^{N} 1_{X_i}$$

for the **empirical distribution** of our sample, where 1_{X_i} is a point mass at X_i. We then use the notation

$$P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \mathbb{E}_{P_n}[f(X)]$$

so that we can think of probability distributions as linear functionals. We also have the "population version" of this: for a probability distribution P we define

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

Example 58

We will be interested in classes of functions $\mathcal{F} \subset \{\text{functions } f : \mathcal{X} \to \mathbb{R}\}$, and we will be interested in the quantity

$$||P_n - P||_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |P_n f - Pf|$$

(that is, how far the sample mean deviates from the population mean).

For example, if our X_i are in $\mathbb{R}^d$, and we consider the set of all bounded linear functionals

$$\mathcal{F} = \left\{ \langle u, x \rangle = u^T x : ||u||_2 \leq 1 \right\}.$$

Then

$$||P_n - P||_{\mathcal{F}} = \sup_{||u||_2 \leq 1} u^T (P_n X - PX),$$

and this is maximized when u points in the same direction and we just get the L^2 distance $||\overline{X}_n - \mathbb{E}[X]||_2$. So this is something we're pretty familiar with, but now we can come up with more sophisticated examples as well: suppose we're considering matrices (for example, covariances) and we let

$$\mathcal{F} = \left\{ x \mapsto \langle u, x \rangle^2 = u^T x x^T u : ||u||_2 \leq 1 \right\}.$$

In this case,

$$||P_n - P||_{\mathcal{F}} = \sup_{||u||_2 \leq 1} u^T (P_n X X^T - PXX^T) u,$$

and now $u^T A u$ is the maximum eigenvalue of the matrix A so that

$$||P_n - P||_{\mathcal{F}} = \lambda_{\max} \left(P_n X X^T - PXX^T \right) = \lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^n \left(X_i X_i^T - \mathbb{E}[XX^T] \right) \right).$$

Very similarly, setting $\mathcal{F} = \left\{ x \mapsto u^T x x^T v : ||u||_2, ||v||_2 \leq 1 \right\}$ now allows us to notice negative eigenvalues as well, so then $||P_n - P||_{\mathcal{F}}$ will be the operator norm of $P_n X X^T - PXX^T$.

There are **two main techniques** we'll see all over the place for this kind of object: the first is **symmetrization**, which is a very general strategy across various areas of analysis that lets us assume that some random variables are symmetric.

Proposition 59

Let X_i be random vectors under any norm (for example we could have $||x|| = \sup_{f \in \mathcal{F}} |f(x)|$ for some class $\mathcal{F}$). Then for any power $p \geq 1$, we have

$$\mathbb{E}\left[\left|\left| \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \right|\right|^p \right] \leq 2^p \mathbb{E}\left[||\sum_{i=1}^n \varepsilon_i X_i||^p \right],$$

where ε_i are Rademacher random variables (iid uniform random signs).

The basic idea of this is that signs are sub-Gaussian, so **conditional on** the sequence $\{X_i\}_{i=1}^n$, the sequence $\sum_{i=1}^n \varepsilon_i X_i$ will be a sub-Gaussian sum and thus it's much easier to control than weird collections of infinite-dimensional vectors.

Proof. Introduce "independent copies" X_i' of the X_is with the same distribution of X_i. Then

$$\mathbb{E}\left[\left|\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right|\right|^p\right] = \mathbb{E}\left[\left|\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i'])\right|\right|^p\right]$$
$$= \mathbb{E}\left[\left|\left|\sum_{i=1}^n (X_i - X_i'))\right|\right|^p\right]$$

last step by Jensen's inequality because the norm is convex (pulling out the blue expectation). But then $X_i - X_i'$ has the same distribution as $\varepsilon_i(X_i - X_i')$, so in particular this is the same as

$$\mathbb{E}\left[\left|\left|\sum_{i=1}^n \varepsilon_i(X_i - X_i'))\right|\right|^p\right] = 2^p\mathbb{E}\left[\left|\left|\frac{1}{2}\sum_{i=1}^n \varepsilon_i(X_i - X_i'))\right|\right|^p\right],$$

and now using Jensen's again bounds this by $2^p\left(\frac{1}{2}\mathbb{E}\left[\left|\left|\sum_{i=1}^n \varepsilon_i X_i\right|\right|^p\right] + \frac{1}{2}\mathbb{E}\left[\left|\left|\sum_{i=1}^n \varepsilon_i X_i'\right|\right|^p\right]\right)$, as desired. □

We won't see too many applications of this, but it will be something in our toolbox that we want to keep in mind. The second big idea is **covering numbers and metric entropies**:

Example 60

Suppose we have some big space Θ, and we're interested in a collection of functions of the form $\frac{1}{n}\sum_{i=1}^n f_\theta(x_i)$ indexed by Θ. If we can argue that Θ is well-approximated by some finite subset, then we'd be in good shape, and so the game will be to take our space Θ and try to cover it with balls. Then if we can show the process is well-behaved within each ball, we only need to worry about the centers of the balls $\theta_1, \theta_2, \cdots$.

Definition 61

Let (Θ, ρ) be a (semi-)metric space. A finite set $\mathcal{N} = \{\theta_i\}_{i=1}^N$ is an ***ε*-cover of Θ** (sometimes also called an ε-net) if the whole space is covered by balls of radius ε centered at the θ_is; that is, for all $\theta \in \Theta$, there is some $1 \le i \le N$ such that $\rho(\theta, \theta_i) \le \varepsilon$. The ***ε*-covering number** is the size of the smallest $\mathcal{N}$ such that $\mathcal{N}$ is an ε-cover; more precisely,

$$N(\Theta, \rho, \varepsilon) = \inf\{|\mathcal{N}| : \mathcal{N} \text{ an } \varepsilon\text{-cover}\}.$$

We also have a very related notion:

Definition 62

Again let (Θ, ρ) be a (semi-)metric space. A ***δ*-packing** $\mathcal{M} = \{\theta_i\}_{i=1}^M$ of Θ is a collection so that for any $i \neq j$, we have $\rho(\theta_i, \theta_j) \ge \delta$. The ***δ*-packing number** is the size of the largest $\mathcal{M}$ such that $\mathcal{M}$ is a δ-packing:

$$M(\Theta, \rho, \varepsilon) = \sup\{|\mathcal{M}| : \mathcal{M} \text{ a } \delta\text{-packing}\}.$$

The picture in mind to have is that if we draw a $\frac{\delta}{2}$-ball around each point in our δ-packing, those balls will be nonoverlapping. So these two notions are indeed quite related:

Lemma 63

We have $M(2\delta) \le N(\delta) \le M(\delta)$.

(The first inequality is easy, and the second one is a bit more annoying but can be proved with some appropriate contrapositive.) The most standard example to keep in mind is the "volume covering" or "volumetric argument:"

Proposition 64

Let $||\cdot||$ be any norm on $\mathbb{R}^d$, and let $\mathbb{B} = \{u \in \mathbb{R}^d : ||u|| \leq 1\}$ be the unit ball. We then claim that

$$\left(\frac{1}{\delta}\right)^d \leq N(\mathbb{B}, ||\cdot||, \delta) \leq \left(1 + \frac{2}{\delta}\right)^d.$$

Proof. Let $\{v_i\}_{i=1}^N$ be a δ-cover of the ball $\mathbb{B}$. This means that the sum of the individual volumes is at least the total volume, so (here $v_i + \delta\mathbb{B}$ means the ball centered at v_i with radius δ, using the notation for the "Minkowski sum")

$$\text{Vol}(\mathbb{B}) \leq \sum_{i=1}^N \text{Vol}(v_i + \delta\mathbb{B}) = N\delta^d \text{Vol}(\mathbb{B}),$$

and therefore we must have $1 \leq N\delta^d$, proving the first inequality. For the other bound, let $\mathcal{M}$ be a **maximal** δ-packing (meaning that we can't put another point in). Then $|\mathcal{M}| \geq N(\mathbb{B}, ||\cdot||, \delta)$ by the lemma above, and we will bound the size of the packing. Letting $\mathcal{M} = \{v_i\}_{i=1}^M$, we know that the sets $\{v_i + \delta\mathbb{B}\}$ must cover the whole ball, or else we could put a new point into our packing and contradict maximality. Furthermore, by construction, the balls $\{v_i + \frac{\delta}{2}\mathbb{B}\}$ are disjoint. So

$$M\left(\frac{\delta}{2}\right)^d \text{Vol}(\mathbb{B}) \leq \text{Vol}\left(\left(1 + \frac{\delta}{2}\right)^d \mathbb{B}\right)$$

because the most our balls $\{v_i + \frac{\delta}{2}\mathbb{B}\}$ can stick out are $\frac{\delta}{2}$ beyond the radius of 1. Therefore the same simplification as before yields

$$M\left(\frac{\delta}{2}\right)^d \leq \left(1 + \frac{\delta}{2}\right)^d,$$

which is exactly the desired bound. □

As an application, we'll see that random covariance matrices concentrate quite strongly:

Example 65 (Matrix concentration / random covariance matrices)

We call a vector $X \in \mathbb{R}^d$ **σ^2-sub-Gaussian** if for all $u \in \mathbb{R}^d$

$$\mathbb{E}\left[\exp(u^T x)\right] \leq \exp\left(\frac{||u||_2^2 \sigma^2}{2}\right)$$

(for example if u is the d-dimensional centered Gaussian with covariance $\sigma^2 I_d$, then it is σ^2-sub-Gaussian, and uniform random d-dimensional signs are 1-sub-Gaussian). Our goal will be to control (or estimate) the deviation

$$\frac{1}{n}\sum_{i=1}^n X_i X_i^T - \mathbb{E}[XX^T]$$

for such vectors.

We claim that if $\mathcal{N} = \{u_i\}_{i=1}^N$ is an ε-cover of the ℓ^2-ball $\mathbb{B}_2^d = \{u : ||u||_2 \leq 1\}$, then

$$(1 - 2\varepsilon)||A||_{\text{op}} \leq \max_{i,j} u_i^T A u_j \leq ||A||_{\text{op}}$$

for any operator A. Clearly $\max_{i,j} u_i^T A u_j \leq \sup_{||u||_2, ||v|_2 \leq 1} uAv = ||A||_{\text{op}}$, so the upper bound is clear. But for the lower

bound, for any vector u, v we have

$$\begin{aligned} u^T A v &= (u - u_i)^T A v + u_i^T A v \\ &= (u - u_i)^T A v + u_i^T A u_j + u_i^T A (v - u_j), \end{aligned}$$

and now if we choose u_i, u_j to be ε-close to u and v (since we're assuming we have a covering) we get

$$u^T A v \le u_i^T A u_j + 2\varepsilon ||A||_{\text{op}}.$$

Now choosing u and v to get arbitrarily close to the operator norm of A on the left-hand side yields that $\max_{i,j} u_i^T A u_j \ge ||A||_{\text{op}}(1 - 2\varepsilon)$, as desired.

This claim lets us get the following result:

Proposition 66

Let X_i be σ^2-sub-Gaussian random vectors in $\mathbb{R}^d$. Then there are constants $c > 0, C < \infty$ such that

$$\mathbb{P}\left(||P_n XX^T - PXX^T||_{\text{op}} \ge t\right) \le \exp\left(-c \min\left(\frac{nt^2}{\sigma^4}, \frac{nt}{\sigma^2}\right) + Cd\right).$$

Rearranging this, if we solve for when the right-hand side is equal to some error bound δ, we get that for

$$t = \max\left(\sigma^2 \sqrt{\frac{1}{c} \cdot \frac{Cd + \log\frac{1}{\delta}}{n}}, \sigma^2 \cdot \frac{1}{c} \cdot \frac{Cd + \log\frac{1}{\delta}}{n}\right),$$

we have $\left|\left|\frac{1}{n}\sum_{i=1}^N X_i X_i^T - \mathbb{E}[XX^T]\right|\right|_{\text{op}} \le O(1)\sigma^2\left(\sqrt{\frac{d+\log\frac{1}{\delta}}{n}} + \frac{d+\log\frac{1}{\delta}}{n}\right)$ with probability at least $1 - \delta$. Thus with extremely high probability, we get fluctuations of order $\sigma^2 \max(\sqrt{\frac{d}{n}}, \frac{d}{n})$, and typically we expect the sample size n to be much larger than d so that the first term is what matters. (We should think of σ^2 as something like the largest eigenvalue of the covariance matrix.)

Proof. Consider a $\frac{1}{4}$-cover of the ℓ^2 ball $\mathbb{B}_2^d$. We know that we can choose such a cover with at most 9^d elements, and by the operator norm bounds we just proved we have

$$\begin{aligned} \frac{1}{2}||P_n XX^T - PXX^T||_{\text{op}} &\le \max_{i,j} (P_n XX^T - PXX^T)u \\ &\le ||P_n XX^T - PXX^T||_{\text{op}}, \end{aligned}$$

which is nice because now we reduce to just analyzing concentration of random sums. So rearranging one part of that inequality yields

$$\mathbb{P}(||(P_n - P)XX^T||_{\text{op}} \ge t) \le \mathbb{P}\left(\max_{i,j} u_i^T (P_n - P) XX^T u_j \ge \frac{t}{2}\right),$$

and by a union bound over i this can be written

$$\mathbb{P}(||(P_n - P)XX^T||_{\text{op}} \ge t) \le 2 \cdot 9^d \max_j \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^N (u_j^T X_i)^2 - \mathbb{E}[(u_j^T X)^2] \ge \frac{t}{2}\right).$$

(**Warning**: In this last step, we actually used the fact that $\frac{1}{2}\lambda_{\max}(A) \le \max_i u_i^T A u_i \le \lambda_{\max}(A)$ and also that $\lambda_{\min}(A) \le \min_i u_i^T A u_i \le 2\lambda_{\min}(A)$ because we have symmetric matrices – so in fact we really didn't need to use both u_i and u_j for this argument.) But now squares of sub-Gaussian are sub-Exponential – specifically $(u_i^T X_i)^2$ is

$O(1)(\sigma^4, \sigma^2)$-sub-Exponential, and so

$$\mathbb{P}(||(P_n - P)XX^T||_{\text{op}} \geq t) \leq 2 \cdot 9^d \exp\left(-c \min\left(\frac{nt^2}{\sigma^4}, \frac{nt}{\sigma^2}\right)\right)$$

as desired. □

So the point is that an infinite-dimensional supremum became a finite maximum, and we have strong concentration for any given element of the cover.

Example 67

We'll finish today with a further application to M-estimation problems. A common problem in machine learning and statistics is that we want to estimate a parameter $\theta(P) = \text{argmin}_\theta\{L_p(\theta) = \mathbb{E}_P[\ell(\theta, Z)]\}$, where $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ is some loss function which is convex in θ.

We can think of examples like **logistic regression**, where the data comes in pairs $z = (x, y) \in \mathbb{R}^d \times \{\pm 1\}$, and the loss is

$$\ell(\theta; x, y) = \log(1 + e^{-yx^T\theta}).$$

Or we can consider something like **robust regression**, where the data now comes in pairs $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}$ and we have loss

$$\ell(\theta; x, y) = \log(1 + e^{\theta^T x - y}) + \log(1 + e^{y - \theta^T x})$$

(so that we want y to be close to $\theta^T x$, and the loss grows on either side – this is like a smooth version of the absolute value).

Definition 68

With the notation above, the **M-estimator** is defined as

$$\hat{\theta}_n = \text{argmin}_\theta \left\{L_{P_n}(\theta) : \mathbb{E}_{p_n}[\ell(\theta, Z)]\right\}.$$

where $P_n = \frac{1}{n}\sum_{i=1}^n Z_i$ is the empirical distribution coming from n iid samples of P.

The idea is that we want to show convergence, and we do so by first showing that L_p will grow around $\theta^* = \text{argmin}L(\theta)$ (e.g. quadratically as $\lambda||\theta - \theta^*||^2$). We then show that gradients of the empirical loss are small near θ^*, so that we can't optimize the empirical loss (which is some perturbation of the population loss) and also be far away from the true minimizer, since the quadratic growth will dominate any linear fluctuations we have. (And if we know how to control random vectors and random matrices – Hessians – it shouldn't be that surprising that we'll be able to do these kinds of bounds.)

6 October 9, 2025

We'll continue our discussion of M-estimation today and then discuss more concentration inequalities (via divergences), particularly applied to learning procedures (PAC-Bayes).

Last time, we saw that for $X \in \mathbb{R}^d$ satisfying $\mathbb{E}[e^{X^T u}] \leq e^{||u||_2^2\sigma^2/2}$, we have with probability at least $1 - e^{-t}$ that

the empirical covariance matrix is close to its expectation:

$$||P_n XX^T - \mathbb{E}[XX^T]||_{\text{op}} \le O(1)\sigma^2 \left[\sqrt{\frac{d+t}{n}} + \frac{d+t}{n}\right].$$

We then set the notation for M-estimation: suppose we have a loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ and want to study the empirical loss $L_{P_n}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta_i, z_i)$. Specifically, we want to give convergence of the **minimizer** $\hat{\theta}_n = \text{argmin}_\theta L_{P_n}(\theta)$ to $\theta^* = \text{argmin}_\theta L_P(\theta)$. And intuitively, we said that this is because our loss function grows quadratically, which will eventually outweigh the linear deviations coming from the gradient (so if $||\theta - \theta^*||$ is too far away, it must be sub-optimal). That particular thing is illustrated by the following technical lemma, which we won't prove:

Lemma 69

Suppose f is a convex function and satisfies the conditions

$$||\nabla f(\theta_0)||_2 \le \gamma, \quad \nabla^2 f(\theta) \ge \lambda I \text{ for } ||\theta - \theta_*|| \le \varepsilon$$

for some $\varepsilon \ge \frac{2\gamma}{\lambda}$. Then $\hat{\theta} = \text{argmin}_\theta f(\theta)$ exists, and $||\hat{\theta} - \theta_0||_2 \le \frac{2\gamma}{\lambda}$.

So the idea to demonstrate convergence is to show that $||\nabla L_{P_n}(\theta^*)||_2$ is small (but we know how to do that because we know how to show sums of random vectors converge), and also argue that $\nabla^2 L_{P_n}(\theta)$ is large enough near θ^* (which we developed the tools for last time). Here are some concrete assumptions we'll make:

(a) We'll use generalized linear model losses (e.g. logistic regression, linear regression), so that

$$\ell(\theta; x, y) = h(\theta^T x, y).$$

(For example $h(t, y) = (y - t)^2$ yields linear regression, and $h(t; y) + \log(1 + e^{-ty})$ yields logistic regression.) Then we have

$$\nabla_\theta \ell(\theta; x, y) = x h'(\theta^T x, y), \quad \nabla^2 \ell(\theta; x, y) = x x^T h''(\theta^T x, y).$$

(b) Also, we'll make some smoothness assumptions: assume that $||\nabla \ell(\theta^*; x, y)||$ is bounded by some M_0 for all x, y. Also assume that the mapping $\theta \mapsto \nabla^2 \ell(\theta; x, y)$ is M_2-Lipschitz continuous – that is, $||\nabla^2 \ell(\theta; x, y) - \nabla^2 \ell(\theta'; x, y)||_{\text{op}} \le M_2 ||\theta - \theta'||_2$. Additionally, assume "upward curvature" at θ^* of the form $\nabla^2 L_P(\theta^*) \ge 2\lambda I$.

(c) Finally, assume that h'' is bounded and X are σ^2-sub-Gaussian random vectors.

These are much stronger assumptions than we need – we really only need moment assumptions for a lot of this, but it would complicate the proof a bit more. And we can also consider robust losses instead so that outliers can't cause problems.

Theorem 70

With the notation and assumptions above, assume that $\frac{d+t}{n} < 1$ and $\sigma^2\sqrt{\frac{d+t}{n}} < 1$. Then we have with probability at least $1 - e^{-t}$ that

$$||\hat{\theta} - \theta^*||_2 \le O(1) \cdot \frac{||\nabla L_{P_n}(\theta^*)||_2}{\lambda} \le O(1) \cdot \frac{M_0}{\lambda\sqrt{n}}\sqrt{1+t}.$$

Proof. We'll do the two steps we mentioned above. First to show that gradients are small, we just use bounded differences: because we've assumed that individual gradient norms are bounded, we have $||\nabla L_{P_n}(\theta^*)|| \le C\frac{M_0}{\sqrt{n}}\sqrt{1+t}$ with probability at least $1 - e^{-t}$ by directly plugging into our bounded differences result. And for the Hessian, we can

say that because X is sub-Gaussian, we have with probability at least $1 - e^{-t}$ that (comparing P_n and P)

$$\nabla^2 L_{P_n}(\theta^*) \geq 2\lambda I - C\sigma^2\sqrt{\frac{d+t}{n}}I.$$

We want this to hold at all θ around θ^*, so the Lipschitz assumption tells us that

$$\nabla^2 L_{P_n}(\theta) \geq \lambda I - C\sigma^2\sqrt{\frac{d+t}{n}}I$$

if $||\theta - \theta^*|| \leq \frac{\lambda}{M_2}$. So we can just apply the technical lemma above and we're done. □

We'll use these kinds of results quite a bit, but for now we're just seeing some convergence results, and for the next part of the class we'll see how to use divergences and information-theoretic inequalities to develop more "stability and concentration" procedures. Somehow, the point is that sample behavior should match population behavior if methods are insensitive to perturbation of underlying data.

Our starting point is a particular representation of the KL-divergence:

Theorem 71 (Donsker-Varadhan representation)
We have

$$D_{\mathrm{KL}}(P||Q) = \sup_{g:\mathbb{E}_Q[e^g]<\infty} \left(\mathbb{E}_P[g] - \log \mathbb{E}_Q[e^g]\right).$$

Furthermore, it suffices to take the supremum over simple functions $g(x) = \sum_{i=1}^m c_i 1\{x \in A_i\}$ (for m finite).

Proof. Without loss of generality, assume that P, Q have densities $p(x), q(x)$. First we show that the left-hand side is at most the right-hand side by choosing an appropriate function g – indeed, if we pick $g(x) = \log \frac{p(x)}{q(x)}$, then

$$\mathbb{E}_P[g] - \log \mathbb{E}_Q[e^g] = \mathbb{E}_P\left[\log \frac{p}{q}\right] - \log \mathbb{E}_Q\left[\frac{p}{q}\right] = D_{\mathrm{KL}}(P||Q) - \log 1 = D_{\mathrm{KL}}(P||Q).$$

For the other direction, we'll use an "exponential tilting" technique where we construct a new distribution by adding in an exponential factor. For any g with $\mathbb{E}_Q[e^g]$ finite, we can define the variable

$$Z_g(x) = \frac{e^{g(x)}}{\mathbb{E}_Q[e^g]}.$$

We have $\mathbb{E}_Q[Z_g(x)] = \frac{\mathbb{E}_Q[e^g]}{\mathbb{E}_Q[e^g]}1$, so $Z(x)q(x)$ is a density "exponentially reweighted in the direction of g." Furthermore we have $\mathbb{E}_P[\log Z] = \mathbb{E}_P[g] - \log(\mathbb{E}_Q[e^g])$ (which is exactly the quantity inside the supremum), so we just want to show that $\mathbb{E}_P[\log Z] \leq D_{\mathrm{KL}}(P||Q)$. But

$$\begin{aligned}\mathbb{E}_P[\log Z] &= \mathbb{E}_P\left[\log \frac{p}{q}\right] + \mathbb{E}_P\left[\log\left(\frac{q}{p}Z\right)\right] \\ &\leq D_{\mathrm{KL}}(P||Q) + \log \mathbb{E}_P\left[\frac{q}{p}Z\right]\end{aligned}$$

by Jensen's inequality, and then this last term is just $\log \mathbb{E}_Q[Z] = \log 1 = 0$, as desired. So taking a supremum over all g yields the other inequality.

Finally, we can optimize g over simple functions because KL-divergence is the supremum over all finite partitions $\mathcal{A}$ of the state space, so we can indeed assume everything is supported on finitely possible values. □

This will turn out to help us get some strong concentration guarantees – we have a moment generating function

in the supremum, and so we'll be able to use this formula to control moment generating functions in terms of KL-divergences.

Example 72

Consider a setting where we have a function class $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$, and we have a "prior" π_0 on functions in $\mathcal{F}$, as well as a "posterior" π on functions in $\mathcal{F}$ **which may depend on some data** $P_n = X_1^n$. Consider drawing f according to π (function chosen after observed data) versus according to π_0 (no-knowledge choice). We want to think about whether we can control the functional $g(f) = P_n f - Pf$.

If we put this function g into Donsker-Varadhan, then (to clarify notation below, **everything is conditional on P_n, which we should think of as fixed**, f is drawn according to π, and g is some fixed function)

$$\int (P_n f - Pf) d\pi(f) = \mathbb{E}_\pi[g(f)]$$
$$\leq D_{\mathrm{KL}}(\pi || \pi_0) + \log \mathbb{E}_{\pi_0}\left[\exp\left(P_n f - Pf\right)\right].$$

But now the latter term depends only on π_0 and is independent of our observed data, so we should be able to get that $P_n f - Pf$ is small with this reasoning.

Theorem 73 (PAC (probably approximately correct) Bayes)

Consider the following standing setting: $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ is some collection of functions, and assume that each function $f \in \mathcal{F}$ is σ^2-sub-Gaussian (so $\mathbb{E}[e^{\lambda(f(X)-Pf)}] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$. Let Π be the collection of all possible distributions π on $\mathcal{F}$. Then with probability at least $1-\delta$ over the sample P_n (that is, with X_i iid drawn from P), we get

$$\int (P_n f - Pf)^2 d\pi(f) = \mathbb{E}_{f\sim\pi}[(P_n f - Pf)^2 | P_n] \leq \frac{8\sigma^2}{3} \cdot \frac{D_{\mathrm{KL}}(\pi||\pi_0) + \log\frac{2}{\delta}}{n}$$

simultaneously for all $\pi \in \Pi$.

That is, no matter what distribution we pick on the function space (we could take all point masses, for example), the squared error between $P_n f$ and Pf is bounded by how far apart π is from π_0 with some logarithm penalty. And the magic of this is that given some prior π_0, we get to pick π however we'd like arbitrarily in terms of our sample data. And if π doesn't depend that much on the underlying sample, we get strong concentration guarantees.

Proof. We've seen already that squares of sub-Gaussians are sub-Exponential, and specifically we proved that for all $\lambda \geq 0$ and any fixed $f \in \mathcal{F}$, we have (expectation taken over P_n here)

$$\mathbb{E}[\exp(\lambda(P_n f - Pf)^2)] \leq \left(1 - \frac{2\lambda\sigma^2}{n}\right)_+^{-1/2}$$

(because $P_n f$ is $\frac{\sigma^2}{n}$-sub-Gaussian). If we average this over $f \in \mathcal{F}$, we thus get that

$$\mathbb{E}\left[\int \exp(\lambda(P_n f - Pf)^2) d\pi_0(f)\right] \leq \left(1 - \frac{2\lambda\sigma^2}{n}\right)_+^{-1/2}.$$

Taking $\lambda = \lambda_n = \frac{3n}{8\sigma^2}$ makes the right-hand side 2. And we can now apply Markov to the quantity inside our expectation: we get that

$$\mathbb{P}\left(\int \exp(\lambda(P_n f - Pf)^2) d\pi_0(f) \geq \frac{2}{\delta}\right) \leq \delta.$$

But now we can apply Donsker-Varadhan: plugging in the specific function $g(f) = \lambda(P_n f - Pf)^2$ yields that for any posterior distribution π,

$$\frac{1}{\lambda}\int g(f)d\pi(f) \le \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\pi||\pi_0) + \log\int \exp\left(g(f)\right)d\pi_0(f)\right],$$

and we just showed that the blue term is at most $\frac{2}{\delta}$ with probability at least $1-\delta$. And this is exactly what we wanted to show if we plug the value of λ back in. □

Corollary 74

By just applying Jensen's inequality, with probability at least $1-\delta$, simultaneously for all posteriors $\pi \in \Pi$ on $\mathcal{F}$ we have

$$\mathbb{E}_{f\sim\pi}\left[|P_n f - Pf|\right] \le \sqrt{\frac{8\sigma^2}{3}\cdot\frac{D_{\mathrm{KL}}(\pi||\pi_0) + \log\frac{2}{\delta}}{n}}.$$

So "if we don't collect too much information from the sample, we get great generalization behavior." Let's do an example of this with loss minimization:

Example 75

Suppose we have classifiers with a zero-one loss. That is, suppose we have a predictive setting where f wants to predict a label $y \in \mathcal{Y}$ from some data $x \in \mathcal{X}$:

$$\ell(f(x), y) = 1\{f(x) \neq y\}.$$

Define the population loss

$$L(f) = \mathbb{E}[\ell(f(X), Y)] = \mathbb{P}\left(f(X) \neq Y\right)$$

and the sample loss

$$L_n(f) = P_n\ell(f(X), Y).$$

We wish to show that these two quantities are close.

So we have a set of predictors $\mathcal{F} \subset \{f : \mathcal{X} \to \mathcal{Y}\}$, and the function class of interest is the set of all loss functions $\mathcal{L} = \{(x, y) \mapsto \ell(f(x), y)\}_{f\in\mathcal{F}}$. Since all of our functions in $\mathcal{L}$ take on value either 0 or 1, this class is $\frac{1}{4}$-sub-Gaussian.

Supposing that we can actually describe our functions (to not worry about axiom-of-choice stuff), assume we can encode our predictors $\mathcal{F}$, meaning that there are complexity measures $c(f) \in \mathbb{N}$ such that $\sum_{f\in\mathcal{F}} 2^{-c(f)} \le 1$.(So $c(f)$ is the number of bits it takes to describe the function f – this is called Kraft's inequality, and the "picture proof" is to write down a binary decision tree of what function we have and add up the $2^{-\text{depths}}$ over the leaves.)

So now to apply PAC-Bayes, we choose a prior on $\mathcal{L}$ and look at all possible posteriors. We let

$$\pi_0(f) = \frac{2^{-c(f)}}{\sum_{f'} 2^{-c(f')}},$$

and for the posteriors we can take all posteriors of point mass form at any $f \in \mathcal{F}$ – more precisely, we are choosing point masses of specific loss functions $\mathcal{L}$. Then

$$D_{\mathrm{KL}}(\pi||\pi_0) = \log\frac{1}{\pi_0(f)} = \log\left(2^{c(f)}\sum_{f'} 2^{-c(f')}\right) \le c(f)\log 2.$$

So therefore with probability at least $1-\delta$, we see that **simultaneously for all classifier functions** $f \in \mathcal{F}$, we have

$$(L_n(f) - L(f))^2 \le \frac{2}{3n} \cdot \left[c(f) \log 2 + \log \frac{2}{\delta} \right].$$

The intuition for this is that we can think of encoding d-dimensional vectors $f(x) = \operatorname{sgn}(x^T\theta)$ for some $\theta \in \mathbb{R}^d$. This will use something like $d \log 32$ bits, and so what this tells us is that $(L_n(f) - L(f))^2 \lesssim \frac{d}{n}$ with high probability. And this technique with priors and posteriors is "the only technique we have" to give nontrivial generalization guarantees for things like large neural networks!

7 October 14, 2025

Today will begin **interactive data analysis** (also called **adaptive data analysis**) – we'll give some definitions of a certain "interaction game" framework and then see how information theoretic tools help us address those ideas. This is an interesting recent development which impacts what we do today.

Last time, we were finding applications of the Donsker–Varadhan variational representation (writing KL-divergence as a supremum over functions), and we thought about cases where we took functions to be of the form $g(f) = \lambda(P_n f - Pf)^2$ for f sub-Gaussian. This allowed us to get bounds over empirical loss simultaneously for all posterior distributions π. We'll see some more applications today in a different framework – the starting point is some standard Fisherian or Neyman–Pearson hypothesis testing.

Example 76

Suppose we choose a null hypothesis H_0 and compute some test statistic $T : \mathcal{X} \to \mathbb{R}$. We then observe some value $t_{\text{obs}} = T(X)$, rejecting the null hypothesis if the p-value $p = \mathbb{P}_{H_0}(T \ge t_{\text{obs}})$ is sufficiently small.

This is a very basic framework, but there are lots of problems it doesn't address in modern statistics framework. For one thing, it doesn't address issues with high-dimensional problems (for example when we have many hypothesis tests) – we won't talk much about that in this course, though. Furthermore, we often have a lot of **dataset reuse** – in machine learning, we publish new algorithms by trying to beat existing algorithms on a given benchmark, and that drives a lot of our decision-making. And even with things like the medical data of UK biobank, we have a few thousand papers being written on the same set. And we also have often have a situation with a "garden of forking paths" where if we make a lot of decisions, there are lots of different tests we could have done.

To put this into a more rigorous setting, we'll consider **statistical queries** where our goals are to estimate certain means of the form $P\phi = \mathbb{E}_p[\phi(X)]$. This could mean the mean of a random variable (with $\phi_1(x) = x$) or the variance (by querying for $\phi_2(x) = x^2$ and then calculating $P\phi_2 - (P\phi_1)^2$).

In classical statistics, we would **pre-specify** the list of potential queries $\Phi = \{\phi\}$, and suppose that these ϕ are σ^2-sub-Gaussian. Then we would know that

$$\mathbb{P}\left(\max_{\phi \in \Phi} |P_n\phi - P\phi| \ge \sqrt{\frac{2\sigma^2}{n} \log \frac{|\Phi|}{\delta}} \right) \le \delta.$$

But we might also ask what happens if we choose our collection Φ after some initial data exploration, and there are many examples where we might want to do that (in particular any example of dataset reuse). So we want to address a situation where Φ can depend on previous answers. Another example where this occurs is gradient descent: if we want to minimize some loss function $L_n(\theta) = P_n\ell(\theta, x)$, then one step of gradient descent is $\theta^{k+1} = \theta^k - \alpha \nabla L_n(\theta^k)$. This is

a statistical query but depends on the past (where we've gone), so it's an issue for applying any of the concentration tools we've developed so far.

Example 77

We'll now describe the interactive setting: suppose we have some (possibly very large) collection of functions $\phi_t : \mathcal{X} \to \mathbb{R}$, indexed by some set $\mathcal{T}$ (so that the $\log|\Phi|$ term we had above is basically way too big for any reasonable consideration). First sample X_1^n iid from P, and then repeat the following for iterations $k = 1, 2, \cdots$:

- Choose some query $T_k \in \mathcal{T}$ that we are interested in and let $\phi = \phi_{T_k}$.
- We have some mechanism which responds with an answer A_k, which we **hope** is approximately $P\phi = \mathbb{E}_P[\phi]$. (We'll be designing algorithms so that the answers are accurate even if T_k can be chosen depending on the past.)

The idea now is that from Donsker–Varadhan, we've seen that if the information between T and X_1^n is small (for example, completely independent of our sample), then we expect the empirical average $P_n\phi_T$ will be approximately $P\phi_T$. Said differently, $P_n\phi_T$ should be able to generalize to the population that we're drawing data from, rather than just the sample we collected.

Theorem 78

Suppose $\{\phi_t\}_{t\in\mathcal{T}}$ are σ^2-sub-Gaussian functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Then for all $\lambda \geq 0$ and any random variable $T \in \mathcal{T}$,

$$\mathbb{E}\left[(P_n\phi_T - P\phi_T)^2\right] \leq \frac{1}{\lambda}\left[I(T; X_1^n) - \frac{1}{2}\log\left(1 - \frac{2\lambda\sigma^2}{n}\right)_+\right],$$

and we have a bias guarantee

$$|\mathbb{E}[P_n\phi_T - P\phi_T]| \leq \sqrt{\frac{2\sigma^2}{n} I(T; X_1^n)}.$$

So if there's very little information, then our sample mean is quite close to being unbiased. And so the point is that we will choose mechanisms that do manage to limit the mutual information.

Proof. Without loss of generality, we'll assume all variables $P\phi_t$ are mean zero. We know by σ^2-sub-Gaussianity that $P_n\phi_t$ is $\frac{\sigma^2}{n}$-sub-Gaussian, so

$$\mathbb{E}\left[\exp\left(\lambda(P_n\phi_t)^2\right)\right] \leq \left(1 - \frac{2\lambda\sigma^2}{n}\right)_+^{-1/2}.$$

Thus in Donsker–Varadhan, if we let our posterior be $\pi = T|X_1^n$ and let π_0 be the prior T when the data is not observed, then we can plug in and get

$$\lambda\mathbb{E}\left[\mathbb{E}_\pi[(P_n\phi_T)^2|P_n]\right] \leq \mathbb{E}\left[D_{\mathrm{KL}}(\pi||\pi_0)\right] + \mathbb{E}\left[\log\mathbb{E}_{\pi_0}[\exp(\lambda(P_n\phi_T)^2)]\right]$$

where the outer expectations are over the data X_1^n. But now by Jensen we can pull the log outside of the expectation to get

$$\begin{aligned}\lambda\mathbb{E}\left[\mathbb{E}_\pi[(P_n\phi_T)^2|P_n]\right] &\leq \mathbb{E}\left[D_{\mathrm{KL}}(\pi||\pi_0)\right] + \log\mathbb{E}_{\pi_0, X_1^n}[\exp(\lambda(P_n\phi_T)^2)] \\ &\leq \mathbb{E}\left[D_{\mathrm{KL}}(\pi||\pi_0)\right] - \frac{1}{2}\log\left(1 - \frac{2\lambda\sigma^2}{n}\right)_+.\end{aligned}$$

So now we just need to deal with the expected KL-divergence, but we claim that if we pick π_0 to be the marginal distribution on T then we actually just have the mutual information:

Lemma 79

If π_0 is the marginal probability on T (over X_1^n) and π is the posterior distribution $T|X_1^n$, then

$$\mathbb{E}\left[D_{\mathrm{KL}}(\pi(\cdot|X_1^n)||\pi_0)\right] = I(T; X_1^n).$$

Proof of lemma. Writing $S = X_1^n$ for simplicity, we have

$$\begin{aligned}\mathbb{E}\left[D_{\mathrm{KL}}(\pi(\cdot|S)||\pi_0)\right] &= \iint \pi(t|s)\log\frac{\pi(t|s)}{\pi_0(t)}P(s)\\ &= \iint \pi(t,s)\frac{\log\pi(t,s)}{\pi_0(t)P(s)}\end{aligned}$$

but this is exactly the definition of the mutual information. □

So we've established the squared-error part of the theorem, and now we do the bias part. Notice that

$$\lambda\mathbb{E}[P_n\phi_T] \le \mathbb{E}[D_{\mathrm{KL}}(\pi||\pi_0)] + \log\mathbb{E}_{X_1^n,\pi_0}\left[e^{\lambda P_n\phi_T}\right],$$

but by sub-Gaussianity the rightmost expectation is at most $\exp\left(\frac{\lambda^2\sigma^2}{2n}\right)$. But then using the lemma again the right-hand side is at most $I(X_1^n; T) + \frac{\lambda^2\sigma^2}{2n}$, and then dividing by λ and optimizing yields the result. □

We can "stick numbers into this" to get useful quantitative bounds:

Corollary 80

We have

$$\mathbb{E}\left[(P_n\phi_T - P\phi_T)^2\right] \le \frac{2e\sigma^2}{n}I(T; X_1^n) + \frac{5\sigma^2}{4n}.$$

So the mutual information gives us some potentially small penalty in addition to the usual variance bound (which we would expect to be $\frac{\sigma^2}{n}$). (To get this, we just plug in $\lambda = \frac{n}{2e\sigma^2}$ in our theorem, and observe that $\frac{1}{2}\log\left(1 - \frac{2\lambda\sigma^2}{n}\right) = \frac{1}{2}\log\left(1 - \frac{1}{2e}\right) \ge -0.102$.)

Let's now see how we can use these results in interactive data analysis, starting with a funny trivial "case of failure:"

Example 81

Suppose we wanted to discover a linear association between gene expressions $X \in \{\pm 1\}^k$ (where 1 means it is expressed and -1 not), and we have a phenotype $Y \in \{\pm 1\}$ of interest. We then want to seek directions $v \in \mathbb{S}^{k-1} = \{u : ||u||_2 = 1\}$ in the unit sphere so that $\mathrm{Cov}(v^T X, Y)$ is large.

In a setting where there is no relationship and nothing to discover, everything is just uniformly random signs. "If we were good scientists," we'd pre-register some finitely many $v_1, \cdots, v_m \in \mathbb{S}^{k-1}$, and we would find that (sub-Gaussian maxima grow as $\log n$)

$$\mathbb{E}\left[\max_{j\le m}(v_j^T P_n XY)^2\right] \le O(1).$$

But now if we allow ourselves just one round of interaction, in which we specify a query (choose one v) that depends on previous answers, things can go very wrong. In particular, suppose we've already observed $P_n X^T v_j Y = e_j^T(P_n XY)$ by querying the basis vectors $v_1 = e_1, \cdots, v_k = e_k$. Then we can just set $v_{k+1} = \frac{P_n XY}{||P_n XY||}$ in the unit sphere, and we'll get that $\mathbb{E}[v_{k+1}^T P_n XY] = \mathbb{E}[||P_n XY||_2^2] = \frac{k}{n}$, which is exponentially worse than if we didn't have any interactions ($\log k$ vs k). So then we can "nefariously discover correlations that don't exist."

So to address this, we'll develop **stable procedures**. The idea is that if we have some statistic that we're computing and small changes to the input don't change it much, then it should be easy to limit information. The two things we thus want are the following:

- Adaptive composition: additional analysis only adds a small additional cost or loss of accuracy.
- Bounds on mutual information between queries and the sample observed $I(T; X_1^n)$ are indeed possible, since that's enough to control $\mathbb{E}[(P_n\phi_T - P\phi_T)^2]$.

Example 82

For example, let T be fittings of hyperparameters of algorithms in a machine learning competition. Then T depends on everyone else's entries on the leaderboard (because we'll choose our algorithm based on what's already doing well), but it won't do so very much and possibly does so randomly. And we're trying to avoid overfitting and saying that we're only using performance of other algorithms, so we're not really using all of the data from the sample X_1^n and thus shouldn't get too much overfitting.

We'll continue to abuse notation and write that if $X \sim P$ and $Y \sim Q$, then $D_{\mathrm{KL}}(X||Y) = D_{\mathrm{KL}}(P||Q)$. The setting for our interactive data analysis is then that we have randomized (analyst / answering) algorithms A which take in a sample from $\mathcal{X}^n$ and returns some random variable $A(X_1^n) \in \mathcal{A}$.

Definition 83

An answering algorithm A is **ε-KL-stable** if for each $i \in [n]$ there exists a randomized algorithm $A_i : \mathcal{X}^{n-1} \to \mathcal{A}$ such that for all inputs $X_1^n \in \mathcal{X}^n$,

$$\frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{KL}}(A(X_1^n)||A_i(X_{\setminus i})) \leq \varepsilon.$$

The intuition here is that these alternative algorithms A_i are basically the original algorithm where we swap in some default value x_i' for the ith data point x_i.

Example 84

Suppose we want to add Gaussian noise and do mean estimation. Suppose $x_i \in [-1, 1]$ and we're interested in analyzing the mean $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$; if our algorithm was just "spit out the sample mean" this would not be stable, since we can change out one coordinate and the KL-divergence between point masses at different points is big. But now if we define $A(x_1^n) = \overline{x}_n + Z$ for $Z \sim N(0, \sigma^2)$, then we will define A_i to be (note that A_i doesn't have to be the algorithm A applied to a smaller set)

$$A_i(X_{\setminus i}) = \overline{x}_n - \frac{1}{n}x_i + Z;$$

that is, we are swapping out x_i with zero in the original algorithm so that A_i in fact does not depend on that coordinate.

We can then calculate the KL-divergence between Gaussians of the same variance

$$D_{\mathrm{KL}}(A||A_i) = \frac{1}{2\sigma^2}\left(\overline{x}_n - (\overline{x}_n - \frac{1}{n}x_i)\right)^2 = \frac{1}{2n^2\sigma^2}x_i^2 \leq \frac{1}{2n^2\sigma^2}.$$

So this is ε-stable for $\varepsilon_{KL} = \frac{1}{2n^2\sigma^2}$. (The idea is that we obscure the information a little bit so that it can't cause too much unexpected dependence, while still having something that is usably informative for algorithms.)

What we'd like to do is apply multiple algorithms in sequence, so we need to understand composition and additivity and chain rules when we try to do all of that. The key idea is that we want to be able to argue that a chain of stable algorithms still limits the KL-divergence (information leaked).

Proposition 85

Let $A : \mathcal{X}^n \to \mathcal{A}_0$ and $A' : \mathcal{X}^n \times \mathcal{A}_0 \to \mathcal{A}$ be two algorithms (where the latter one can be dependently chosen based on our data) which are ε-KL-stable and ε'-KL-stable, and define $A' \circ A(X_1^n) = A'(A(X_1^n), X_1^n)$. Then $((A' \circ A)(X_1^n), A(X_1^n))$ is $(\varepsilon + \varepsilon')$-KL-stable.

Notice that if we compose any number of KL-stable algorithms with stability constants $\varepsilon_1, \cdots, \varepsilon_k$, the point is that even if we have a huge dependence structure, then the entire thing is still $(\varepsilon_1 + \cdots + \varepsilon_k)$-KL-stable. So there's no point where we suddenly "dump all of our information" – we have nice sequential control on everything. We'll prove this next time!

8 October 16, 2025

Last time, we proved a bound on $(P_n\phi_T - P\phi_T)^2$ in terms of the sub-Gaussianity constants within a function class and the mutual information $I(X_1^n; T)$. The idea was then that we can limit the mutual information by only giving noisy answers, so that $A(X_1^n)$ is hopefully approximately $P\phi_T$.

To do this, we then started thinking about KL-stability. Recall that we call a function ε-KL-stable if we have alternative algorithms A_i on all but one coordinates (which are usually "replace x_i with some value x_i^*) so that the average of the KL-divergences of those algorithms with the original one is at most ε. We in particular stated Proposition 85, which says that we can compose together algorithms in a certain way and still have stability ("graceful degradation").

Proof of Proposition 85. Unsurprisingly, compositions mean that we'll be using certain chain rules. Fix some i. Let A_i and A_i' be the "promised" sub-algorithms from the definition of KL-stability. Let $P_{A,A'}$ be the joint distribution on $(A'(A(X_1^n), X_1^n), A(X_1^n))$, and let $Q_{A,A'}$ be the joint distribution on the sub-algorithms $(A_i'(A_i(X_{\setminus i}), X_{\setminus i}), A_i(X_{\setminus i}))$. We now define $P_{A'|a}$ to be the distribution of $A'(a, X_1^n)$ and $Q_{A'|a}$ to be that of $A_i'(a, X_{\setminus i})$ (note that we stick the **same** value into the first argument). By the chain rule we then have

$$D_{\mathrm{KL}}(A, A' \circ A || A_i, A_i' \circ A_i) = D_{\mathrm{KL}}(A, A_i) + D_{\mathrm{KL}}(A' \circ A || A_i' \circ A | A),$$

and now remembering that we integrate against the left variable in KL-divergence, the latter term is exactly averaging over the conditioned value A and so it is $\mathbb{E}_A[D_{\mathrm{KL}}(A'(A, X_1^n) || A_i'(A, X_{\setminus i}))]$. So if we sum over all i, and divide by n, the first term on the left-hand side is at most ε by assumption, and the second term becomes

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{KL}}(A'(A, X_1^n) || A_i'(A, X_1^n))\right]$$

and this is at most ε' by assumption. □

Our next step (of interactive data analaysis) is then to leverage KL-stable algorithms to guarantee bounds on mutual information.

Lemma 86

Suppose X_is are independent and use the same notation as above. Then

$$\begin{aligned} I(A; X_1^n) &\leq \sum_{i=1}^n I(A; X_i | X_{\setminus i}) \\ &= \sum_{i=1}^n \int D_{\mathrm{KL}}(A(x_1^n) || A_i(x_{\setminus i}) dP^n(x_1^n), \end{aligned}$$

where the key is to specifically take $A_i(x_{\setminus i}) = A(x_1^{i-1}, X_i, x_{i+1}^n)$ to be the marginal when we resample the ith coordinate.

Proof. The latter equality holds because we can write the mutual information as an average over the KL-divergence

$$I(X; Y) = \int D_{\mathrm{KL}}(P_{X|Y=y} || P_X) p(y) dy$$

for any X, Y. Thus we just have to prove the inequality. By the chain rule we have

$$\begin{aligned} I(A; X_1^n) &= \sum_{i=1}^n I(A; X_i | X_1^{i-1}) \\ &= \sum_{i=1}^n H(X_i | X_1^{i-1}) - H(X_i | X_1^{i-1}, A) \end{aligned}$$

and we want to bound and rewrite these terms. The first term is $H(X_i) = H(X_i | X_{\setminus i})$ (since all of the X_is are independent anyway), and entropy is decreased if we condition on more variables and thus

$$\begin{aligned} I(A; X_1^n) &\leq \sum_{i=1}^n H(X_i | X_{\setminus i}) - H(X_i | X_{\setminus i}, A) \\ &= \sum_{i=1}^n I(A; X_i | X_{\setminus i}), \end{aligned}$$

as desired. □

Proposition 87

The mutual information between any two random variables X, Y is actually given by

$$\begin{aligned} I(X; Y) &= \inf_Q \int D_{\mathrm{KL}}(P_{X|Y=y} || Q) p(y) dy \\ &= \int D_{\mathrm{KL}}(P_{X|Y=y} || P_X) p(y) dy. \end{aligned}$$

The point now is that we chose a particular marginal for our algorithm A_i above, but in fact we could have chosen any other distribution and gotten the same guarantee.

Proof. Writing out the definition,

$$\begin{aligned} I(X;Y) &= \iint p(x|y)\log\frac{p(x|y)}{p(x)}p(y) \\ &= \iint \left(p(x|y)\log\frac{p(x|y)}{q(x)} + p(x|y)\log\frac{q(x)}{p(x)}\right)p(y) \\ &= \int D_{\mathrm{KL}}(P_{X|Y=y}||Q)p(y) + \iint p(y)p(x|y)\log\frac{q(x)}{p(x)} \\ &= \int D_{\mathrm{KL}}(P_{X|Y=y}||Q)p(y) + \int p(x)\log\frac{q(x)}{p(x)} \\ &= \int D_{\mathrm{KL}}(P_{X|Y=y}||Q)p(y) - D_{\mathrm{KL}}(P_X||Q) \\ &\le \int D_{\mathrm{KL}}(P_{X|Y=y}||Q)p(y) \end{aligned}$$

because KL-divergence is nonnegative. And we have equality if and only if the latter KL term is zero, so $Q = P_X$, as desired. □

So if we have a bunch of KL-stable algorithms, we get some mutual information bound, and here is the important consequence of our calculations:

Proposition 88

Let $A_1, \cdots, A_k$ be ε_i-KL-stable algorithms, and suppose X_1^n are independent random variables. Then

$$\frac{1}{n}I(A_1, \cdots, A_k; X_1^n) \le \sum_{i=1}^k \varepsilon_i.$$

So we can take whatever stable algorithms we want, and we can compose them together and limit the total information.

Proof. Notice that the composition $A_1 \circ A_2 \circ \cdots \circ A_k$ is $\sum_{i=1}^k \varepsilon_i$-KL-stable (by induction), so (here we have the shorthand $A_1^k = (A_1, A_2 \circ A_1, \cdots, A_k \circ \cdots \circ A_1)$)

$$\frac{1}{n}I(A_1^k; X_1^n) \le \int \frac{1}{n}\sum_{i=1}^n D_{\mathrm{KL}}(A_1^k(x_1^n)||A_1^k(x_{\setminus i}))dP(x_1^n)$$

by the fact we just proved, where $A_1^k(x_{\setminus i})$ is the composition of algorithms obtained by marginally resampling over X_i. But then we know that the quantity in the integrand is at most $\sum_{i=1}^k \varepsilon_i$ by KL-stability, since the marginals are the best possible algorithm. □

So adding noise to queries controls stability, and now we just need to put some pieces together to get a nice interactive data analysis procedure.

Example 89

Let's now put together a noise addition scheme in the following situation. We sample X_1^n iid from P, and we have some function $\phi_t : \mathcal{X} \to [-1, 1]$ (think of this as a statistical query). We repeat the following procedure: a data analyst chooses some $T_i \in \mathcal{T}$ to query (possibly dependent on past answers), and we respond with the noisy answer

$$A_i = P_n\phi_{T_i} + Z_i, \quad Z_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

In this setting, T_i is some function of the A_1^{i-1}s, and so we should be thinking about the data processing inequality:

$$\frac{1}{n} I(X_1^n; T_1^k) \le \frac{1}{n} I(X_1^n; A_1^k)$$

(the mutual information is at most the information from the answers that came out). The right-hand is then upper bounded in turn by $\frac{k}{2\sigma^2 n^2}$ from our Gaussian noise calculations from last class, so we get the following result:

Theorem 90

We have a uniform bound over our k queries

$$\mathbb{E}\left[\max_{j \le k} (A_j - P\phi_{T_j})^2\right] \lesssim \frac{k}{\sigma^2 n^2} + \frac{1}{n} + \sigma^2 \log k.$$

In particular, we optimize this by choosing our noise of size $\sigma^2 = \sqrt{\frac{k}{\log k}} \cdot \frac{1}{n}$, so the maximum error from issuing k adaptive queries is

$$\mathbb{E}\left[\max_{j \le k} (A_j - P\phi_{T_j})^2\right] \lesssim \frac{\sqrt{k \log k}}{n}.$$

Notice that a naive strategy with no noise addition can easily get much larger errors: we can instead have

$$\mathbb{E}\left[\max_{j \le k} (P_n\phi_{T_j} - P\phi_{T_j})^2\right] \gtrsim \frac{k}{n}.$$

So we have an improvement over naive strategies for sure. It's also believed that $\frac{\sqrt{k}}{n}$ is mostly optimal, meaning that it's optimal for a family of algorithms which return answers which are both accurate for the samples and for the population. (We don't know of algorithms that are accurate for sample means but not for population means, but the proof seems to require it.)

Remark 91. *The issue with applying this to stochastic gradient descent is that we interact many more times than these naive bounds give us, and it turns out this technique won't do what we want. Privacy analysis does give us sharper bounds in those settings though, since stochastic gradient is picking random subsamples and that actually makes us much more stable.*

Remark 92. *If we think about Bonferroni-type corrections in standard hypothesis testing, those usually completely break down when we have any adaptivity. But there were a series of papers on doing "selective inference" in which people found (through a very intricate way) exactly what we condition on when we do sequential decisions in a linear model, though unfortunately those don't really have practical impact.*

We'll develop these techniques more in our next topic of **differential privacy**. First, let's prove this result using something called the **monitor technique**:

Proof. Think of the data analyst as "trying to break everything" and choose adversarial queries which are most different from the population means. Thus, after we've chosen our k queries, we pick the worst one:

$$T_{k+1} = T_{k^*}, \quad k^* = \operatorname{argmax}_{j \le k} (A_j - P\phi_{T_j})^2.$$

Then T_{k+1} is a function of the answers A_1^k that we get, since everything else is a population quantity depending only on the random index (which we know). We've already shown from last class that

$$\mathbb{E}\left[(P_n\phi_{T_{k+1}} - P\phi_{T_{k+1}})^2\right] \le \frac{2e}{n} I(T_{k+1}; X_1^n) + \frac{5}{4n}.$$

But we know by that the mutual information here is at most $k\varepsilon n$ for $\varepsilon = \frac{1}{2\sigma^2 n^2}$, and so this right-hand side is at most $O(1)\left(\frac{k}{n^2\sigma^2} + \frac{1}{n}\right)$. Therefore

$$\mathbb{E}\left[\max_{j\le k}(A_j - P\phi_{T_j})^2\right] = \mathbb{E}[(A_{k^*} - P\phi_{T_k^*})^2]$$
$$\le 2\mathbb{E}\left[(P_n\phi_{T_{k+1}} - P\phi_{T_{k+1}})^2\right] + 2\mathbb{E}[\max_{j\le k} Z_j^2]$$

where we used that $(a+b)^2 \le 2a^2+2b^2$ in the last line. So we have a bound on the first term here, and the maximum of k squared Gaussians can be bounded by $O(1)\sigma^2 \log k$. This completes the proof. □

Example 93

Suppose we want to ask Professor Duchi an embarrassing yes-or-no question, but then a die is rolled and the answer is a lie if the die rolls 1 or 2. This turns out to be optimal for information control certain classes of questions, and this connects to the idea of privacy.

The basic idea is that if we have released statistics which are stable to changes in the underlying sample, then we can maintain privacy. For example, if we have average counts of SNPs in a dataset (which tells us how many people in the sample have a variation on genes 1, 2, and so on), it turns out this breaches quite a lot of privacy; people started to back out when this kind of data was released.

So suppose we have some sample P_n and get some Z (which we can think of as a mechanism or channel), and any definition of privacy should satisfy the following properties:

1. Graceful degradation (no catastrophic collapse in privacy if a user participates in three different studies),
2. Downstream processing of outputs should not do anything, meaning that $P_n \to Z \to Y$ should not increase privacy risk over $P_n \to Z$,
3. Resilience against side information (for example, no catastrophic loss if a separate credit card database is hacked)

The first two of these points can be dealt with using information-theoretic techniques (data processing, chain rule), while the third is a bit trickier. (2) might seem like it's always true, but it just means we have to be careful with the definition – we can't say that privacy requires "we cannot learn anything from the sample at all," for example.

To set notation for this part of the course, let $\mathcal{P}_n$ denote the set of all empirical distributions on n data points, so that $P_n \in \mathcal{P}_n$ if $P_n = \frac{1}{n}\sum_{i=1}^n 1_{x_i}$ for some x_i.

Definition 94

A **mechanism** M takes in a sample (that is, an empirical distribution in $\mathcal{P}_n$) and outputs some value in $\mathcal{Z}$. We say that a mechanism is **(ε, δ)-differentially private** if for all possible outputs $A \in \mathcal{Z}$,

$$\mathbb{P}(M(P_n) \in A) \le e^\varepsilon \mathbb{P}(M(P_n') \in A) + \delta$$

whenever P_n, P_n' are neighboring (meaning that $||P_n - P_n'||_{\text{TV}} \le \frac{1}{n}$, or equivalently there is one change in data point).

This turns out to satisfy those three properties above, and in fact we can develop examples of this. But first let's define a stronger version of privacy:

Definition 95

Let $M : \mathcal{X} \to \mathcal{Z}$ just take in values from our input space. We say that M is **ε-locally-differentially private** if

$$Q(Z_i \in A | X_i) \le e^\varepsilon Q(Z_i \in A | X_i');$$

that is, individuals privatize their own data. Note that we sometimes allow Z_i here to depend on the past Z_1^{i-1}, so there are interactive versions of this as well.

Example 96

The earliest version of this is **randomized response**, coming from Warner in 1965 when working on survey sampling. Suppose that we have two groups represented by $\{0, 1\}$, and we want to know what proportion of individuals belong to 0 versus 1.

The idea is to give individuals a spinner with some region labeled 0 and some region labeled 1, and they respond with a "yes" if they are in the labeled group and "no" otherwise. Then because we know the area of the two regions, we can calculate the proportion ourselves. Mathematically, each individual has a bit $x \in \{0, 1\}$ and they release $1 - x$ with probability $\frac{1}{1+e^\varepsilon}$ and x with probability $q = \frac{e^\varepsilon}{1+e^\varepsilon}$. We then can't compute what each individual's response is, but we know that

$$\mathbb{E}[Z|x] = (2q - 1)x + (1 - q)$$

and thus if we set $\hat{x} = \frac{Z-(1-q)}{2q-1}$ we have $\mathbb{E}[\hat{x}|x] = x$ and the individual variances are $\mathrm{Var}(\hat{x}) \le \frac{1}{4(2q-1)^2} = \left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)^2$. So from this we can get estimates on population proportions: if we want to estimate $p = \mathbb{P}(X = 1)$, we can set $\hat{p} = \frac{1}{n}\sum_{i=1}^n \hat{x}_i$ and thus find that

$$\mathbb{E}[(\hat{p} - p)^2] \le \frac{1}{n}\left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1}\right)^2 + \frac{1}{4n} \asymp \frac{1}{(\varepsilon^2 \wedge 1)n}$$

(variance of the individual terms plus the standard statistical variance). So we indeed get an unbiased estimate where we can control the variance.

9 October 21, 2025

We'll continue our discussion of differential privacy, developing some mechanisms and mentioning alternate definitions. Then we'll move into "composition of private release."

Recall that a mechanism is a randomized mechanism $M : \mathcal{P}_n \to \mathcal{Z}$ from empirical distributions to some output space, and we call it (ε, δ)-differentially private if the probability of outputting some set A satisfies

$$\mathbb{P}(M(P_n) \in A) \le e^\varepsilon \mathbb{P}(M(P_n') \in A) + \delta$$

for neighboring datasets P_n' (differing in only one example). The idea is then that we cannot really distinguish between the outputs of $M(P_n)$ versus $M(P_n')$ since they are pretty close, which means that any individual's response is somewhat hidden.

Our main task is often to design mechanisms that are private but don't lose much over their non-private counterparts (so we want to compute statistics or some function of our sample in a useful way). A useful building block here is the following:

Definition 97

For a function $f : \mathcal{P}_n \to \mathbb{R}^d$ of our sample, we define its **global sensitivity** with respect to the ℓ_p norm

$$\mathrm{GS}_p(f) = \sup_{||P_n - P_n'||_{\mathrm{TV}} \leq 1/n} ||f(P_n) - f(P_n')||_p .$$

This is essentially a "Lipschitz constant" in mathematical language, and the key idea is that adding noise at the scale of the global sensitivity should be sufficient for privacy (since any given individual will be washed away in that noise).

Example 98

The **Laplace mechanism** is the "standard textbook" privacy example, and it works as follows. Recall that the **Laplace distribution** has density $p(w) = \frac{1}{2}\exp(-|w|)$ on $\mathbb{R}$. We add noise to our function by using

$$M(P_n) = f(P_n) + \frac{\mathrm{GS}_1(f)}{\varepsilon} W$$

for W with iid Laplace distributed coordinates. (That is, we add noise proportional to ℓ_1-sensitivity.)

Note that this mechanism is $(\varepsilon, 0)$-differentially-private, which we can see by writing down a ratio of densities. If q_0, q_1 denote the densities of $M(P_n)$ and $M(P_n')$, then

$$\begin{aligned}\frac{q_0(z)}{q_1(z)} &= \exp\left(-\frac{\varepsilon}{\mathrm{GS}_1(f)}||z - f(P_n)||_1 + \frac{\varepsilon}{\mathrm{GS}}||z - f(P_n')||_1\right) \\ &\leq \exp\left(\frac{\varepsilon}{\mathrm{GS}_1(f)}||f(P_n) - f(P_n')||_1\right)\end{aligned}$$

by the triangle inequality, and this is at most $\exp\left(\frac{\varepsilon}{\mathrm{GS}_1(f)} \cdot \mathrm{GS}_1(f)\right)$ by definition.

The point is that this is sometimes great, but at other points it adds way too much noise to be a practical algorithm. Let's see some examples of that:

Example 99

Suppose we want to estimate a bounded mean which we know lies in the range $[-b, b]$, meaning $f(P_n) = \overline{x}_n$. Then $\mathrm{GS}_f = \frac{2b}{n}$ is the maximum we can change the mean with one coordinate, so our mechanism is

$$M(P_n) = \overline{x}_n + \frac{2b}{n\varepsilon}\mathrm{Lap}(1).$$

We can evaluate the expected squared error of this mechanism: we have

$$\mathbb{E}\left[(M(P_n) - \overline{x}_n)^2\right] = \frac{4b^2}{n^2\varepsilon^2}\mathrm{Var}(\mathrm{Lap}(1)) = \frac{8b^2}{n^2\varepsilon^2}.$$

Generalizing this to higher dimensions, let's say that "bounded" now means $||X_i||_2 \leq b$. The issue now is that the ℓ_1-sensitivity gains a factor of $\sqrt{d}$ (because we can place our points at $\left(\frac{b}{\sqrt{d}}, \cdots, \frac{b}{\sqrt{d}}\right)$) and so $\mathrm{GS}_1(f) = \frac{2b}{n}\sqrt{d}$. In this case our mechanism will be

$$M(P_n) + \overline{x}_n + \frac{2b\sqrt{d}}{n\varepsilon} W, \quad W_j \overset{\text{iid}}{\sim} \mathrm{Lap}(1).$$

Then the squared error, by the same calculation, becomes

$$\mathbb{E}\left[||M(P_n) - \overline{x}_n||_2^2\right] = \frac{4b^2}{\varepsilon^2 n^2} d \sum_{j=1}^{d} \mathrm{Var}(W_j) = \frac{8b^2 d^2}{n^2 \varepsilon^2}$$

This quadratic dependence $\frac{d^2}{n^2}$ is actually typical (and unavoidable) for "pure differential privacy" with $\delta = 0$, and that can feel a bit annoying.

Example 100

In histogram estimation (that is, estimating the probability of a multinomial), the quantity of interest is the number of data points within each bucket: suppose $h(P_n) \in \mathbb{N}^k$ with the coordinates being

$$h_j(P_n) = nP_n(X = j) = \#\{X_i = j\}.$$

So in other words, the X_i are in one of k possible categorical states. This was actually the original privacy motivation – we want to compute counts of things happening, say in a database, but we don't want any individual's data to be leaked. This is actually a great setting for ℓ_1-sensitivity, because $\mathrm{GS}_1(h) = 2$ (changing one example can only make one bar go down and one bar go up by 1). So the natural mechanism now is

$$M(P_n) = h(P_n) + \frac{2}{\varepsilon} W$$

for W_j iid Laplace as usual. Then the maximum error of any given coordinate doesn't scale too badly, since any individual bucket won't have counts changing by much. Specifically we have for all $t \geq 0$ by a union bound that

$$\mathbb{P}\left(||M(P_n) - h(P_n)||_\infty \geq \frac{t}{\varepsilon}\right) \leq 2k \exp\left(-\frac{t}{2}\right),$$

and inverting everything shows that if we set $t = 2 \log \frac{2k}{\delta}$, then

$$||M(P_n) - h(P_n)|| \leq \frac{2}{\varepsilon} \log \frac{2k}{\delta} \text{ with probability at least } 1 - \delta.$$

So we're not losing too much for most buckets, and our empirical probabilities are pretty close if we normalize by n.

We'll now try to turn to some alternative definitions of privacy that will let us do more powerful things (and think about sequential composition as well). The basic idea is that we want to control the likelihood ratio

$$L(z) = \log \frac{q(z|P_n)}{q(z|P_n')|}$$

for q the density of the mechanism $Z = m(P_n)$ – if this is always small then we have privacy. For this, we'll make use of a certain quantity useful in information theory (coming up originally in optimal testing exponents):

Definition 101

The **Rényi-α-divergence** between two distributions P, Q is given by

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \int \left(\frac{p}{q}\right)^\alpha q = \frac{1}{\alpha - 1} \log \int \left(\frac{p}{q}\right)^{\alpha - 1} p.$$

(This is defined for $\alpha \in [0, \infty]$ by taking appropriate limits at $0, 1, \infty$.)

Notice that if we define $f(t) = t^\alpha - 1$, then

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log\left(1 + D_f(P||Q)\right),$$

and so at least for $\alpha \geq 1$ we inherit things from f-divergences like data processing inequalities and so on. We can now enumerate some other properties (which follow from various inequalities and identities, so we won't do the proofs here):

- The function $\alpha \mapsto D_\alpha(P||Q)$ is nondecreasing in α.
- We have the limits

$$D_0(P||Q) = -\log Q(p(X) > 0),$$

$$D_1(P||Q) = D_{\mathrm{KL}}(P||Q),$$

$$D_\infty(P||Q) = \sup_{q(x)>0} \log \frac{p(x)}{q(x)}.$$

Definition 102

A mechanism M is **(α, ε)-Rényi-differentially private** if $D_\alpha(M(P_n)||M(P_n')) \leq \varepsilon$ for all neighboring samples P_n, P_n'.

(This is saying that the moments of the likelihood ratio are small, or in other words the densities are typically close.) And this makes sense – privacy should say that tests can't tell between P_n and P_n', and recall that the likelihood ratio test (Neyman-Pearson) is optimal for this kind of thing.

The reason we care about this is that it often allows for cleaner (simpler) mechanisms and easier composition guarantees, since we can use our information theoretic chain rules.

Example 103

We have the Rényi-divergence between Gaussians

$$D_\alpha(N(\mu_0, \Sigma)||N(\mu_1, \Sigma)) = \frac{\alpha}{2}(\mu_0 - \mu_1)^T \Sigma^{-1}(\mu_0 - \mu_1).$$

Now if we have control over normal distributions, it tells us that adding Gaussian noise is a good idea, and that motivates the following:

Example 104

Gaussian mechanisms are essentially the building block behind all current industrial deployments of differential privacy (things like Android phones developing text prediction, or iPhones giving App Store recommendations). Recalling the definition of GS_2 from Definition 97, our mechanism is to take

$$M(P_n) = f(P_n) + \mathrm{GS}_2(f) \cdot N\left(0, \frac{\alpha}{2\varepsilon} I\right),$$

and this is in fact (α, ε)-Rényi-differentially private.

Indeed, substituting into the definition above yields

$$D_\alpha\left(f(P_n) + N\left(0, \frac{\mathrm{GS}_2^2 \alpha}{2\varepsilon} I\right) || f(P_n') + N\left(0, \frac{\mathrm{GS}_2^2 \alpha}{2\varepsilon} I\right)\right) = \frac{\alpha}{2} \frac{||f(P_n) - f(P_n)'||_2^2}{(\mathrm{GS}_2^2 \cdot \alpha)/(2\varepsilon)} \leq \varepsilon.$$

So again turning back to bounded means in higher dimensions, suppose we have points in $\mathbb{R}^d$ satisfying $||x_i||_2 \le 1$. Then $f(P_n) = \overline{x}_n$ has $\frac{2}{n}$-ℓ_2-sensitivity, so we actually have that

$$M(P_n) = \overline{x}_n + N\left(0, \frac{\alpha}{2\varepsilon}\left(\frac{2}{n}\right)^2 I\right)$$

is (α, ε)-Rényi-differentially private, and we find that the expected squared error is

$$\mathbb{E}\left[||M(P_n) - \overline{x}_n||_2^2\right] = \frac{d\alpha}{2\varepsilon} \cdot \frac{4}{n^2};$$

in particular we only have one power of d instead of two. So we might be curious what the connection is between these two notions, and we start with the following:

Proposition 105

Let P, Q be any distributions with $D_\infty(P||Q), D_\infty(Q||P) \le \varepsilon$ –that is, $\left|\log \frac{p(x)}{q(x)}\right| \le \varepsilon$. (This is the same as differential privacy.) Then

$$D_\alpha(P||Q) \le \min\left(\frac{3}{2}\alpha\varepsilon^2, \varepsilon\right).$$

(This will come up more when we talk about optimal lower bounds.)

Proof. The bound $D_\alpha(P||Q) \le \varepsilon$ is trivial from our definition of the Rényi-α-divergences. Now for the other bound, we write out the following calculation. Assume without loss of generality that $\varepsilon \le 1$ and $\varepsilon \le \frac{2}{3\alpha}$ (otherwise $D_\alpha(P||Q) \le \varepsilon$ is already a stronger bound), and define $f(z) = \frac{p(z)}{q(z)} - 1$. We have $|f(z)| \le e^\varepsilon - 1$ by assumption, and since ε is small this is approximately ε. Then

$$(1 + f(z))^\alpha = 1 + \alpha f(z) + \frac{\alpha(\alpha-1)}{2}(1+t)^{\alpha-2} f(z)^2$$

for some $t \in [e^{-\varepsilon} - 1, e^\varepsilon - 1]$ by Taylor's theorem. Thus

$$\begin{aligned}\int \left(\frac{p}{q}\right)^\alpha q &= \int (1 + f(z))^\alpha q(z) \\ &\le \int \left(1 + \alpha f + \text{const} \cdot \alpha(\alpha-1)(e^\varepsilon - 1)^2\right) q(z).\end{aligned}$$

But since $\int fq = \int (p - q) = 0$, this integral above evaluates to $1 + C\alpha(\alpha-1)(e^\varepsilon - 1)^2$, and so plugging things in we get that

$$D_\alpha(P||Q) \le \frac{1}{\alpha - 1}\log\left(1 + C\alpha(\alpha-1)(e^\varepsilon - 1)^2\right) \le C\alpha(e^\varepsilon - 1)^2 \lesssim \alpha\varepsilon^2$$

and we can track the constant $\frac{3}{2}$ by being more careful. □

So differential privacy implies a certain flavor of Rényi-differential privacy, and we can also go backwards (which allows us to argue that we can compose lots of differentially private data releases together and still maintain privacy).

Proposition 106

If $D_\alpha(P||Q) \le \varepsilon$, then we have for any set A that

$$P(A) \le \min\left(\exp\left(\varepsilon + \frac{1}{\alpha - 1}\log\frac{1}{\delta}\right) Q(A), \delta\right)$$

for all $\delta > 0$.

That is, P-probabilities of sets are either bounded by a constant multiple of Q-probabilities, or the sets are very unlikely to start with.

Proof. By data processing, we have that

$$\varepsilon \geq D_\alpha(P||Q) \geq \frac{1}{\alpha-1}\log\left(\frac{P(A)^\alpha}{Q(A)^\alpha}Q(A)\right)$$

where we've replaced P with the indicator $x \mapsto 1\{x \in A\}$. Now if we set $p = P(A)$ and $q = Q(A)$ for shorthand, we get

$$\begin{aligned}
&\frac{\alpha}{\alpha-1}\log\frac{p}{q} \leq \varepsilon + \frac{1}{\alpha-1}\log\frac{1}{q} \\
\iff &\log\frac{p}{q} \leq \frac{\alpha-1}{\alpha}\varepsilon + \frac{1}{\alpha}\log\frac{1}{q} \\
\iff &p \leq \exp\left(\frac{\alpha-1}{\alpha}\varepsilon\right) q^{(\alpha-1)/\alpha}.
\end{aligned}$$

So now we just break into cases: if $q = Q(A) \leq e^{-\varepsilon}\delta^{\alpha/(\alpha-1)}$, then substituting it in yields $p \leq \delta$. Otherwise if $q > e^{-\varepsilon}\delta^{\alpha/(\alpha-1)}$, then first rearranging the boxed expression and then plugging in our bound yields

$$\begin{aligned}
p &\leq \left(\frac{\alpha-1}{\alpha}\varepsilon + \frac{1}{\alpha}\log\frac{1}{q}\right) q \\
&\leq \exp\left(\frac{\alpha-1}{\alpha}\varepsilon + \frac{1}{\alpha}\varepsilon + \frac{1}{\alpha-1}\log\frac{1}{\delta}\right) q,
\end{aligned}$$

which is what we wanted to show. □

The thought now is that Rényi-divergences are a bit hard to interpret, but (ε, δ) feels more doable. Let's see how we can use that:

Corollary 107

Suppose we have a mechanism which is $(\alpha, \varepsilon_{\text{RDP}})$-Rényi private with $\varepsilon_{\text{RDP}} = \frac{\varepsilon}{2}$ and $\alpha = \frac{2}{\varepsilon}\left(1 + \log\frac{1}{\delta}\right)$. Then M is also (ε, δ)-differentially-private.

Proof. We want to get ε-differential-privacy out of $\varepsilon_{\text{RDP}} + \frac{1}{\alpha-1}\log\frac{1}{\delta}$, and those values above are exactly what make this latter quantity at most $\frac{\varepsilon}{2} + \frac{\varepsilon}{2}$. □

Example 108

Going back to our Gaussian mechanisms, suppose $\text{GS}_2(f)$ is finite. Then we know that $M(P_n) = f(P_n) + \text{GS}_2(f) \cdot N\left(0, \frac{\alpha}{2\varepsilon_{\text{RDP}}}I\right)$ is $(\alpha, \varepsilon_{\text{RDP}})$-Rényi-private. So if we take $\alpha = \frac{2}{\varepsilon}\left(1 + \log\frac{1}{\delta}\right)$ and $\varepsilon_{\text{RDP}} = \frac{\varepsilon}{2}$, then evidently our mechanism

$$M(P_n) = f(P_n) + \text{GS}_2(f) \cdot N\left(0, \frac{1 + \log\frac{1}{\delta}}{2\varepsilon^2}I\right)$$

is now (ε, δ)-differentially-private.

Example 109

Applying this again to our multidimensional sample means, suppose $x_i \in B_2^d$ (meaning $||x_i||_2 \leq 1$) and again we care about $f(P_n) = \overline{x}_n$ so that $\text{GS}_2(f) = \frac{2}{n}$. Then the mechanism

$$M(P_n) = \overline{x}_n + \frac{O(1)}{n} \cdot N\left(0, \frac{1 + \log\frac{1}{\delta}}{\varepsilon^2} I\right)$$

will be (ε, δ)-differentially-private, and the expected squared error is at most $O(1) \cdot \frac{d \log\frac{1}{\delta}}{n^2\varepsilon^2}$. So now we've gotten d instead of d^2 by allowing say a probability $\delta = 10^{-10}$ of accidentally releasing all of the information.

We could directly compute the differential privacy parameters for a Gaussian too, and that actually gets us tighter bounds. But the point is that we're going through this other method instead because it will work better with sequential mechanisms, which we'll see next time!

10 October 23, 2025

We'll finish up our discussion of privacy today, moving more into Rényi privacy and the "privacy game" (the way we set up the idea that multiple sequential data releases can maintain some guarantees, similar to adaptive data analysis). We'll then discuss a mechanism with inverse sensitivity which is optimal in some sense.

Recall that we have two different ways of measuring privacy now: one of them relates the quantities $\mathbb{P}(M(P_n) \in A)$ and $\mathbb{P}(M(P_n') \in A)$, and the other uses the Rényi-α-divergence. We also showed some results relating the two with various parameters, though recall that we needed $\delta = 0$ in our ε-DP to get Rényi-DP, but we need some positive probability of catastrophic error δ to go in reverse. (The reason we do this transformation is that likelihood ratios and Rényi-divergences work better with sequential releases.)

We'll make one more connection now, basically saying that if $P(A) \leq e^\varepsilon Q(A) + \delta$ for all A, then this inequality "nearly holds" with $\delta = 0$ for all A. Define (this isn't actually a divergence despite the notation)

$$D_\infty^\delta(P||Q) = \sup_A \left\{ \log \frac{P(A) - \delta}{Q(A)} : P(A) > \delta \right\}.$$

From the definition, we have that $D_\infty^\delta(P||Q) \leq \varepsilon$ if and only if $P(A) \leq e^\varepsilon Q(A) + \delta$ for all A.

Lemma 110

For $\varepsilon > 0$ and $\delta \in (0, 1)$, the following hold:

- $D_\infty^\delta(P||Q) \leq \varepsilon$ if and only if there exists some probability measure R with $||P - R||_{\text{TV}} \leq \delta$ and $D_\infty(R||Q) \leq \varepsilon$. (That is, we can tweak P on a tiny fraction of its mass and get bounded likelihood ratio.)
- We have both $D_\infty^\delta(P||Q) \leq \varepsilon$ and $D_\infty^\delta(Q||P) \leq \varepsilon$ if and only if there exist P_0, Q_0 such that $||P - P_0||_{\text{TV}} \leq \frac{\delta}{1+e^\varepsilon}$, $||Q - Q_0||_{\text{TV}} \leq \frac{\delta}{1+e^\varepsilon}$, and $\max(D_\infty(P_0||Q_0), D_\infty(Q_0||P_0)) \leq \varepsilon$.

We won't prove this (it's tedious and technical), but the proof can be found in the book. We basically just move the mass around to shrink the likelihood in the problem parts, but the algebraic details are annoying.

We'll now leverage this to get composition guarantees (graceful degradation in information release). For this, we'll need chain rules for Rényi-divergence similar to the ones we've already established.

Theorem 111

Let P, Q be any two distributions on some product set $\mathcal{Z}^n$, and suppose that $P_i(\cdot|Z_1^{i-1})$ (this is the distribution of Z_i given Z_1^{i-1}) and $Q_i(\cdot|Z_1^{i-1})$ satisfy

$$D_\alpha(P_i(\cdot|Z_1^{i-1})||Q_i(\cdot|Z_1^{i-1})) \le \varepsilon_i.$$

Then $D_\alpha(P||Q) \le \sum_{i=1}^n \varepsilon_i$.

Proof. Use the usual chain rule for densities. We have

$$\begin{aligned} D_\alpha(P||Q) &= \frac{1}{\alpha-1}\log\left(\int \prod_{i=1}^n \left(\frac{p_i(z_i|z_1^{i-1})}{q_i(z_i|z_1^{i-1})}\right)\right)^\alpha q_i(z_i|z_1^{i-1})dz_1^n \\ &= \frac{1}{\alpha-1}\log\int_{\mathcal{Z}^{n-1}}\left[\int_{\mathcal{Z}}\left(\frac{p_n(z_n|z_1^{n-1}}{q_n(z_n|z_1^{n-1}}\right)^\alpha q_n(z_n|z_1^{n-1})dz_n\right]\prod_{i=1}^{n-1}\left(\frac{p_i}{q_i}\right)^m q_i dz_1^{n-1}. \end{aligned}$$

But by assumption the inner bracketed integral is at most $e^{\alpha-1}\varepsilon_n$, so we get that

$$D_\alpha(P||Q) = \varepsilon_n + D_\alpha(P_{Z_1^{n-1}}||Q_{Z_1^{n-1}})$$

and induct to get the result. □

So the point is that if we can control Rényi-divergence, then we can keep things bounded, and the question is now how to connect this to adaptive data composition and think about multiple data releases.

Example 112

We'll set up a "game" between nature and an adversary, where the adversary wants to figure out the information from nature. Nature chooses a single bit $b \in \{0,1\}$, and the adversary doesn't observe the bit but gets to pick two samples differing in one observation and a private mechanism. Nature then releases the output of the mechanism applied to sample b.

Slightly more formally, we have a family of channels $\mathcal{Q}$, where each Q maps from empirical distributions $\mathcal{P}_n$ to some output. The adversary does not know the value of $b \in \{0,1\}$, and we repeat the following for $k = 1, 2, \cdots$:

- The adversary picks some channel / mechanism $Q_k \in \mathcal{Q}$ and some datasets $X_k^{(0)}, X_k^{(1)}$ with Hamming distance $d_{\text{Ham}}(X_k^{(0)}, X_k^{(1)}) \le 1$. (These choices can depend on all previous outputs.)
- The adversary then observes a draw $Z_k \sim Q_k(\cdot|X_k^{(b)})$.

The point then is that if we still have bounded likelihood ratio no matter what we do, then we've guaranteed privacy. For this, let $Q^{(0)}$ and $Q^{(1)}$ be the possible resulting joint distributions on Z_1^k depending on what the bit b was

Definition 113

The family $\mathcal{Q}$ is **(α, ε)-Rényi-private under k-fold adaptive composition** if

$$\max_b D_\alpha(Q^{(b)}||Q^{(1-b)}) \le \varepsilon,$$

and similarly it is (ε, δ)-differentially private if $\max_{b\in\{0,1\}} D_\infty^\delta(Q^{(b)}||Q^{(1-b)}) \le \varepsilon$.

We've conveniently built up our tools so that if each of our channels Q_i is (α, ε_i)-Rényi-private, then the composition is $(\alpha, \sum_{i=1}^k \varepsilon_i)$-Rényi-private.

Corollary 114
Suppose each channel is $(\varepsilon, 0)$-differentially private. Then the k-fold composition is $(k\varepsilon, 0)$-differentially-private and also $\left(\frac{3k}{2}\varepsilon^2 + \sqrt{6k \log \frac{1}{\delta}}\varepsilon, \delta\right)$-differentially-private.

Proof. The first claim is automatic by taking $\alpha = +\infty$ in our observation. For the second claim, recall that ε-differential-privacy gets us squared ε when we move to Rényi-privacy. Each channel is $\frac{3}{2}\alpha\varepsilon^2$-Rényi-differentially-private for any α, so

$$D_\alpha(Q^{(b)}||Q^{(1-b)}) \leq \frac{3k\alpha}{2}\varepsilon^2 \quad \text{for all } \alpha \geq 0,$$

which is the same as being $\left(\frac{3k\alpha}{2}\varepsilon^2 + \frac{1}{\alpha-1}\log\frac{1}{\delta}, \delta\right)$-differentially-private for any $\alpha > 1, \delta > 0$; now just take $\alpha = 1 + \frac{1}{\varepsilon}\sqrt{\frac{1}{k}\log\frac{1}{\delta}}$ to get the result. □

The point is that if $\varepsilon \ll 1$, then the latter degradation is much stronger than the naive $k\varepsilon$ that we might expect – it grows as $\sqrt{k}$ rather than k. In industry, there have been a number of papers trying to get this as small as possible, and various numerical calculations give exact values at various ranges of k, ε, δ.

Corollary 115
The k-fold composition of (ε, δ)-differentially-private channels is $(k\varepsilon, k\delta)$-differentially-private and also $\left(\frac{3k}{2}\varepsilon^2 + \sqrt{6k\log\frac{1}{\delta_0}}\varepsilon, \delta_0 + \frac{k\delta}{1+e^\varepsilon}\right)$ -differentially-private for all $\delta_0 > 0$.

Proof. This time we have a sequence of divergence measures we have to control more carefully. For each i, we have by assumption that

$$D_\infty^\delta(Q_i(\cdot|x^{(b)})||Q_i(\cdot|x^{(1-b)})) \leq \varepsilon,$$

so by our technical lemma there are $Q_i^{(0)}$ and $Q_i^{(1)}$ with $D_\infty(Q_i^{(b)}||Q_i^{(1-b)}) \leq \varepsilon$ and $||Q_i^{(b)} - Q_i(\cdot|x^{(b)})||_{\text{TV}} \leq \frac{\delta}{1+e^\varepsilon}$. If we now apply the previous result on these purely-bounded likelihood ratio channels, we get

$$D_\alpha(Q_1^{(b)} \circ \cdots \circ Q_k^{(b)}||Q_1^{(1-b)} \circ \cdots \circ Q_k^{(1-b)}) \leq \min\left(\frac{3k}{2}\alpha\varepsilon^2, k\varepsilon\right)$$

where the two arguments in the α-divergence are the distributions of the releases Z_1^k under $Q_i(b)$ and $Q_i^{(1-b)}$, respectively. But now by the triangle inequality for total variation distance over the coordinates, we get

$$||Q_1^{(b)} \circ \cdots \circ Q_k^{(b)} - Q^{(b)}||_{\text{TV}} \leq \frac{k\delta}{1+e^\varepsilon},$$

and therefore $Q_1^{(b)} \circ \cdots \circ Q_k^{(b)}$ is $\left(\frac{3k}{2}\alpha\varepsilon^2 + \frac{1}{\alpha-1}\log\frac{1}{\delta_0}, \delta_0\right)$=differentially-private for all $\delta_0 > 0$, and then we add in the total variation term to get the same for the actual Qs, as desired. □

In practice, b would be like "was the last person participating in the study me or you, given that the adversary knows everything about us?", but this more powerful technique comes out of security (we pick a threat model and show that we are safe under that model).

Remark 116. *We should think of δ as the probability of a "catastrophic failure" situation. The census also implements privacy because it has to protect information, and the census thinks of δ as the probability of encryption schemes being*

cracked, or social engineering getting into the system. In a sample of size n, we can maybe typically think of setting $\delta = \frac{1}{n^3}$.

Example 117

For one more alternative perspective, we can give a different flavor of "why we should expect these kinds of results." Imagine the same privacy game, but now define the likelihood ratio

$$L_i = \log \frac{q_i(z_i|x^{(0)})}{q_i(z_i|x^{(1)})}.$$

We'll work with pure differential privacy, so that we know the L_is are bounded in $[-\varepsilon, \varepsilon]$, and furthermore

$$\mathbb{E}_{q(\cdot|x^{(0)})}[L_i] = D_{\mathrm{KL}}\left(Q_i(\cdot|x^{(0)})||Q_i(\cdot|x^{(1)})\right) \lesssim \min(\varepsilon, \varepsilon^2).$$

So we have bounded random variables with means bounded by ε^2, and we should expect some kind of concentration inequality because we have a martingale difference sequence $L_i - \mathbb{E}[L_i|Z_1^{i-1}]$. Thus

$$\mathbb{P}\left(\sum_{i=1}^{k} L_i - \mathbb{E}[L_i|Z_1^{i-1}] \geq t\right) \leq \exp\left(-\frac{t^2}{k\varepsilon^2}\right).$$

In particular, since the mean is upper bounded by $k\varepsilon^2$, we can solve for the probability by solving $\delta = \exp\left(-\frac{t^2}{k\varepsilon^2}\right) \implies t = \varepsilon\sqrt{k \log \frac{1}{\delta}}$. Thus

$$\sum_{i=1}^{k} L_i \lesssim k\varepsilon^2 + \varepsilon\sqrt{k \log \frac{1}{\delta}} \quad \text{with probability at least } 1-\delta,$$

and that looks a lot like the bound we just showed.

Example 118

We'll close with a family of algorithms that end up being strongly optimal for a lot of problems in privacy; there are some connections with stable estimators and this is an active research area.

Suppose we want to compute some statistic $f(P_n)$ which is real-valued (think of this as one coordinate of a linear model, like a treatment effect). Abstractly, think of most private mechanisms as computing some associated quantity that's stable under perturbations and computing that quantity by adding some noise. The nicest stable function is the following (invented by Asi in 2020):

Definition 119

Define the **inverse sensitivity** for the function f by

$$d(t; P_n) = \inf\{d_{\mathrm{Ham}}(P_n, P_n') : f(P_n') = t\}$$

(in words, this is how many observations we need to alter get our function to output a given value).

Notice that this function is 1-Lipschitz with respect to changing the underlying dataset (since we can just change the value back after P_n changes to P_n'). So our mechanism should be to release the value t with probability proportional to $\exp\left(-\frac{\varepsilon}{2}d(t; P_n)\right)$; this is ε-differentially-private, since if we look at the likelihood ratios we have

$$\frac{dQ(M(P_n) = t)}{dQ(M(P_n') = t)} = \exp\left(-\frac{\varepsilon}{2}(d(t; P_n) - d(t; P_n')\right) \cdot \frac{\int \exp\left(-\frac{\varepsilon}{2}d(t; P_n')\right) dt}{\int \exp\left(-\frac{\varepsilon}{2}d(t; P_n)\right) dt}$$

(the latter fraction is the normalizing constants). But then the first term is bounded by $\exp(\frac{\varepsilon}{2})$ and so is the ratio of the integrands, so we get an overall bound of $\exp(\frac{\varepsilon}{2}) \cdot \exp(\frac{\varepsilon}{2}) = e^{\varepsilon}$, as desired.

Example 120 (Median mechanism)

Suppose our function f is the median of our dataset. If we want to move our median to some new place, then we need to move all data points between our true and target median to the new value t; in particular the inverse sensitivity is given by

$$d(t; P_n) = \#\{x_i : x_i \in [\text{Med}(P_n), t]\}.$$

So for the data release, we have a bunch of "shells" around our true median, where S_k is the set of values t with $d(t, P_n) = k$ (so S_0 is just the median itself, S_1 is the two adjacent intervals to it, S_2 is the next two outer intervals, and so on). We then have equal probability of releasing anything within S_k, with a weighting factor of $e^{-\varepsilon k/2}$. Thus if $V_k = \text{Vol}(S_k)$, we choose an index k with probability $\frac{V_k e^{-\varepsilon k/2}}{\sum_k V_k e^{-\varepsilon k/2}}$, and then we uniformly pick a value within that particular shell S_k.

For a heuristic accuracy guarantee of this mechanism strategy, define the modulus of continuity

$$\omega_k(P_n) = \sup_{P_n'} \{|f(P_n') - f(P_n)| : d_{\text{Ham}}(P_n, P_n') \le k\}.$$

(So ω_k might be big in some parts of the distribution space where f is less stable, but small in parts where it is stable.) Now again define these shells S_k of points, which is the set of targets we get to in k steps:

$$S_k = \{t : \omega_{k-1}(P_n) < |f(P_n) - t| \le \omega_k(P_n)\}.$$

Then the error can be summed over all shells:

$$\mathbb{E}[|M(P_n) - f(P_n)|] \lesssim \sum_{k=0}^{n} \omega_k(P_n) \mathbb{P}(M(P_n) \in S_k).$$

Here's the handwavy heuristic part: somehow $\mathbb{P}(M(P_n) \in S_k)$ should be proportional to the density $e^{-k\varepsilon}$ if things are nice, and thus we can bound this by $\sum_{k=0}^{n} \omega_k \frac{e^{-k\varepsilon}}{\text{normalization}}$ and only keep the "largest contribution" around $k \sim \frac{1}{\varepsilon}$. Thus this is something like $\omega_{1/\varepsilon} + \text{error}$. But this should be optimal as a bound, because if we have two samples P_n, P_n' differing in exactly $\frac{1}{\varepsilon}$ entries, then under ε-differential privacy we have

$$\frac{\mathbb{P}(M(P_n) \in A)}{\mathbb{P}(M(P_n') \in A)} \le \exp\left(\varepsilon \cdot \frac{1}{\varepsilon}\right) = e,$$

so up to a constant error we can't distinguish with any accuracy. So it makes sense that this is the threshold for how far we can move and still get guarantees.

11 October 28, 2025

We'll begin lower bounds today (chapter 9) – instead of demonstrating that things work, we'll demonstrate that things can't possibly work. This is nice because it's fundamentally interesting but also often shows us what the limits of various procedures look like, so that we can demonstrate optimality and find ways to circumvent those bounds.

The setup is as follows: we'll have some family of distributions $\mathcal{P}$ and some parameter that we want to estimate, which we think of as a mapping $\theta : \mathcal{P} \to \Theta$ (abstractly we don't really care if this is a parametric problem or not).

We'll then measure distances via some metric $\rho : \Theta \times \Theta \to \mathbb{R}_+$, and we have some loss function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ which is nondecreasing (so the more wrong we are, the worse the loss is).

Example 121

For example, we may want to measure the mean $\theta(P) = \mathbb{E}_P[X]$ (and we're not saying that the mean fully characterizes the distribution or anything). Or in M-estimation, we may have some loss function ℓ, and we want the parameter

$$\theta(P) = \operatorname{argmin}_\theta \mathbb{E}_P[\ell(\theta, Z)].$$

The idea is that we will have some estimator $\hat{\theta}$ whose **risk** (or **expected loss**) will be given by

$$\mathbb{E}_P\left[\Phi(\rho(\hat{\theta}(X_1^n), \theta(P)))\right]$$

for X_1^n iid from P as usual. We need a way to measure loss without making it per-distribution (otherwise we can just set $\hat{\theta}$ to be $\theta(P)$ itself), and Wald introduced the **minimax principle** for this in the 1940s, wanting us to measure instead the maximum risk

$$\sup_{P \in \mathcal{P}} \left[\Phi(\rho(\hat{\theta}, \theta(P)))\right],$$

so that uniformly across the set of possible distributions we need to do well. So the best possible estimator under this notion is the **minimax risk** of the problem, which we denote as

$$\mathcal{M}_n(\theta(P), \Phi \circ \rho) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\Phi(\rho(\hat{\theta}(X_1^n), \theta(P)))\right].$$

Here $\hat{\theta}$ "knows" (or "can depend on") the values X_1 through X_n, but also the family of possible distributions $\mathcal{P}$, as well as the loss Φ and metric ρ.

We'll develop some information theoretic tools to help develop lower bounds, and the first step is a **reduction from estimation to testing**. The picture (which should sound familiar) in mind is that we have a bunch of points $\theta^1, \cdots, \theta^k$ in our space of possible parameters Θ. If all of these points have radius-δ balls around them that are disjoint, and we can estimate our parameter to accuracy δ, then we should be able to test which distribution actually generates our data. The math version of this is the following:

Definition 122

A family $\{P_v\}_{v \in \mathcal{V}}$ of distributions induces a **2δ-packing** if the parameters $\theta_v = \theta(P_v)$ satisfy $\rho(\theta_v, \theta_{v'}) \geq 2\delta$ for all $v \neq v'$. From this, we can construct the following **canonical hypothesis test**: pick $V \in \mathcal{V}$ uniformly at random and draw $X \sim P_v$ conditional on $V = v$. We then "test the index v."

Proposition 123 (Estimation-to-testing lower bound)

Suppose that $\{P_v\}_{v \in \mathcal{V}}$ induces a 2δ-separation. Then

$$\mathcal{M}_n(\theta(P), \Phi \circ \rho) \geq \Phi(\delta) \inf_\psi \mathbb{P}(\psi(X) \neq V),$$

where $\mathbb{P}(\psi(X) \neq V)$ is exactly the probability of error in the canonical hypothesis test.

Proof. Let $\theta_v = \theta(P_v)$ be the induced parameters in our packing set. We have the trivial lower bound for any P that

$$\mathbb{E}_P\left[\Phi(\rho(\hat{\theta}(X), \theta(P)))\right] \geq \Phi(\delta)\mathbb{P}_P(\rho(\hat{\theta}, \theta) \geq \delta)$$

because if $\rho(\hat{\theta}(X), \theta(P)) \geq \delta$, then Φ of that quantity is at least $\Phi(\delta)$. Also we have that the maximum of a set is at least as big as the average, and in particular

$$\max_v \mathbb{P}_{P_v}\left(\rho(\hat{\theta}, \theta) \geq \delta\right) \geq \frac{1}{|\mathcal{V}|}\sum_{v \in V} \mathbb{P}_{P_v}(\rho(\hat{\theta}, \theta_v) \geq \delta).$$

So now define the "result of our canonical test" $\psi(\hat{\theta}) = v$ if $\rho(\hat{\theta}, \theta_v) < \delta$ and otherwise output some arbitrary element in $\mathcal{V}$. Because of the 2δ-separation, this is well-defined (only one possible θ_v can satisfy this). Thus in terms of probabilities we have

$$P_v(\rho(\hat{\theta}, \theta_v) \geq \delta) \geq P_v(\psi \neq v),$$

and plugging back in yields

$$\max_v \mathbb{P}_{P_v}\left(\rho(\hat{\theta}, \theta) \geq \delta\right) \geq \inf \frac{1}{|\mathcal{V}|}\sum_{v \in \mathcal{V}} P_v(\psi \neq v) = \inf_\psi \mathbb{P}(\psi(X) \neq V),$$

so plugging that back into our bound above yields the result. □

The idea is that we want a tradeoff where we have a family of distributions with large separation δ, but where we still have a constant probability of error $\mathbb{P}(\psi(X) \neq V)$ in the canonical test. So we'll basically come up with situations where we keep that probability of mistake $\frac{1}{2}$, because we do already know how to prove lower bounds on errors of tests from earlier in the course.

Example 124

In **Le Cam's 2-point method**, the idea is that in low-dimensional problems it is often sufficient for rates of convergence to consider only two distributions P_0, P_1. We know already that the probability of error satisfies (by Le Cam's inequality)

$$\frac{1}{2}\inf_\psi \{P_0(\psi \neq 0) + P_1(\psi \neq 1)\} = \frac{1}{2}\left(1 - ||P_0 - P_1||_{\mathrm{TV}}\right),$$

so if we can come up with a pair of distributions with separation $\rho(\theta(P_0), \theta(P_1)) \geq 2\delta$, then we get a bound on the minimax risk for n observations

$$\mathcal{M}_n(\theta(P), \Phi \circ \rho) \geq \frac{\Phi(\delta)}{2}(1 - ||P_0^n - P_1^n||_{\mathrm{TV}}).$$

The game plan will thus be to find the largest δ such that we still have $||P_0^n - P_1^n||_{\mathrm{TV}} \leq \frac{1}{2}$. We typically will use Pinsker's inequality

$$||P_0^n - P_1^n||_{\mathrm{TV}}^2 \leq \frac{1}{2}(P_0^n||P_1^n) = \frac{n}{2}D_{\mathrm{KL}}(P_0||P_1),$$

and often D_{KL} will scale like something like $O(\delta^2)$. So we can then choose something like $\delta = \frac{1}{\sqrt{n}}$.

Example 125

Suppose we have a family of normal distributions $\mathcal{P} = \{N(\theta, \sigma^2)\}_{\theta \in \mathbb{R}}$. We'll get a lower bound by picking $P_0 = N(0, \sigma^2)$ and $P_1 = N(2\delta, \sigma^2)$.

We can calculate that

$$D_{\mathrm{KL}}(P_0||P_1) = \frac{(2\delta)^2}{2\sigma^2} = \frac{2\delta^2}{\sigma^2},$$

so the squared error satisfies (that is, our loss function is $x \mapsto x^2$)

$$\begin{aligned}\max_{P\in\{P_0,P_1\}} \mathbb{E}_P\left[(\hat{\theta}(X_1^n)-\theta)^2\right] &\geq \frac{\delta^2}{2}\left(1-||P_0^n - P_1^n||_{\mathrm{TV}}\right)\\ &\geq \frac{\delta^2}{2}\left(1-\sqrt{\frac{n}{2}D_{\mathrm{KL}}(P_0||P_1)}\right)\\ &= \frac{\delta^2}{2}\left(1-\sqrt{\frac{n\delta^2}{\sigma^2}}\right).\end{aligned}$$

Thus choosing $\delta = \frac{\sigma^2}{4n}$ makes the right-hand side $\frac{\delta^2}{4} = \frac{\sigma^2}{16n}$, meaning that for all sample sizes, we have

$$\inf_{\hat{\theta}} \max_{\theta\in\{0,2\delta\}} \mathbb{E}[(\hat{\theta}-\theta)^2] \geq \frac{\sigma^2}{16n}.$$

That is, for all estimators $\hat{\theta}$, we will have expected squared error $\frac{\sigma^2}{16n}$ for at least one of the two specified Gaussians (and so in particular we'll have that error over all possible Gaussians in our $\mathcal{P}$). And we can achieve this up to numerical constants by just actually taking the sample mean, which gives expected squared error $\frac{\sigma^2}{n}$.

Remark 126. *In fact $\frac{\sigma^2}{n}$ is the correct lower bound – there was a lot of work in the 1960s on how many points you need to include to get the right asymptotic constant of 1 instead of $\frac{1}{16}$, but it's really not so important. And instead of having a maximum over $\{0, 2\delta\}$, we could also just draw θ from a Gaussian distribution and do some Bayesian calculations.*

But often we do care about high-dimensional problems, and then we'll need something more powerful:

Example 127

In the **local Fano method**, the big-picture principles are the same: we often have to pack in a lot more potential solutions. Recall that for a Markov chain $V \to X_1^n \to \hat{V}$ for V uniform on some set $\mathcal{V}$, we have by Fano's inequality

$$\mathbb{P}(V \neq \hat{V}) \geq 1 - \frac{I(X_1^n; V) + \log 2}{\log|\mathcal{V}|}$$

a bound in terms of "how many bits we can store of V."

Furthermore, since our X_1^n are iid from P_v given $V = v$ in the canonical hypothesis test, we have

$$\begin{aligned}I(V; X_1^n) = \sum_{i=1}^n I(X_i; V|X_1^{i-1}) &= \sum_{i=1}^n H(X_i|X_1^{i-1}) - H(X_i|V, X_1^{i-1})\\ &\leq \sum_{i=1}^n H(X_i) - H(X_i|V)\\ &= \sum_{i=1}^n I(X_i; V),\end{aligned}$$

so we really only need to worry about single individual observations in our bounds. And we know we can write mutual information as a KL-divergence, so

$$I(X; V) = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} D_{\mathrm{KL}}(P_v||\overline{P})$$

where $\overline{P}$ is the average of P_v uniformly over all v. By convexity of KL, this thus means

$$I(X;V) \le \frac{1}{|\mathcal{V}|^2}\sum_{v,v'} D_{\mathrm{KL}}(P_v||P_{v'}),$$

and this is why we call this a "local method:" we'll typically have that each KL term is less than something like $\kappa^2\delta^2$ where δ is the separation of choice. Then scaling δ appropriately so that probability of error becomes a constant again gives us the bound we want, since we get that the minimax risk is

$$\mathcal{M}_n(\theta(P), \Phi\circ\rho) \ge \Phi(\delta)\cdot\left(1 - \frac{\kappa^2 n\delta^2 + \log 2}{\log|\mathcal{V}|}\right).$$

In such a situation, we then choose $\delta^2 = \frac{1}{2}\frac{\log|\mathcal{V}|}{n\kappa^2}$. And when we do this, what's important is that the packing elements are fixed and we shrink them down to zero, so the cardinality $|\mathcal{V}|$ doesn't depend on δ.

Example 128

Suppose we want to estimate the mean in the normal location family $\{N(\theta,\sigma^2 I_d)\}_{\theta\in\mathbb{R}^d}$ **using the ℓ^∞ metric**. So first we need to choose our packing elements, and we'll do so by setting $\mathcal{V} = \{\pm e_j\}_{j=1}^d$. Associated to $\mathcal{V}$, we choose $\theta(P_v) = 2\delta v$ for δ to be chosen later.

Then we have $||\theta_v - \theta_{v'}||_\infty \ge 2\delta$ in all cases, and the KL-divergence between observations

$$D_{\mathrm{KL}}(P_v||P_v') = \frac{(2\delta)^2||v-v'||_2^2}{2\sigma^2} = \frac{2\delta^2}{\sigma^2}||v-v'||_2^2.$$

Thus we get the bound

$$\begin{aligned} I(X_1^n;V) &\le n\frac{1}{|\mathcal{V}|^2}\sum_{v,v'} D_{\mathrm{KL}}(P_v||P_{v'}) \\ &= \frac{2n\delta^2}{\sigma^2}\mathbb{E}[||V-V'||_2^2] \\ &= \frac{4n\delta^2}{\sigma^2}, \end{aligned}$$

so that

$$\inf_{\hat\theta}\sup_{\theta\in\mathbb{R}^d}\mathbb{E}_\theta\left[||\hat\theta(X_1^n)-\theta||_\infty\right] \ge \delta\cdot\left(1 - \frac{\frac{4n\delta^2}{\sigma^2}+\log 2}{\log(2d)}\right),$$

so choosing $\delta = \frac{\sigma^2\log(2d)}{8n}$ gives us a lower bound of $\frac{\delta}{2}$; thus,

$$\inf_{\hat\theta}\sup_{\theta\in\mathbb{R}^d}\mathbb{E}_\theta\left[||\hat\theta(X_1^n)-\theta||_\infty\right] \gtrsim \sqrt{\frac{\sigma^2\log(2d)}{n}}.$$

And the sample mean does achieve this up to numerical constants, because we're measuring in the ℓ^∞ norm.

Example 129

We'll now show a more general recipe of construction involving "taking a fixed packing and rescaling by δ."

One way we can construct such packings is to make use of the **probabilistic method** to demonstrate existence, and the following came from coding theory:

Lemma 130 (Gilbert-Varshamov bound)
There exists a packing in "bit-space" $\mathcal{V} \subset \{\pm 1\}^d$ with $||v - v'||_1 \geq \frac{d}{2}$ for all v, v' and $|\mathcal{V}| \geq \exp(cd)$ for some numerical constant c.

Proof. Let v, v' be iid on the hypercube. Each differing coordinate is 2 apart in ℓ^1 norm, so $\mathbb{E}[||v - v'||_1] = d$ and thus by sub-Gaussian concentration it's very likely that this quantity is at least $\frac{d}{2}$: we have iid Bernoullis $B_i = 1\{V_i \neq V_i'\}$ which are each $\frac{1}{4}$-sub-Gaussian, so $||v - v'||_1 = 2\sum_{i=1}^d B_i$ is sub-Gaussian with constant d and thus

$$\mathbb{P}\left(||v - v'||_1 < \frac{d}{2}\right) = \mathbb{P}\left(2\sum_{i=1}^d B_i - d \leq -\frac{d}{2}\right) \leq \exp\left(-\frac{1}{2}\left(\frac{d}{2}\right)^2 \cdot \frac{1}{d}\right) = e^{-d/8}.$$

So now if we take M iid such $V^{(i)}$ uniform on the hypercube, then the probability that any pair i, j satisfies $||V^{(i)} - V^{(j)}||_1 \leq \frac{d}{2}$ is at most $\binom{M}{2}e^{-d/8} \leq \frac{M^2}{2}\exp(-\frac{d}{8}$ by a union bound, so some packing exists as long as this quantity is at most 1. Thus we can set $M = \sqrt{2\exp\left(\frac{d}{8}\right)} = \sqrt{2}\exp\left(\frac{1}{16}d\right)$, as desired. □

We can now use this to prove lower bounds, using linear regression as an example:

Example 131
Given a matrix $X \in \mathbb{R}^{n\times d}$ of covariates and letting $Y = X\theta + \xi$ for $\xi \sim N(0, \sigma^2 I_n)$ noise, we take the packing $\theta_v = \delta v$ where v is in the hypercube packing $\mathcal{V}$ guaranteed from the above lemma.

In our canonical estimation setting, we then have $V \mapsto Y = X\theta_v + \xi \mapsto \hat{\theta}$, and the information satisfies (we're just using a max instead of a sum)

$$I(V;Y) \leq \max_{v,v'\in\mathcal{V}} D_{\mathrm{KL}}(N(\delta X\theta_v, \sigma^2 I)||N(\delta X\theta_{v'}, \sigma^2 I)) = \max_{v,v'} \frac{1}{2\sigma^2}||X(v - v')||_2^2\delta^2 \lesssim \frac{d\delta^2}{\sigma^2}\lambda_{\max}(X^TX).$$

Our separation between the δ_vs in this setting is $\delta^2||v - v'||_2^2 \gtrsim d\delta^2$ because of our choice of packing. Thus our lower bound in ℓ^2 error is

$$\mathcal{M}_n(\theta(P), ||\cdot||_2^2) \gtrsim d\delta^2\left(1 - \frac{I(V;Y) + \log 2}{\log(|\mathcal{V}|)}\right) \gtrsim d\delta^2\left(1 - \frac{\frac{d\delta^2}{\sigma^2}||X||_{\mathrm{op}}^2 + \log 2}{cd}\right).$$

So to make the probability of error constant, we choose $\delta^2 = \frac{\sigma^2}{||X||_{\mathrm{op}}^2}$, so that up to a numerical constrant we have

$$\mathcal{M}_n(\theta(P), ||\cdot||_2^2) \gtrsim \frac{d\sigma^2}{||X||_{\mathrm{op}}^2} = \frac{1}{||\frac{1}{\sqrt{n}}X||_{\mathrm{op}}^2}\frac{d\sigma^2}{n}.$$

There's some looseness and sharpness here: if X is perfectly well-designed "orthogonal design" whose columns' ℓ^2-norms are all $\sqrt{n}$ and orthogonal (that is, $X^TX = nI$), then the lower bound is $\frac{d\sigma^2}{n}$, which is completely sharp.

12 October 30, 2025

Today's lecture will be given by **Saminul Haque**, the TA for the course.

Last time, we defined the minimax error $\mathcal{M}_n(\theta(P), \Phi\circ\rho)$ to be the worst-case error of our expected loss function. We worked out last time that we can reduce this to a testing problem: if $\{P_v\}_{v\in\mathcal{V}}$ is a set of distributions with the separation condition $\rho(\theta_v, \theta_{v'}) \geq 2\delta$, then we get a lower bound of $\Phi(\delta)\inf_\psi \mathbb{P}(\psi(X) \neq V)$ (that is, $\Phi(\delta)$ times the

smallest probability of guessing incorrectly in a test where V is uniformly chosen on $\mathcal{V}$). We then established Le Cam's method for bounding this in the binary case $V = \{0, 1\}$, so that the infimum just becomes the total variation distance between the two distributions, as well as a more general local Fano method where we choose some rescaled packing.

What we'll do today is that instead of getting dimensionality through a set V, we'll try to "use Le Cam's over each coordinate."

Example 132

In **Assouad's method**, we will fix our set $\mathcal{V}$ to be the hypercube $\{\pm 1\}^d$. Suppose we have an "additive condition" on our loss in the following way: suppose there exists some tester $\hat{v} : \Theta \to \mathcal{V}$ from parameters to the hypercube such that for all $v \in \mathcal{V}$ and $\theta \in \Theta$, we have the **δ-Hamming separation condition**

$$\Phi(\rho(\theta, \theta_v)) \geq \delta \sum_{j=1}^{d} 1\{\hat{v}_j(\theta) \neq v_j\}.$$

In other words, for all vertices of the hypercube, the error of θ is at least δ times the number of wrong coordinates.

As a concrete example, we can let $P_v = N(\delta v, \sigma^2 I)$ and let our parameter be the mean $\theta_v = \delta v$. With the ℓ^1 error, the loss would be

$$||\theta - \theta_v||_1 = \sum_{j=1}^{d} |\theta_j - \delta v_j|.$$

We're trying to guess v_j from θ, and we always know that v_j is either 1 or -1, so a good guess would be the sign of θ (so if we guess the right one, we get zero loss, and otherwise our loss is at least δ because we're on the wrong side of the number line. Therefore

$$||\theta - \theta_v||_1 \geq \delta \sum_{j=1}^{d} 1\{\text{sgn}(\theta_j) \neq v_j\}.$$

Similarly, we can do the ℓ^2 squared error as well with

$$||\theta - \theta_v||_2^2 = \sum_{j=1}^{d} |\theta_j - \delta v_j|^2.$$

But we can do the same thing and take $\text{sgn}(\theta_j)$ (and in fact usually we have a sign prediction problem with this method), and we'll then get a δ^2-Hamming separation instead.

So now that we have this additive component, v_j is binary and so for a fixed j, this term basically corresponds to a binary testing problem between vertices where v_j is $+1$ or -1. So just like Le Cam, we get some total variation terms but we have to average it over all distributions. Writing this more formally, define

$$P_{+j} = X_1^n | v_j = 1, \quad P_{-j} = X_1^n | v_j = -1.$$

In other words, remembering that X_1^n are generated from P_v given V, we have

$$P_{+j} = \frac{1}{2^d} \sum_{V \in \{\pm 1\}^d} P_{v,+j}.$$

Thus we get that the reduction is

$$\mathcal{M}_n \geq \delta \sum_{j=1}^{d} \left(1 - ||P_{+j} - P_{-j}||_{\text{TV}}\right).$$

The difference between this and Le Cam is that we have an explicit dimension dependence, and also Fano instead gets δ scaling with dimension if our points are well-separated (but here δ comes from a single coordinate and will thus be

dimension-independent).

The quantity $||P_{+j} - P_{-j}||_{\text{TV}}$ is often ugly to work with, so we can often get some upper bounds. For one thing, plugging in the definitions of P_{+j} and P_{-j} and using convexity yields

$$||P_{+j} - P_{-j}||_{\text{TV}} \leq \frac{1}{2^d} \sum_v ||P_{v,+j} - P_{v,-j}||_{\text{TV}},$$

but often this is a bit too coarse. So instead we can incorporate the different js together a bit more nicely by observing that

$$\begin{aligned}
\sum_{j=1}^d ||P_{+j} - P_{-j}||_{\text{TV}} &\leq \sqrt{d} \left(\sum_{j=1}^d ||P_{+j} - P_{-j}||_{\text{TV}}^2 \right)^{1/2} \\
&\leq \sqrt{d} \left(\frac{1}{2^d} \sum_{j=1}^d \sum_v ||P_{v_j,+j} - P_{v,-j}||_{\text{TV}}^2 \right)^{1/2} \\
&= d \left(\frac{1}{d2^d} \sum_{j=1}^d \sum_v ||P_{v_j,+j} - P_{v,-j}||_{\text{TV}}^2 \right)^{1/2}
\end{aligned}$$

by Jensen in the middle step. This yields (going back to our formula for $\mathcal{M}_n$) that

$$\mathcal{M}_n \geq \delta d \left(1 - \left(\frac{1}{d2^d} \sum_{j=1}^d \sum_v ||P_{v,+j} - P_{v,-j}||_{\text{TV}}^2 \right)^{1/2} \right).$$

So all we need to do is come up with a parametrization and find the Hamming separation, and then we can find the TV distance (we typically just want to choose the TV on the inside to be $\frac{1}{2}$). Let's do an example of this:

Example 133

Returning to the linear regression model from last time, we have some fixed design matrix $X \in \mathbb{R}^{n \times d}$ and observe $Y = X\theta + \xi$ for $\xi \in N(0, \sigma^2 I_n)$. We wish to estimate θ in $|| \cdot ||_2^2$.

By what we said before, we can take $\theta_v = \delta v$ so that our Hamming separation is δ^2 under this loss. So the only thing that remains is to compute, by Pinsker,

$$\begin{aligned}
||P_{v,+j} - P_{v,-j}||_{\text{TV}}^2 &\leq \frac{1}{2} D_{\text{KL}}(P_{v,+j} || P_{v,-j}) \\
&= \frac{1}{2} D_{\text{KL}}(N(\delta X v^{+j}, \sigma^2 I) || N(\delta X v^{-j}, \sigma^2 I))
\end{aligned}$$

where v^{+j}, v^{-j} just mean v but with the jth coordinate set to 0 or 1. But we have the same variance on both sides, and the difference in means is exactly $2e_j$. Thus we get

$$||P_{v,+j} - P_{v,-j}||_{\text{TV}}^2 \leq \frac{1}{2} \frac{||\delta X(v^{+j} - v^{-j}||_2^2}{2\sigma^2} = \frac{\delta^2 ||X_{e_j}||_2^2}{\sigma^2},$$

and we could just take an operator norm bound here to get a uniform bound on δ. But we get a tighter bound by plugging this into our sum: we find that (remember we have δ^2-Hamming separation)

$$\mathcal{M}_n \geq \delta^2 d \left(1 - \left(\frac{1}{d2^d} \sum_{j=1}^d \sum_v \frac{\delta^2 ||X_{e_j}||_2^2}{\sigma^2} \right)^{1/2} \right).$$

Now there is no more v-dependence so we can cancel the sum over v and the 2^d factor, and $\sum_j ||Xe_j||_2^2 = \sum_j \text{tr}(e_j^T X^T X e_j) = \sum_j \text{tr}(X^T X e_j e_j^T) = \text{tr}(X^T X)$ by linearity of trace. Thus we get

$$\mathcal{M}_n \geq \delta^2 d \left(1 - \left(\frac{\delta^2}{\sigma^2}\frac{\text{tr}(X^T X)}{d}\right)^{1/2}\right).$$

Choosing $\delta^2 = \frac{\sigma^2 d}{4\text{tr}(X^T X)}$ then shows that

$$\mathcal{M}_n \geq \frac{\sigma^2 d^2}{8\text{tr}(X^T X)}.$$

Compare this to Fano's method last time, where we found that $\mathcal{M}_n \gtrsim \frac{\sigma^2 d}{||X^T X||}$. Notice that $\text{tr}(X^T X) \leq d||X^T X||$ (since the left-hand side is the sum of all eigenvalues and the operator norm is the largest eigenvalue). So our bound we're getting now is at least as good as the one we had last time, and the tightness comes from how we didn't take a uniform operator norm in our estimation of TV distance. It turns out this is actually still not tight: the actual order is something like $\sigma^2 \text{Tr}((X^T X)^{-1})$, but more careful applications of Assouad can get us there and we won't worry about that further here.

So given that Fano and Assouad are kind of similar, **Assouad turns out to be more useful for adaptive or interactive settings** since it's hard to reason about information over multiple rounds. We'll see some exericses for this.

Example 134

Next, we'll apply these methods to parametric estimation to get the Fisher score lower bound. Assume we have a parametric family of distributions $\{p_\theta\}_{\theta \in \mathbb{R}}$, and we have the **Fisher score**

$$\dot{\ell}_\theta(x) = \nabla_\theta \log p_\theta(x) = \frac{\dot{p}_\theta(x)}{p_\theta(x)}.$$

Our goal is to think about and why for "accurate estimators" $\hat{\theta}$, the error $\hat{\theta} - \theta$ should somehow be related to the empirical score $P_n \dot{\ell}_\theta$ because of some "asymptotic minimax." We'll be able to analyze this for the squared error.

We'll introduce the **Fisher information matrix**

$$J(\theta) = \mathbb{E}_\theta\left[\dot{\ell}_\theta \dot{\ell}_\theta^T\right] = \text{Cov}_\theta(\dot{\ell}_\theta(X)).$$

We'll assume enough regularity conditions to not worry about swapping integrals and derivatives. Define the risk

$$L(\beta) = \mathbb{E}_\theta[-\log p_\beta(x)],$$

so that the excess risk is exactly given by

$$L(\beta) - L(\theta) = \mathbb{E}_\theta\left[\log \frac{p_\theta(x)}{p_\beta(x)}\right] = D_{\text{KL}}(P_\theta || P_\beta).$$

We can do a Taylor expansion and write this as

$$\nabla L(\theta)^T(\beta - \theta) + (\beta - \theta)^T \nabla^2 L(\theta)(\beta - \theta) + O(||\beta - \theta||^3)$$

and the first term is just zero because our loss is minimized at θ. So we really care about the quadratic term – picking a critical threshold for KL is like estimating the range of smallness, and we need to connect $\nabla^2 L(\theta)$ to our Fisher

information matrix. We have

$$\begin{aligned}\nabla^2 L(\theta) &= -\int (\nabla_\theta^2 \log p_\theta(x)) p_\theta(x) dx \\ &= -\int \left(\frac{\ddot{p}_\theta}{p_\theta} - \frac{\dot{p}_\theta \dot{p}_\theta^T}{p_\theta^2} \right) p_\theta(x) dx \\ &= 0 + \int \dot{\ell}_\theta \dot{\ell}_\theta^T p_\theta(x) dx \\ &= J(\theta).\end{aligned}$$

So here's where things become more heuristic: we've seen in high-dimensional problems that we can't do better than a KL-divergence of $\frac{d}{n}$. So if we take our alternatives of the form

$$\beta = \theta + J(\theta)^{-1/2} U$$

for U drawn from Unif$(\sqrt{\frac{d}{n}}\mathbb{S}^{d-1})$, then we're left with D_{KL} on the order of $\frac{d}{n}$, and the parameter separation we have then yields

$$\mathcal{M}_n \gtrsim \mathbb{E}\left[||J(\theta)^{-1/2} U||_2^2 \right].$$

But this works out to just $\frac{\mathrm{tr}(J(\theta)^{-1})}{n}$ (since covariance on the unit sphere is $\frac{1}{d}\mathrm{Id}$). So we skew the noise in the $J^{-1/2}$ direction to be isotropic in KL-space. (In a nice well-conditioned problem like mean estimation, $J(\theta)$ is just the identity so the numerator is just d. But the worse the conditioning, the smaller the eigenvalues, and so the larger the numerator will be because we can move the parameter a lot before the KL changes.)

In the last section of today, we'll talk about some **classical lower bounds**:

Fact 135 (Cramér-Rao)

In one dimension, if $\hat{\theta}$ is an unbiased estimator for $\theta \in \mathbb{R}$, then

$$\mathbb{E}\left[(\hat{\theta}(X) - \theta)^2 \right] \geq \frac{1}{J(\theta)}$$

In some sense this is the "biggest con in statistics" since we can always actually do better in some sense, but we'll prove a more "real version" of this:

Theorem 136 (Van-Trees, aka Bayesian Cramér-Rao)

Given a parametric family $\{p_\theta\}_{\theta \in \mathbb{R}}$ and a prior π on $[a, b]$ with $\pi(a) = \pi(b) = 0$, define a score

$$J(\pi) = \int_a^b \frac{\pi'(t)^2}{\pi(t)} dt = \int_a^b (\log \pi(t))')^2 \pi(t) dt.$$

Then if θ is drawn according to π and x is drawn according to p_θ, then the mean-squared error satisfies

$$\mathbb{E}_\pi \left[\mathbb{E}_\theta \left[(\hat{\theta}(X) - \theta)^2 \right] \right] \geq \frac{1}{\mathbb{E}_\pi[J(\theta)] + J(\pi)}.$$

This is pretty similar to the usual proof for Cramér-Rao if we've seen that before:

Proof. Define the new score

$$\dot{\ell}_{\theta,\pi}(x) = \dot{\ell}_\theta(x) + \frac{\pi'(\theta)}{\pi(\theta)}.$$

We then have

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\theta}(X)-\theta\right)\dot{\ell}_{\theta,\pi}(X)\right] &= \int_a^b \int_{\mathcal{X}} (\hat{\theta}_x-\theta)\dot{\ell}_{\theta,\pi}(x)p_\theta(x)\pi(\theta)dxd\theta \\ &= \int_a^b \int_{\mathcal{X}} (\hat{\theta}_x-\theta)\left(\dot{p}_\theta(x)\pi(\theta)+p_\theta(x)\pi'(\theta)\right)dxd\theta \\ &= \int_{\mathcal{X}} \int_a^b (\hat{\theta}_x-\theta)\left[\frac{d}{d\theta}(p_\theta(x)\pi(\theta))\right]d\theta dx \\ &= \int_{\mathcal{X}} \left((\hat{\theta}(x)-\theta)p_\theta(x)\pi(\theta)\Big|_a^b + \int_a^b p_\theta(x)\pi(\theta)d\theta\right)dx\end{aligned}$$

by integration by parts. But by assumption the boundary term is zero, so we're just left with $\int_{\mathcal{X}}\int_a^b p_\theta(x)\pi(\theta)d\theta dx = 1$. On the other hand, we can use Cauchy-Schwarz to find that

$$1 = \mathbb{E}\left[\left(\hat{\theta}(X)-\theta\right)\dot{\ell}_{\theta,\pi}(X)\right] \le \mathbb{E}\left[(\hat{\theta}(X)-\theta)^2\right]^{1/2}\mathbb{E}\left[\dot{\ell}_{\theta,\pi}(X)^2\right]^{1/2},$$

but we can calculate by the definition of $\dot{\ell}_{\theta,\pi}$ that

$$\begin{aligned}\mathbb{E}\left[\dot{\ell}_{\theta,\pi}(X)^2\right] &= \mathbb{E}[\dot{\ell}_\theta(X)^2] + 2\mathbb{E}\left[\dot{\ell}_\theta(X)\frac{\pi'(\theta)}{\pi(\theta)}\right] + \mathbb{E}\left[\frac{\pi'(\theta)^2}{\pi(\theta)^2}\right] \\ &= \mathbb{E}_\pi[J(\theta)] + 2\mathbb{E}\left[\dot{\ell}_\theta(X)\frac{\pi'(\theta)}{\pi(\theta)}\right] + J(\pi),\end{aligned}$$

and we claim this middle term is actually zero. Indeed, when calculating $\mathbb{E}\left[\dot{\ell}_\theta(X)\frac{\pi'(\theta)}{\pi(\theta)}\right]$ we can tower rule condition on θ and pull out $\dot{\ell}_\theta(X)$, and $\mathbb{E}[\dot{\ell}_\theta(X)|\theta]$ is always zero because we have a parametric family:

$$\mathbb{E}_{x\sim p_\theta}[\dot{\ell}_\theta(x)] = \int_{\mathcal{X}} \frac{\dot{p}_\theta(x)}{p_\theta(x)}p_\theta(x)dx = \frac{d}{d\theta}\int_{\mathcal{X}} p_\theta(x)dx = \frac{d}{d\theta}1 = 0.$$

So then we just have $\mathbb{E}_\pi[J(\theta)] + J(\pi)$, and substituting that back in yields the desired inequality. □

13 November 6, 2025

We'll finish up our discussion of squared error and Van Trees and then start talking about constrained risk inequalities, in particular the strong (quantitative) data processing inequalities and lower bounds on private estimation.

We saw last time that for estimators $\hat{\theta}: \mathcal{X} \to \mathbb{R}$, on average the true value θ must correlate with the Fisher score $\dot{\ell}_\theta(x) = \frac{\partial}{\partial\theta}\log p_\theta(x)$. The idea is that the Fisher information quantity gives us lower bounds: for a family $\{p_\theta\}_{\theta\in\mathbb{R}^d}$ we assume the quantity

$$\dot{\ell}_\theta(x) = \nabla_\theta \log p_\theta(x) = \frac{\dot{p}_\theta(x)}{p_\theta(x)}$$

exists, and then we have the Fisher information matrix

$$J(\theta) = \mathbb{E}_\theta\left[\dot{\ell}_\theta(x)\dot{\ell}_\theta(X^T)\right] = \mathrm{Cov}_\theta(\dot{\ell}_\theta) = -\mathbb{E}_\theta[\nabla^2 \log p_\theta(x)].$$

(Checking these equalities is kind of a routine exercise showing that we can swap around integrals and derivatives because boundary terms go to zero in integration by parts – for example, we do always have $\mathbb{E}_\theta[\dot{\ell}_\theta(X)] = 0$. There are some regularity conditions, but they really should hold for most problems.)

Fact 137

The intuition here is that if we define $L_\theta(\beta) = \mathbb{E}_\theta[\log p_\beta(x)]$ to be the expected log-loss when we stick in β instead of θ, then we have

$$L_\theta(\theta) - L_\theta(\beta) = D_{\mathrm{KL}}(p_\theta || p_\beta)$$

so this quantity is positive unless $\theta = \beta$, and assuming smoothness in the parameter we have "first derivative zero" and $\nabla^2_\beta L_\theta(\beta)$ is given by the negative of the Fisher information matrix. So $J(\theta)$ describes local curvature near $\beta = \theta$.

More formally, we have that

$$D_{\mathrm{KL}}(P_\theta || P_\beta) = \frac{1}{2}(\theta - \beta)^T J(\theta)(\theta - \beta) + O(|\theta - \beta|^3),$$

so we kind of believe that the KL-divergence should tell us how hard it is to estimate and thus the Fisher information should as well. Furthermore, for iid sampling (meaning that we observe $X_1^n \sim p_\theta^n$), the Fisher information for p_θ^n is just $J_n(\theta) = nJ(\theta)$.

Theorem 138 (Van Trees in higher dimensions)

Let T be drawn from some prior distribution π, and conditional on $T = \theta$ we draw $X \sim p_\theta$. Assume the prior is a product $\pi = \pi_1 \times \cdots \times \pi_d$ with independent coordinates for simplicity. Also assume that π_j is supported on an interval $[a_j, b_j]$ with $\pi_j(a_j) = \pi_j(b_j) = 0$. Then for any positive definite matrix A, we have

$$\mathbb{E}_\pi \left[\mathbb{E}[(\hat{\theta} - T)^T A^{-1} (\hat{\theta} - T) | T] \right] \geq \frac{d^2}{\mathbb{E}_\pi[\mathrm{tr}(AJ(\theta))] + \mathrm{tr}(AJ(\pi))},$$

where we have the prior Fisher information

$$J(\pi) = \mathbb{E}\left[\nabla_\theta \log \pi(T) \cdot \nabla_\theta \log \pi(T)^T \right].$$

The left-hand side of this inequality can be thought of as $\int_\Theta \mathbb{E}_\theta[(\hat{\theta} - T)^T A^{-1}(\hat{\theta} - T)]\pi(\theta) d\theta$ (so we have the average error over our prior).

Proof. This is basically integration by parts, but unfortunately it really only works for quadratic error. Define the joint log-likelihood

$$\ell_{\theta,\pi}(x) = \log(p_\theta(x)\pi(\theta)).$$

Then

$$\begin{aligned}
\langle \hat{\theta}(x) - \theta, \dot{\ell}_{\theta,\pi}(x) \rangle &= \sum_{j=1}^d (\hat{\theta}_j(x) - \theta_j) \cdot \frac{\partial}{\partial \theta_j} \log(p_\theta(x)\pi(\theta)) \\
&= \sum_{j=1}^d (\hat{\theta}_j(x) - \theta_j) \left(\frac{\frac{\partial}{\partial \theta_j} p_\theta(x)}{p_\theta(x)} + \frac{\pi_j'(\theta_j)}{\pi_j(\theta_j)} \right)
\end{aligned}$$

by the product rule and using that the prior is a product distribution. But then this is just

$$\sum_{j=1}^d (\hat{\theta}_j(x) - \theta_j) \left(\frac{\frac{\partial}{\partial j}(p_\theta(x)\pi_j(\theta_j))}{p_\theta(x)\pi_j(\theta_j)} \right),$$

so if we integrate any individual term for a fixed j we get

$$\int_{a_j}^{b_j} (\hat{\theta}_j(x) - \theta_j) \left(\frac{\frac{\partial}{\partial j}(p_\theta(x)\pi_j(\theta_j)}{p_\theta(x)\pi_j(\theta_j)} \right) \pi_j(\theta_j) p_\theta(x) d\theta_j dx = \int_{a_j}^{b_j} (\hat{\theta}_j(x) - \theta_j) \frac{\partial}{\partial \theta_j}(p_\theta(x)\pi_j(\theta_j)) d\theta_j dx$$
$$= (\hat{\theta}_j(x) - \theta_j)(p_\theta(x)\pi_j(\theta_j))\Big|_{a_j}^{b_j} + \int_{a_j}^{b_j} 1 \cdot p_\theta(x)\pi_j(\theta_j) d\theta_j dx$$

by integration by parts in θ_j. The boundary term is just zero and the integral will just be 1, so now taking the joint expectation over $T \sim \pi$ and $X \sim p_\theta$ (given $T = \theta$) yields

$$\mathbb{E}\left[\langle \hat{\theta}(x) - T, \dot{\ell}_{T,\pi}(x) \rangle\right] = \sum_{j=1}^{d} \iint (\hat{\theta}_j(x) - \theta_j) \frac{\partial}{\partial j} (\log(p_\theta(x)\pi(\theta)) p_\theta \pi(\theta) = d,$$

and now the rest of the proof is just Cauchy-Schwarz. For fixed T,

$$\langle \hat{\theta} - \theta, \dot{\ell}_{\theta,\pi} \rangle = \langle A^{-1/2}(\hat{\theta} - \theta), A^{1/2}\dot{\ell}_{\theta,\pi} \rangle \leq ||A^{-1/2}(\hat{\theta} - \theta)||_2 ||A^{1/2}\dot{\ell}_{\theta,\pi}||_2,$$

so now taking expectations yields

$$d \leq \mathbb{E}\left[||A^{-1/2}(\hat{\theta} - T)||_2 ||A^{1/2}\dot{\ell}_{T,\pi}||_2 \right]$$
$$\leq \mathbb{E}\left[||A^{-1/2}(\hat{\theta} - T)||_2^2 \right]^{1/2} \mathbb{E}\left[||A^{1/2}\dot{\ell}_{T,\pi}||_2^2 \right]^{1/2}$$

by Cauchy-Schwarz, and now the latter expectation is just $\mathbb{E}[\text{tr}(AJ(T)) + \text{tr}(AJ(\pi))]$ by an explicit computation. Rearranging and squaring yields the result. □

What's interesting is to try to get useful corollaries out of this: we instantiate with different models and any particular model that we want, usually supported on a small neighborhood of some particular point of interest θ^*. We'd like to have a per-parameter lower bound, and the starting point for that is to let π_0 be some fixed density on $[-1, 1]$ with $\pi_0(-1) = \pi_0(1) = 0$ (it actually really doesn't matter what we choose, but we can do something like $\exp\left(-\frac{1}{(|t|-1)^2}\right)$ and consider the sequence of priors π_n

$$\pi_n(t) = \prod_{j=1}^{d} r_n \pi_0\left(r_n(t_j - \theta_j^*)\right)$$

where $r_n \to \infty$ is any rate. So basically we're taking increasingly tight regions around some θ^*, and by a u-substitution calculation we can check that

$$J(\pi_n) = r_n^2 I_d J(\pi_0).$$

We then automatically have a generic lower bound: for any θ^* of interest, the expected error over a box

$$\int_{\theta^* + [-\frac{1}{r_n}, \frac{1}{r_n}]^d} \mathbb{E}_\theta\left[(\hat{\theta} - \theta)^T A^{-1}(\hat{\theta} - \theta)\right] \pi_n(\theta) d\theta \geq \frac{d^2}{\int \text{tr}(A(J(\theta))\pi_n(\theta) d\theta + O(r_n^2)}.$$

For one more simplification, suppose that we actually get iid observations $X_1^n \sim p_\theta$, so that $J_n(\theta) = nJ_1(\theta)$. We then get

$$\int_{\theta^* + \frac{1}{r_n}[-1,1]^d} \mathbb{E}\left[(\hat{\theta}(X_1^n) - \theta)^T A^{-1}(\hat{\theta}(X_1^n) - \theta)\right] \geq \frac{d^2}{n\int \text{tr}(AJ(\theta)) d\pi_n(\theta) + O(r_n^2)}.$$

Our goal now is to get shrinking neighborhoods of θ^*, so that the average error near θ^* is at least $\frac{d^2}{n\,\text{tr}(AJ(\theta^*))} + o(\frac{1}{n})$. So take any rate with $1 \ll r_n \ll \sqrt{n}$ and we do indeed get that:

Corollary 139

Observe X_1^n iid from P_θ, and suppose the single distribution p_θ has Fisher information matrix $J(\theta)$ continuous in θ near θ^*. Then there exist prior distributions supported on $\theta^* + [-\frac{1}{r_n}, \frac{1}{r_n}]^d$ (with $1 \ll r_n \ll \sqrt{n}$) with average squared error

$$\int_{\theta^*+[-\frac{1}{r_n},\frac{1}{r_n}]^d} \mathbb{E}_\theta \left[(\hat{\theta}_n - \theta)^T A^{-1} (\hat{\theta} - \theta)\right] \pi_n(\theta) d\theta \geq \frac{d^2}{n \operatorname{tr}(A(J(\theta^*))} + o\left(\frac{1}{n}\right).$$

Note that if r_n scales exactly as $\sqrt{n}$, our lower bound falls apart and actually we can do better: the neighborhood size does need to be somewhat large. But this gives us almost a per-instance squared error bound, and typically we'll be choosing the matrix A to be the inverse Fisher information matrix so that

$$\mathbb{E}\left[(\hat{\theta}_n - \theta)^T J(\theta^*)(\hat{\theta} - \theta)\right] \geq \frac{d}{n} + o\left(\frac{1}{n}\right)$$

for almost all θ near θ^*. And in fact this is the correct scaling and the correct numerical constant.

Example 140

Consider the setting of linear regression, where we have $Y = X\theta + N(0, \sigma^2 I_n)$ where $X \in \mathbb{R}^{n \times d}$ is some matrix with rows $X_1^T, \cdots, X_n^T$.

Here the Fisher information is constant and completely independent of θ: we have

$$\begin{aligned} J(\theta) &= \mathbb{E}\left[\left(\sum_{i=1}^n (Y_i - X_i^T\theta) X_i\right)\left(\sum_{i=1}^n (Y_i - X_i^T\theta) X_i\right)^T\right] \cdot \frac{1}{\sigma^4} \\ &= \sum_{i=1}^n \mathbb{E}\left[(Y_i - X_i^T\theta)^2 X_i X_i^T\right] \frac{1}{\sigma^4} \\ &= \frac{1}{\sigma^2} X^T X. \end{aligned}$$

because $Y - X\theta$ has mean-zero coordinates. (Indeed, if X gets bigger we get more information on θ, but if we have more noise we get less information.) Therefore by Van Trees,

$$\inf_{\hat{\theta}} \int \mathbb{E}\left[(\hat{\theta} - \theta)^T X^T X (\hat{\theta} - \theta)\right] \pi(\theta) d\theta \geq d\sigma^2 + o(1)$$

as the size of the matrix n grows. And we can achieve this with the least-squares "optimal estimator" $\hat{\theta} = (X^T X)^{-1} X^T Y = \theta + (X^T X)^{-1} X^T \varepsilon$ for $\varepsilon = N(0, \sigma^2 I_n)$.

This idea of saying "a score function has to correlate with a probability distribution to have good performance" still gets used even today, for example proving lower bounds for private estimators and in terms of understanding "what data is memorized by machine learning."

We'll now turn to strong data processing and constrained risk. Often our estimators have constraints (for example communication or privacy or memory use), and we'd like to be able to prove optimality results for these constrained classes. Specifically, the plan is to write out some information-theoretic consequences of these constraints so that we can use them to prove various optimality results.

Definition 141

View processing as a channel $Q : X \to Z$. Given a distribution P on X, let $Q \circ P$ denote the marginal on Z over drawing $X \sim P$ and then $Z|X = x \sim Q(\cdot|x)$ (this is the induced distribution). We say that Q satisfies the **strong data processing inequality (SDPI)** for an f-divergence D_f if for all P_0, P_1,

$$D_f(Q \circ P_0||Q \circ P_1) \leq \alpha D_f(P_0||P_1)$$

for some $\alpha < 1$. Define the SDPI constant for Q

$$\alpha_f(Q) = \sup_{P_0 \neq P_1} \frac{D_f(Q \circ P_0||Q \circ P_1)}{D_f(P_0||P_1)}$$

(where we really should ensure the denominator is nonzero).

(The inequality above always holds for $\alpha = 1$ by ordinary data processing.) This first came up in the theory of mixing of Markov chains – often there we work with total variation distance, and we get a specific name:

Definition 142

The **Dobrushin coefficient** of the channel Q is given by

$$\alpha_{\text{Dob}}(Q) = \sup_{x,y} ||Q(\cdot|x) - Q(\cdot|y)||_{\text{TV}}.$$

Fact 143

For total variation, the worst-case P_0, P_1 in the data processing inequality will be given by point masses, so in fact we actually have

$$\alpha_{\text{Dob}}(Q) = \alpha_{\text{TV}}(Q) = \sup_{P_0 \neq P_1} \frac{||Q \circ P_0 - Q \circ P_1||_{\text{TV}}}{||P_0 - P_1||_{\text{TV}}}.$$

Remark 144. *For any f-divergence, if we have a Markov chain generated via $X_{k+1} \sim Q(\cdot|X_k)$ and we have the strong data processing inequality $\alpha_f(Q) < 1$, we know that this Markov chain mixes geometrically fast:*

$$D_f(Q^k \circ P_0||Q^k \circ P_1) \leq \alpha_f(Q)^k D_f(P_0||P_1).$$

Theorem 145

For any f-divergence (with f convex and $f(1) = 0$), we have $\alpha_{\text{TV}}(Q) \geq \alpha_f(Q)$.

The heuristic proof sketch of this result is the following. We can assume $f \geq 0$ without loss of generality by taking the tilt $g(t) = f(t) - f'(1)(t-1)$ (and we can check that $D_g = D_f$). But now we can approximate convex functions of this form by tangent lines, meaning we can approximate by positive linear combinations of functions of one of the two forms

$$h(t) = a_i(t - c_i)_+ \quad \text{or} \quad h(t) = a_i(-t + c_i)_+$$

for some $c_i \geq 1$ and $a_i \geq 0$. But recall now that the total variation distance satisfies

$$||P_0 - P_1||_{\text{TV}} = \int (p_0 - p_1)_+ = \int \left(\frac{p_0}{p_1} - 1\right)_+ p_1,$$

which corresponds to $h(t) = (t-1)_+$. So if we know that we have an SDPI for this particular h by assumption, then shifts and scalings of this base function will also satisfy strong data processing inequalities with constant at worst α_{TV}, and so if f is approximately some linear combination of them then it will do so as well.

14 November 11, 2025

We'll finish up constrained risk inequalities today, giving an example for private estimation, and then switch gears to the next part of our course, which is **proper losses**.

Last time, we were thinking about channels $Q : \mathcal{X} \to \mathcal{Z}$ satisfying strong data processing inequalities with SDP constants $\alpha_f(Q)$ (which tell us about contraction in the space of probability measures). The point is that processing reduces information and yields stronger minimax bounds, and we'll just give the big-picture ideas here. We have our usual setup for lower bounds where we have some separation

$$|\theta(P_0) - \theta(P_1)| = 2\delta, \quad D_{\mathrm{KL}}(P_0||P_1) \le \kappa^2\delta^2.$$

But now when we process data via Q, we have $D_{\mathrm{KL}}(Q \circ P_0||Q \circ P_1) \le \alpha\kappa^2\delta^2$, so our minimax bounds now look like

$$\inf_{\hat{\theta}} \sup_{P} \mathbb{E}_{Q\circ P}\left[|\hat{\theta}(Z_1^n) - \theta(P)|\right] \ge \delta\left(1 - ||(Q\circ P_0)^n - (Q\circ P_1)^n||_{\mathrm{TV}}\right)$$

(where X is drawn from P_0 and Z comes from passing that through Q), and now using Pinsker and data processing like before yields a lower bound of

$$\delta\left(1 - \sqrt{\frac{n}{2}D_{\mathrm{KL}}(Q\circ P_0||Q\circ P_1)}\right) \ge \delta\left(1 - \sqrt{\frac{n\alpha\kappa^2\delta^2}{2}}\right)$$

so that the minimax bound improves by $\frac{1}{\sqrt{\alpha}}$ over the usual $\frac{1}{\kappa\sqrt{n}}$.

Proposition 146

Suppose we have an ε-locally-differentially private channel, meaning that we have $Q : X \to Z$ with $\frac{Q(A|x)}{Q(A|x')} \le e^\varepsilon$ for all x, x'. Then we have the strong data processing inequality

$$D_{\mathrm{KL}}(M_0||M_1) + D_{\mathrm{KL}}(M_1||M_0) \le 4(e^\varepsilon - 1)^2||P_0 - P_1||_{\mathrm{TV}}^2$$

for any distributions P_0, P_1 on $\mathcal{X}$, where $M_0 = Q \circ P_0, M_1 = Q \circ P_1$ are the marginal distributions.

This is particularly notable because we've actually bounded KL by total variation rather than the other way around, and in particular TV is always bounded by 1.

Proof. Writing out the left-hand side,

$$D_{\mathrm{KL}}(M_0||M_1) + D_{\mathrm{KL}}(M_1||M_0) = \int (m_0(z) - m_1(z)) \log\frac{m_0(z)}{m_1(z)},$$

and now observe that for any scalars a, b we have $(a-b)\log\frac{a}{b} = (a-b)\log(1+\frac{a}{b}-1)$ if $a \ge b$ and $(b-a)\log(1+\frac{b}{a}-1)$ if $b > a$, so we can upper bound logs by $\frac{a}{b} - 1$ and $\frac{b}{a} - 1$ in the two cases, meaning that in all cases

$$(a-b)\log\frac{a}{b} \le \frac{(a-b)^2}{\min(a,b)}.$$

Therefore plugging this back in,

$$D_{\mathrm{KL}}(M_0||M_1) + D_{\mathrm{KL}}(M_1||M_0) \le \int \frac{(m_0(z) - m_1(z))^2}{\min(m_0(z), m_1(z))}.$$

Now define $q^*(z) = \inf_x q(z|x)$; by assumption this is at least $e^{-\varepsilon} q(z|x)$ for any x, and therefore when we expand out the difference in our integrand above we can write

$$\begin{aligned} m_0(z) - m_1(z) &= \int q(z|x)(p_0(x) - p_1(x)) \\ &= \int (q(z|x) - q^*(z))(p_0(x) - p_1(x)) \\ \implies |m_0(z) - m_1(z)| &\le \int |q(z|x) - q^*(z)|\,|p_0(x) - p_1(x)|, \end{aligned}$$

and now $|q(z|x) - q^*(z)| \le q^*(z)(e^\varepsilon - 1)$ so that

$$|m_0(z) - m_1(z)| \le q^*(z)(e^\varepsilon - 1) \int |p_0(x) - p_1(x)| = 2q^*(z)(e^\varepsilon - 1)||P_0 - P_1||_{\mathrm{TV}},$$

so substituting this in and noticing that $m_0(z), m_1(z) \ge q^*(z)$ yields

$$D_{\mathrm{KL}}(M_0||M_1) + D_{\mathrm{KL}}(M_1||M_0) \le 2^2||P_0 - P_1||_{\mathrm{TV}}^2(e^\varepsilon - 1)^2 \int \frac{(q^*(z))^2}{q^*(z)}$$

and the integral is at most 1 so we get our result. □

Of course, we want to tensorize this and do chain rule because we often observe multiple data points, and so we will move now to a slightly more elaborate setting.

Theorem 147

Suppose we have samples of size n now where each Z_i depends on an independent X_i, but also on the previous Zs. (We can think of this as saying that "based on survey answers from the first half of a population, we change the surveys for the second half", or having interaction through a stochastic gradient algorithm; this is called being **sequentially interactive**) Suppose that each $Q : X_i \to Z_i$ is ε-differentially-private, meaning that $\frac{Q(A|x_i, z_1^{i-1})}{Q(A|x_i', z_1^{i-1})} \le e^\varepsilon$ for all x_i, x_i', z_1^{i-1}. Let M_0^n and M_1^n be the marginal distributions over Z_1^n when X_1^n are iid drawn from P_0 or P_1. Then

$$D_{\mathrm{KL}}(M_0^n||M_1^n) \le 4n(e^\varepsilon - 1)^2||P_0 - P_1||_{\mathrm{TV}}^2.$$

Remember that M_0^n and M_1^n are not product distributions, but we still get a nice "tensorization-type" inequality (though we can only get a one-sided KL bound here). And the channels Q can be different and we can replace iid from P_v with product distributions $P_{v,i}$, though there don't really seem to be applications of that generality.

Proof. Suppose $M_{0,i}(\cdot|z_1^{i-1})$ and $M_{1,i}(\cdot|Z_1^{i-1})$ are the induced marginals on Z_i under X drawn from either P_0 or P_1, respectively. By chain rule for KL-divergence,

$$D_{\mathrm{KL}}(M_0^n||M_1^n) = \sum_{i=1}^n \mathbb{E}_{M_0^{i-1}}\left[D_{\mathrm{KL}}(M_{0,i}(\cdot|Z_1^{i-1})||M_{1,i}(\cdot|Z_1^{i-1}))\right]$$

with Z_1^{i-1} drawn from the first distribution M_0^{i-1}. But now we can apply our original proposition to each term inside the expectation, since we have marginals over a single channel use: for either index $v \in \{0, 1\}$ we have by definition

of conditioning that

$$M_{v,i}(A|z_1^{i-1}) = \int Q(A|x_i, x_{\setminus i}, z_1^{i-1}) dP_v(x_i, x_{\setminus i}|z_1^{i-1})$$
$$= \int Q(A|x_i, z_1^{i-1}) dP_v(x_i|z_1^{i-1})$$

since conditioned on the previous Zs and X_i, Z_i does not depend on the remaining $X_{\setminus i}$ and so we can remove the conditioning in Q and then marginalize over those variables afterward. And now X_i don't depend on the Z_1^{i-1} (they're iid from our base distribution) so this is in fact equal to $\int Q(A|x_i, z_1^{i-1}) dP_v(x_i)$ and thus we have the standard marginal. So now applying Proposition 146 yields

$$D_{\mathrm{KL}}(M_0^n||M_1^n) \leq \sum_{i=1}^n \mathbb{E}_{M_0^{i-1}}\left[4(e^\varepsilon - 1)^2 ||P_0 - P_1||_{\mathrm{TV}}^2\right]$$

and now the expectation doesn't matter and we sum over i to get the result. □

Example 148 (Optimality of randomized response)

Suppose $X_i \in \{0, 1\}$ are sensitive and iid from $\mathrm{Ber}(\theta)$ (these are "answers we don't want to give away"). Recall that if we set $Z_i|X = x$ to be x with probability $q = \frac{e^\varepsilon}{1+e^\varepsilon}$ and $1-x$ with probability $1-q$, then we can successfully estimate X via the unbiased estimator

$$\hat{X} = \frac{1}{2q-1}(Z - (1-q))$$

and therefore can define $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \hat{X}_i$ and get squared error $\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[((\hat{\theta} - \overline{X}_n) + (\overline{X}_n + \theta))^2]$. We can then break this up because the two halves of this are conditionally independent, and we get that

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \asymp \frac{1}{(e^\varepsilon - 1)^2 n} + \frac{\theta(1-\theta)}{n},$$

and we claim this is actually sharp.

To get a lower bound, fix some $\theta \in [0, \frac{1}{2}]$ and let $\delta \geq 0$ be chosen later. Let P_0 be $\mathrm{Ber}(\theta)$ and P_1 be $\mathrm{Ber}(\theta + \delta)$; we have $|\theta(P_0) - \theta(P_1)| = \delta$ and also $||P_0 - P_1||_{\mathrm{TV}} = \delta$, so by our general Le Cam method we get

$$\inf_{\hat{\theta}} \sup_P \mathbb{E}\left[(\hat{\theta}(Z_1^n) - \theta)^2\right] \gtrsim \delta^2 \left(1 - ||M_0^n - M_1^n||_{\mathrm{TV}}\right),$$

and by Pinsker and the theorem we just proved, this is at least

$$\delta^2\left(1 - \sqrt{\frac{1}{2}D_{\mathrm{KL}}(M_0^n||M_1^n)}\right) \geq \delta^2\left(1 - \sqrt{2n(e^\varepsilon - 1)^2 ||P_0 - P_1||_{\mathrm{TV}}^2}\right) = \delta^2(1 - \sqrt{2n(e^\varepsilon - 1)^2\delta^2})$$

and choosing $\delta^2 = \frac{1}{8n(e^\varepsilon - 1)^2}$ proves the lower bound that we want.

To unpack this, notice that something funny is happening compared to estimating the mean of a Bernoulli. The expected squared error between sample mean and θ is $\frac{\theta(1-\theta)}{n}$, which is zero when $\theta = 0$ so we "adapt to the variance for free." But under privacy, no matter what our initial parameter was, we cannot adapt to "problem difficulty" and need to use the worst-case scenario because we've moved from KL to total variation.

Example 149

We'll now switch over to our new topic of **proper losses**; the idea is that we want to measure performance of estimators or predictions and there are often many connections to information-theoretic ideas. The motivating problem to keep in mind is to predict probabilities or probability distributions, incentivizing playing correctly. This comes up primarily in weather prediction or forecasting (predicting distribution of rainfall or other climatological outcomes) or finance (predicting distributions of prices of assets tomorrow so we can make decisions related to risk).

The setting will be the following: a player "plays" a distribution Q, and nature reveals an outcome Y, which makes the player suffer some loss $\ell(Q, Y)$ on the realization of what happened.

Definition 150

A loss is **proper** if $\mathbb{E}_P[\ell(P, Y)] \leq \mathbb{E}_P[\ell(Q, Y)]$ for any distributions P, Q, and it is **strictly proper** if the inequality is strict for $Q \neq P$.

Example 151

Assume our distributions have densities (this is a bit of a strong assumption in cases like rainfall where there's a large point mass at zero, but let's ignore that for now). Then the **log loss** $\ell(Q, y) = -\log q(y)$ satisfies $\mathbb{E}_P[\ell(Q, Y)] - \mathbb{E}_P[\ell(P, Y)]$ is exactly the KL-divergence $D_{\mathrm{KL}}(P||Q)$, so log loss is a strictly proper loss.

Example 152

Similarly (common in classification problems) the **zero-one loss for classification** $\ell(Q, y) = 1\{q(y) \leq \max_{j \neq y} q(j)\}$ (which gives a loss if y is not the largest assigned probability) is proper but not strictly proper, since

$$\mathbb{E}_P[\ell(Q, Y)] - \mathbb{E}_P[\ell(P, Y)] = \max_y P(Y = y) - P(Y = \mathrm{argmax}_y q(y))$$

is nonnegative but if we make Q just a point mass at the most likely outcome it will also achieve zero loss.

The end goal will be to give a characterization of all proper losses, and to do so we need some convex analysis (the book does full proofs from full principles if we want more details).

Definition 153

A set C is **convex** if it contains all lines between points in the set, meaning that $x, y \in C$ implies $\lambda x + (1-\lambda)y \in C$ for all $\lambda \in [0, 1]$. A function $f : \mathbb{R}^n \to \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ is **convex** if the domain $\{x : f(x) < \infty\}$ is convex and $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for all $\lambda \in [0, 1]$ and all x, y (where if x, y are outside the domain the inequality is trivial); we say it is **strictly convex** if this inequality is strict for all $\lambda \in (0, 1)$ and $x \neq y$.

There's a connection between convex functions and convex sets which is foundationally important to convex analysis:

Definition 154

The **epigraph** of a function is the set of points above the graph of the function

$$\mathrm{epi}\ f = \{(x, t) : f(x) \leq t\}.$$

We can see from the definition that f is convex if and only if epi f is convex.

Definition 155

A vector s is a **subgradient** of f at a point x if

$$f(y) \geq f(x) + \langle s, y - x \rangle \quad \forall y \in \mathbb{R}^n.$$

The **subdifferential** is the collection of all subgradients at a certain point:

$$\partial f(x) = \{s : f(y) \geq f(x) + \langle s, y - x \rangle \text{ for all } y\}.$$

Theorem 156

If x is in the interior of the domain of f, then $\partial f(x)$ is nonempty and is a compact convex set. Furthermore, x minimizes f if and only if $0 \in \partial f(x)$.

The picture to have is that points in the subdifferential let us globally underestimate the function (or equivalently this means the vector $(s, -1)$ supports the epigraph of f, meaning it points straight out from the set); at a smooth point this is just the ordinary gradient, but at a kink we may have a variety of points. For example, the subdifferential at 0 for the absolute value function $f(x) = |x|$ is the full interval $[-1, 1]$.

Fact 157

For any arbitrary collection of convex functions $\{f_\alpha\}_{\alpha \in \mathcal{A}}$, the supremum $f(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$ is convex, and the subdifferential of f satisfies

$$\partial f(x) \supseteq \text{Conv}\{\partial f_\alpha(x) : f_\alpha(x) = f(x)\}$$

where Conv denotes the convex hull. The picture is that if we have two convex functions f_1, f_2, then their maximum may have some kinks where f_1 and f_2 intersect and we must have lots of different possible subdifferentials there.

Theorem 158

If f is a **closed** convex function (meaning that epi f is a closed set), then $f(x) = \sup\{g(x) : g \text{ affine function}, g \leq f\}$.

In other words, f is the maximum of supporting lines underneath it. The key object we'll play with for understanding optimality is the following:

Definition 159

The **convex conjugate** (also called the **Legendre-Fenchel transform**) of a function f is

$$f^*(s) = \sup_x \{\langle s, x \rangle - f(x)\}.$$

This is a closed convex function because it is the supremum of affine functions of s, and the picture is that we draw a line through the origin at slope s and look at where we have the biggest gap to f. The gap we get is then $f^*(s)$. Now by subtracting off that gap, notice that at the point x where this occurs, we have the linear function $g(x) = \langle s, x \rangle - f^*(s)$ which is a supporting line for the original function f. Indeed, we have that

$$g(x) = \langle s, x \rangle - f^*(s) \leq \langle s, x \rangle - (\langle s, x \rangle - f(x)) = f(x),$$

and we actually get the following:

Theorem 160 (Fenchel-Young inequality)
We always have $f(x) + f^*(s) \geq \langle s, x\rangle$, and equality holds if and only if $x \in \partial f^*(s)$, which holds if and only if $s \in \partial f(x)$.

Theorem 161 (Conjugate duality)
If f is a closed convex function, meaning that its epigraph is a closed set, then $f(x) = f^{**}(x)$.

Indeed, the biconjugate is like looking at all possible functions g and taking the biggest one, so that just gets us back our function f again. And the other relevant piece of intuition here is that if C is a convex closed set, we can write C as the intersection of all halfspaces H containing C (so in particular, the epigraph is the intersection of all half-spaces traced out by linear functions).

15 November 13, 2025

We started talking about proper losses last time; we'll characterize them today and establish some dualities (with connections to generalized notions of entropy), and then we'll move into calibration of predictors. Recall that we have a "forecasting game" in which we play a distribution Q, nature instead draws from $Y \sim P$, and we suffer a loss $\ell(Q, Y)$. We want to understand the losses where this is optimized when we actually pick $P = Q$.

We will sometimes use the notation $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ or $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$; sometimes we will allow our losses to take on the value $+\infty$ (for things like the log loss with zero probability). Let's now describe a characterization of proper losses, starting with the multinomial / categorical case:

Theorem 162 (Gneiting–Raftery–Savage, special case)
Suppose $Y \in \{1, \cdots, k\}$, so we can identify probability distributions with vectors $p \in \Delta_k$ in the probability simplex $\Delta_k = \{v \in \mathbb{R}^k_{\geq 0} : \langle v, 1\rangle = 1\}$. A loss $\ell : \Delta_k \times [k] \to \overline{\mathbb{R}}$ is (strictly) proper if and only if there exists some (strictly) convex $\Omega : \Delta_k \to \mathbb{R}$ such that for some $\Omega'(q) \in \partial\Omega(q)$, we have $\ell(q, y) = -\Omega(q) - \langle \Omega'(q), e_y - q\rangle$, where e_y is the yth standard basis vector for $1 \leq y \leq k$.

We will call Ω the **negative generalized entropy** associated to the loss. Implicit in the theorem statement is that we can construct this function Ω, and it is in fact subdifferentiable on its domain (so there exists some $\Omega'(q)$).

Proof. For the backward direction (which is easier), suppose we can write our loss in this form. Then the gap between the loss at p and at q is (our expectations are taken for the random variable Y)

$$\mathbb{E}_P[\ell(q, Y) - \ell(p, Y)] = \Omega(p) - \Omega(q) - \langle \Omega'(q), \mathbb{E}_p[e_y] - q\rangle + \langle \Omega'(p), \mathbb{E}_p[e_y] - p\rangle,$$

and now $\mathbb{E}_p[e_y] = \sum_y p(y) e_y$ is exactly the distribution p, so that

$$\mathbb{E}_P[\ell(q, Y) - \ell(p, Y)] = \Omega(p) - \Omega(q) - \langle \Omega'(q), p - q\rangle,$$

and this is exactly the Bregman divergence $D_\Omega(p, q)$, which is nonnegative. (Indeed, this is how much the function is above the first-order approximation coming from $\Omega'(p)$.) And if this function is strictly convex, then in fact D_Ω is strictly positive for all $p \neq q$.

For the other direction, we can do a direct construction using propriety (properness) of the loss. Define the generalized entropy

$$H_\ell(P) = \inf_Q \mathbb{E}_P\left[\ell(Q, Y)\right].$$

As a function of P, $\mathbb{E}_P[\ell(Q, Y)]$ is linear for each fixed Q (since it is an expectation), so the infimum of these quantities is a concave function. Thus we can define $\Omega(P) = -H_\ell(P)$, and in fact by propriety we know that we must actually have $\Omega(P) = -\mathbb{E}_P[\ell(P, Y)] = \sum_j p_j \ell(p, j)$. But now

$$\begin{aligned}\ell(p, y) &= -\Omega(p) + \Omega(p) + \ell(p, y) \\ &= -\Omega(p) - \sum_j p_j \ell(p, j) + \ell(p, y) \\ &= -\Omega(p) + \langle \ell(p, j)_{j=1}^k, e_y - p\rangle.\end{aligned}$$

But now writing out the definitions, $-\ell(p, j)_{j=1}^k$ is indeed a subdifferential, because properity implies that $-\Omega(p) \leq \mathbb{E}_P[\ell(q, Y)] = -\Omega(q) + \langle \ell(q, j)_{j=1}^k, p - q\rangle$, and rearranging says that $\Omega(p) \geq \Omega(q) + \langle (-\ell(q, j))_{j=1}^k, p - q\rangle$, which is what we want. And all inequalities become strict if we have strict propriety, which is the same as saying we have strict convexity. □

Example 163

For the log loss $\ell(q, y) = -\log q(y)$ (everything has a probability mass function here so we don't have to worry about technical details), we already saw that our loss is proper and

$$H_\ell(p) = -\sum p_j \log p_j$$

is the usual Shannon entropy, so $\Omega(p) = \sum_j p_j \log p_j$ on the probability simplex. Then $\nabla\Omega(p) = (1 + \log p_j)_{j=1}^k$, and we can check that indeed $\mathbb{E}_p[\ell(q, Y) - \ell(p, Y)] = D_\Omega(P, q) = D_{\mathrm{KL}}(P||q)$, as we did last time.

Example 164

For the **Brier score** (that is, the squared error), we have $\ell(q, y) = \frac{1}{2}||q - e_y||_2^2 = \frac{1}{2}||q||_2^2 - \langle q, e_y\rangle + \frac{1}{2}$. Then

$$\mathbb{E}_p[\ell(q, Y)] = \frac{1}{2}||q||_2^2 - \langle q, p\rangle + \frac{1}{2};$$

this is minimized at $q = p$ by taking a derivative and so

$$H_\ell(p) = \inf_q \mathbb{E}_p[\ell(q, Y)] = \frac{1}{2} - \frac{1}{2}||p||_2^2,$$

which is concave. So $\Omega(p) = \frac{1}{2}||p||_2^2 - \frac{1}{2}$ and $D_\Omega(p, q) = \frac{1}{2}||p - q||_2^2$; this is indeed a strictly proper loss.

We can now do the general statement:

Theorem 165 (Gneiting–Raftery–Savage, general statement)

Let $\mathcal{P}$ be a family of distributions on $\mathcal{Y}$, and consider functions of the form $\Omega : \mathcal{P} \to \mathbb{R}$ and $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$. We will index elements of $\partial\Omega(Q)$ with functions of y which we think of as vectors (doing a bit of abuse of notation) and write $\Omega'(Q) = [\Omega'(P, y)]_{y \in \mathcal{Y}}$. Define the inner product of such a function with a distribution as

$$\langle \Omega'(Q), P \rangle = \mathbb{E}_P[\Omega'(Q, y)] = \int \Omega'(Q, y) dP(y).$$

Let 1_y denote the point mass at y. Then $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is (strictly) proper if and only if there exists a (strictly) convex $\Omega : \mathcal{P} \to \mathbb{R}$ with

$$\ell(P, y) = -\Omega(P) - \langle \Omega'(P), 1_Y - P \rangle$$

for some $\Omega'(P)$ is the subdifferential at P.

Proof. The proof is exactly the same as before; we define the entropy $H_\ell(P)$ in the same way and we still have $\Omega(P) = -\mathbb{E}_P[\ell(p, Y)]$. The only changes are that instead of $\sum_j p_j \ell(p, j)$ we must write $\int \ell(P, y) dP_y$, and instead of $(\ell(p, j))_{j=1}^k$ and e_y we have the infinite-dimensional vector $(\ell(p, y))_{y \in \mathcal{Y}}$ and the point mass 1_y. □

The way to think about $\Omega'(P)$ is as an "influence function" of y on our loss: if it exists, then we can write

$$\Omega'(P, y) = \lim_{t \downarrow 0} \frac{\Omega((1-t)P + t1_y) - \Omega(P)}{t}$$

in basically all reasonable cases.

Example 166

In the **continuous ranked probability score** (CRPS), assume we are playing cumulative distribution functions ($F : \mathbb{R} \to [0, 1]$ which are nondecreasing with $\lim_{t \to -\infty} F(t) = 0$ and $\lim_{t \to \infty} F(t) = 1$). For example, if we are predicting rainfall, the cdf will be 0 to $t < 0$, and then it will jump to some positive value at 0 and then increase up to 1. We then define

$$\ell_{\text{CRPS}}(F, y) = -\int_{-\infty}^{\infty} (F(t) - 1(t \le y))^2 dt$$

(this is the difference between F and the cdf coming from the point mass at y).

Suppose the true CDF is F and we instead play G. The gap between the expected scores (expectation with respect to F) is

$$\begin{aligned}
\mathbb{E}\left[\ell_{\text{CRPS}}(G, y) - \ell_{\text{CRPS}}(F, y)\right] &= \int \left(G(t)^2 - F(t)^2 - 2(G(t) - F(t))\mathbb{E}[1(Y \le t)]\right) dt \\
&= \int \left(G(t)^2 - F(t)^2 - 2(G(t) - F(t))F(t)\right) dt \\
&= \int (F(t) - G(t))^2 dt,
\end{aligned}$$

so this is exactly the squared L^2 distance $||F - G||_2^2$. This is zero if and only if the cdfs are the same by right-continuity, so we get a strictly proper loss. And we can explicitly compute $\Omega'(F, y)$ if we want it as well.

We'll now use these generalized entropies to get us loss functions and back, and the idea is that our representation $\ell(p, y) = -\Omega(p) - \langle \nabla\Omega(p), e_y - p \rangle$ is not easy to optimize because the inner product term is "whackadoodle" and the first term is actually concave instead of convex. So we want to use a surrogate function $\phi(s, y)$ instead which is

actually convex in those "predictive scores" s and hopefully is equivalent to ℓ, and we'll do this using convex duality. For simplicity, we will focus on the finite case where $\mathcal{Y} = [k]$.

For any convex function we have a conjugate function

$$\Omega^*(s) = \sup_{p \in \Delta_k} \{\langle s, p\rangle - \Omega(p)\}$$

(in words, pick a slope s and see what the maximum $\langle p, s\rangle$ is above $\Omega(p)$) which is convex closed. Remembering that Ω is also closed convex (because it was defined as a certain infimum), we must actually have

$$\Omega(p) = \sup_{s \in \mathbb{R}^k} \{\langle s, p\rangle - \Omega^*(s)\}$$

(in words, this is saying that we take all of the global underestimators and pull them up to recover the original function). With all of this, we can now define the **surrogate loss** (replacing ℓ)

$$\phi(s, y) = -\langle s, e_y\rangle + \Omega^*(s).$$

This is convex in s, and we can compute the entropy associated with it:

$$\inf_s \mathbb{E}_P[\phi(s, y)] = \inf_{s \in \mathbb{R}^k} \{-\langle s, p\rangle + \Omega^*(s)\} = -\Omega(p).$$

So given any loss function, we can construct the entropy associated with it; from that entropy we can construct a convex surrogate which generates the same entropy (meaning "the uncertainty is the same at optimality"). And the point is that optimizing ϕ is the same as optimizing ℓ because of the following:

Fact 167 (Fenchel duality)
Suppose Ω, Ω^* are differentiable. Then

$$s = \nabla\Omega(p) \iff p = \nabla\Omega^*(s) \iff \langle s, p\rangle = \Omega(p) + \Omega^*(s).$$

This is true more generally and also gives us a way of defining differentials in general, but the intuition is that if slopes are equal then we can go back and forth between the two functions.

So if we define the predicted probability distribution

$$\text{pred}_\Omega(s) = \text{argmax}_p \{\langle s, p\rangle - \Omega(p)\} = \nabla\Omega^*(s),$$

then the associated loss is

$$\begin{aligned}\ell(\text{pred}_\Omega(s), y) &= -\Omega(\text{pred}(s)) - \langle \nabla\Omega(\text{prod}(s)), e_y - \text{pred}(s)\rangle \\ &= \langle -\Omega(\nabla\Omega^*(s)) - \langle s, e_y - \nabla\Omega^*(s)\rangle\end{aligned}$$

by just substituting in definitions. But now $p = \nabla\Omega^*(s)$ and $\langle s, p\rangle = \Omega(p) + \Omega(s^*)$, so we get that in fact we have exactly

$$\ell(\text{pred}_\Omega(s), y) = \Omega^*(s) - \langle s, p\rangle - \langle s, e_y - p\rangle = -\langle s, e_y\rangle + \Omega^*(s) = \phi(s, y).$$

This therefore tells us the following:

Theorem 168

Suppose Ω, Ω^* are differentiable. Then associated to this prediction function, we have

$$\ell(\text{pred}_\Omega(s), y) = \phi(s, y),$$

and furthermore if we define $p_n = \text{pred}_\Omega(s_n)$, then $\mathbb{E}_p[\phi(s_n, y)] \to \inf_s \mathbb{E}[\phi(s, y)]]$ if and only if $\mathbb{E}_p[\ell(p_n, y)] \to \inf_q \mathbb{E}[\ell(q, Y)]$.

So we can minimize our convex thing in s-space if and only if we minimize losses in the probability space.

Example 169

Consider the standard logistic loss $\ell(p, y) = -\log p(y)$. We know that $H(p) = -\sum_j p_j \log p_j$ and the negative entropy is $\Omega(p) = \sum p_j \log p_j$. The conjugate of this function is

$$\Omega^*(s) = \sup_{p : p^T 1 = 1} \left\{ \langle s, p \rangle - \sum_j p_j \log p_j \right\}$$

and we can compute this by Lagrange multipliers.

Defining $\mathcal{L}(p, \lambda) = \langle s, p \rangle - \sum_j p_j \log p_j + \lambda(1^T p - 1)$, we find that $\nabla_p \mathcal{L}(p, \lambda) = s - (1 + \log p_j)_{j=1}^k + \lambda 1$, which means that $p_j \propto \exp(s_j)$ and the constant is chosen so the sum is 1: $p(s) = \left(\frac{e^{s_y}}{\sum_{j=1}^k e^{s_j}} \right)_{y=1}^k$, and we find that

$$\Omega^*(s) = \langle s, p(s) \rangle - \Omega(p(s)) = \log \left(\sum_{y=1}^k e^{s_y} \right)$$

by some algebra. Thus we get the surrogate function

$$\phi(s, y) = -s_y + \log \left(\sum_{j=1}^k e^{s_j} \right) = \log \left(1 + \sum_{j \neq y} e^{s_j - s_y} \right).$$

So if we start with the usual Shannon entropy to define uncertainty of a distribution, the standard loss is the log loss, and the surrogate function is the cross-entropy or the multi-class logistic loss.

16 November 18, 2025

Our topic for today is **calibration**, which is still kind of an active research area. We'll first give some definitions and connections to proper losses, and then we'll understand how to measure it and discuss impossibility and alternatives.

Example 170

The setting we have is the following: we are looking to predict some kind of vector $Y \in \mathbb{R}^k$ from some data $X \in \mathcal{X}$. The two examples to keep in mind are **binary classification**, where $Y \in \{0, 1\} \subset \mathbb{R}$, or **multi-class classification**, where we think of $Y \in \{e_y\}_{y=1}^k \subset \mathbb{R}^k$ as the set of basis vectors. We can think of $f : \mathcal{X} \to \text{Conv}(\mathcal{Y})$ as the possible predictors (that is, f maps into $\{\mathbb{E}_P[Y] : P \text{ a distribution on } Y\}$).

For example in the multi-class classification case, the different components of the k-dimensional vector are the probability mass we assign to each of the k classes

Definition 171
A predictor f is **calibrated** if $f(X) = \mathbb{E}[Y|f(X)]$.

For example, if Y is valued in $\{0, 1\}$ (with 1 indicating rain and 0 indicating no rain), then $f(X)$ is the estimated probability of rain given X. And our predictor being calibrated says that "when we think there is a 70 percent change of rain, it rains 70 percent of the time." And so the space $\mathcal{X}$ doesn't really matter, only the output probabilities. (And in the real world, this does really matter in the context of things like LLMs – it would be great to be able to estimate the probability that these models are spitting out something actually true.)

The bare-bones goals of calibration (which is where research is still at) are to calibrate predictors and measure them. Let's start by making some connections with proper losses – to set some notation, let $\mathcal{M}$ be the convex hull of $\mathcal{Y}$ (so the space of means $Y \in \mathbb{R}^k$). Recall that a loss ℓ is proper if and only if (this is the GRS representation) for some convex function $\Omega : \mathcal{M} \to \overline{\mathbb{R}}$, we have $\ell(\mu, y) = -\Omega(\mu) - \langle \nabla\Omega(\mu), y - \mu\rangle$. We claim that at a high level, **propriety encourages calibration** (so minimizing a proper loss means we are calibrated), and we can always break an expected loss into a prediction error plus calibration error (which is similar to a bias-variance decomposition).

Let $S = f(X)$ be the score (predicted value) of the prediction, which is some random variable. Then we can decompose

$$\mathbb{E}[\ell(S, Y)|S] = \mathbb{E}\left[\ell(\mathbb{E}[Y|S], Y)|S\right] + \mathbb{E}\left[\ell(S, Y) - \ell(\mathbb{E}[Y|S], Y)|S\right],$$

and now the conditional expectation of the blue term is (substituting in the GRS representation above and noting that one of the gradient terms disappears) $\Omega(\mathbb{E}[Y|S]) - \Omega(S) - \langle \nabla\Omega(S), \mathbb{E}[Y|S] - S\rangle$, which is exactly the Bregman divergence $D_\Omega(\mathbb{E}[Y|S], S)$. Thus we get the following:

Theorem 172
If ℓ is a proper loss, then for any predictor f, we can decompose

$$\mathbb{E}[\ell(f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y|f(X)], Y] + \mathbb{E}\left[D_\Omega(\mathbb{E}[Y|f(X)], f(X)\right].$$

Here the first term is the prediction error, and the second is the calibration error. Furthermore, if we define $g(s) = \mathbb{E}[Y|f(X) = s]$ as a "recalibration," then

$$\mathbb{E}[\ell(g \circ f(X), Y)] \leq \mathbb{E}[\ell(f(X), Y)],$$

with strict inequality if the loss is proper and f is not perfectly calibrated.

Indeed, ℓ is strictly proper if and only if Ω is strictly convex, meaning the Bregman divergence term is strictly positive as long as its arguments are not the same. And we'll see how we can use this to improve predictors later today; first we'll take a detour towards understanding how to measure calibration (and determine whether a function f is calibrated for a particular application).

One candidate idea (which turns out to be a problem) is to calculate the **expected calibration error**

$$\text{ece}(f) = \mathbb{E}\left[||\mathbb{E}[Y|f(X)] - f(X)||\right].$$

The issue is that this actually isn't a friendly quantity, since it's provably impossible to estimate, and it's also discontinuous. Discontinuity is actually pretty easy to show:

Example 173

Suppose $Y = X \in \{0, 1\}$ and X is uniform (so probability 0.5 of being either one). Now consider $f_0(x) = \frac{1}{2}$; this is calibrated and the expected calibration error is 0, since conditioning on $f(X)$ doesn't tell us anything. But now for any $\varepsilon > 0$ we can define

$$f_\varepsilon(x) = \begin{cases} \frac{1}{2} - \varepsilon & \text{if } x = 0, \\ \frac{1}{2} + \varepsilon & \text{if } x = 1, \end{cases}$$

and now $\text{ece}(f_\varepsilon) = \frac{1}{2} - \varepsilon$ by a direct calculation. But on any reasonable metric, f_ε approaches f_0, and we're saying that $\text{ece}(f_0) = 0$ is perfectly calibrated but $\text{ece}(f_\varepsilon) \to \frac{1}{2}$ is always bad.

We can also formalize what we mean about the other statement:

Theorem 174

The expected calibration error is impossible to estimate. Let $f : \mathcal{X} \to [0, 1]$ and suppose $f(X)$ contains any neighborhood $[\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon]$ (any open set works). Consider the hypothesis test $H_0 : \text{ece}(f) = 0$ versus $H_\gamma : \text{ece}(f) \geq \frac{1}{2} - \gamma$ for any $\gamma > 0$. Then

$$\liminf_n \inf_{\psi_n \text{ with } n \text{ data points}} \sup_{P_0 : H_0} \sup_{P_1 : H_\gamma} \{P_0(\psi_n \neq 0) + P_1(\psi \neq 1)\} = 1.$$

So we have a hypothesis test of being perfectly calibrated versus being almost as bad as just flipping a coin, and we can't actually distinguish them better than random guessing. This is kind of related to saying how if we wanted to estimate the variance of a function or the number of modes in a probability density, and we can give lower bounds but never provide guaranteed upper bounds.

We won't prove this result (we can look in the book), but instead what we can do is prove what we are able to do. We'll develop a theory of **sound and complete calibration measures** which can actually be estimated. Given a function f and some collection of observations $(X_i, Y_i)_{i=1}^n$, we want to measure how calibrated f is and whether it can be useful for decision-making tasks. (Intuitively, think of this as "we take actions based on the weather forecast" such as "bring an umbrella if the forecast has at least a 20 percent chance of rain.")

Definition 175

A calibration criterion / benchmark $M : \mathcal{F} \to \mathbb{R}_+$ is **sound** if $M(f) = 0$ implies $\mathbb{E}[Y | f(X)] = f(X)$ (so if we think it is calibrated, it actually is calibrated) and **complete** if $\mathbb{E}[Y | f(X)] = f(X)$ implies $M(f) = 0$.

Note that ece does satisfy these conditions, but it's just discontinuous and awful and doesn't have a good reflection of the space of underlying functions. So we want these conditions plus some nice continuity and estimability guarantees.

The key idea is that we will try to **witness miscalibration**: if we imagine making a plot of the value we're predicting $f(X) \in [0, 1]$ on the x-axis and the frequency with which $Y = 1$ given $f(X)$ on the y-axis, then perfect calibration would be the line $y = x$. If we could magically plot this frequency, then we'd be interested in regions where the true frequency is far away from the line $y = x$; if we then make our witness function w large in that region, then we will notice that

$$\mathbb{E}[w(f(X))(Y - f(X))] \text{ is large.}$$

So this gives us a recipe for calibration error:

Definition 176

Let $\mathcal{W}$ be a **symmetric** collection of functions $\mathbb{R}^k \to \mathbb{R}^k$, meaning that if $w \in \mathcal{W}$, then $-w \in \mathcal{W}$ as well. The **calibration error relative to** $\mathcal{W}$ (also sometimes called **weak calibration** or **smooth calibration**) is

$$\mathrm{CE}(f, \mathcal{W}) = \sup_{w \in \mathcal{W}} \mathbb{E}\left[\langle w(f(X)), Y - f(X)\rangle\right].$$

Notice in particular that

$$\mathbb{E}\left[\langle w(f(X)), Y - f(X)\rangle\right] = \mathbb{E}\left[\langle w(f(X)), \mathbb{E}[Y|f(X)] - f(X)\rangle\right]$$

so this is basically witnessing places where the conditional expectation is wrong.

If we choose $\mathcal{W}$ to be too large, then we won't be able to estimate it and things aren't great. For example, if $\mathcal{W}$ is the class of all functions with $||w(s)||_2 \leq 1$, then we just recover the expected calibration error again, since we can just choose w to point in the direction of $\mathbb{E}[Y|f(X)] - f(X)$, which exactly gives us back $\mathrm{ece}(f)$. But there are better choices:

Example 177

For picking binary probabilities, let $\mathcal{W}$ be the set of Lipschitz functions $w : [0,1] \to [-1,1]$. If we were given a sample of n observations (X_i, Y_i), the naive estimator of the calibration error $\mathrm{CE}(f, \mathcal{W})$ would be the empirical average

$$\widehat{\mathrm{CE}}_n(f, W) = \sup_{w \in \mathcal{W}} P_n\langle w(f(X)), Y - f(X)\rangle = \frac{1}{n}\langle w(f(X_i)), Y_i - f(X_i)\rangle.$$

The point is that for Lipschitz functions, we can actually do this and get convergence rates.

Indeed, we have

$$\widehat{\mathrm{CE}}_n(f, W_{\mathrm{Lip}}) = \sup_{||w||_\infty = 1} \left\{ \frac{1}{n}\sum_{i=1}^n w_i(Y_i - f(X_i)) : |w_i - w_j| \leq |f(x_i) - f(x_j)| \right\},$$

which we can solve because we have an n-dimensional linear program with n^2 constraints. (Of course we could have done this with a larger class of $\mathcal{W}$ instead like bounded functions, but the point is that we wouldn't get nice convergence to the calibration error.)

Proposition 178

There is some universal constant $C < \infty$ such that

$$\left|\widehat{\mathrm{CE}}_n(f, \mathcal{W}_{\mathrm{Lip}}) - \mathrm{CE}(f, \mathcal{W}_{\mathrm{Lip}})\right| \leq C\left(\frac{1}{n^{1/3}} + \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$$

with probability at least $1 - \delta$.

Proof. We'll do a covering number argument. Cover the class of Lipschitz functions in the ℓ_∞ norm; we claim that

$$N(\mathcal{W}_{\mathrm{Lip}}, ||\cdot||_\infty, \varepsilon) \lesssim \frac{1}{\varepsilon} \cdot 3^{1/\varepsilon}.$$

Indeed, break up the horizontal interval $[0,1]$ and vertical interval $[-1,1]$ into bins of width ε. Over each horizontal interval, consider the piecewise linear changes where we can either change our function by $+\varepsilon, 0$, or $-\varepsilon$, and we can

convince ourself that any Lipschitz function will stay close to one of these piecewise linear functions in ℓ^∞ norm. And there are $3^{1/\varepsilon}$ possible branches and $\frac{2}{\varepsilon}$ possible starting points.

Thus for any ε, we have some collection of functions $\mathcal{N}(\varepsilon) = \{w^i\}_{i=1}^{N(\varepsilon)}$ so that for any $w \in \mathcal{W}_{\text{Lip}}$ we have $\sup_t |w(t) - w^i(t)| \le \varepsilon$ for some w^i. Let $E_i = Y_i - f(X_i)$ be the error on the ith prediction, and use the shorthand $\langle w, E\rangle_n = \frac{1}{n}\sum_{i=1}^n w_i E_i$. Then

$$|\widehat{\text{CE}}_n(f) - CE(f)| \le \max_{w\in\mathcal{N}(\varepsilon)} |\langle w, E\rangle_n - \mathbb{E}[\langle w, E\rangle_n]| + 2\varepsilon,$$

but now $|\langle w, E\rangle_n - \mathbb{E}[\langle w, E\rangle_n]|$ satisfy $\frac{2}{n}$-bounded differences (because they're sums of independent things whose differences are between -1 and 1), so

$$\begin{aligned}\mathbb{P}(|\widehat{\text{CE}}_n(f) - CE(f)| > t) &< \mathbb{P}\left(\max_{w\in\mathcal{N}(\varepsilon)} |\langle w, E\rangle_n - \mathbb{E}[\langle w, E\rangle_n]| \ge t - 2\varepsilon\right)\\ &\le 2N(\varepsilon)\exp\left(-\frac{n(t-2\varepsilon)_+^2}{2}\right)\end{aligned}$$

by a union bound and our favorite concentration inequality. This in turn is at most $\exp\left(C\cdot\frac{1}{\varepsilon} - \frac{n}{2}(t-2\varepsilon)_+^2\right)$, so if we take $\varepsilon = n^{-1/3}$ and $t = 2\varepsilon + C'\sqrt{\frac{\log\frac{1}{\delta}}{n} + \frac{1}{n^{2/3}}}$, we get exactly δ with the appropriate choice of C' to cancel out the C term. □

With this function family $\mathcal{W}_{\text{Lip}}$, we can confirm that $\text{CE}(f, \mathcal{W}_{\text{Lip}})$ is sound and complete; the completeness is trivial but the soundness requires some more work. The idea is to let $S = f(X)$, and

$$E = E(S) = \mathbb{E}[Y|S] - S$$

is a random variable and where intuitively

$$\sup_{w\in\mathcal{W}_{\text{Lip}}} \mathbb{E}[w(S)E(S)] = 0 \implies E(S) = 0 \text{ with probability 1},$$

as long as S is a Borel random variable (since Lipschitz functions are dense enough in this reasonable sense).

Remark 179. *The book provides many equivalent calibration measures / benchmarks, but we spend a lot of time in the literature asking what the right $\mathcal{W}$s are for different problem classes.*

Finally, we'll discuss how proper losses can help us audit calibration of predictors. Define the **post-processing gap**

$$\text{gap}(f, \ell, \mathcal{W}) = \mathbb{E}\left[\ell(f(X), Y)\right] - \inf_{W\in\mathcal{W}} \mathbb{E}\left[\ell(f(X) + w(f(X)), Y)\right]$$

(this is how much we can improve a loss by tweaking a prediction via some function class).

Proposition 180

Suppose $\ell(\mu, y) = \frac{1}{2}||\mu - y||_2^2$ is the squared error, and suppose $\mathcal{W}$ is some convex, bounded, and symmmetric collection of functions. Define the radius $R^2(f) = \sup_{w\in\mathcal{W}} \mathbb{E}\left[||w(f(X))||_2^2\right]$. Then the calibration error satisfies the two-sided bounds

$$\frac{1}{2}\min\left\{\text{CE}(f, \mathcal{W}), \frac{\text{CE}(f, \mathcal{W})^2}{R^2(f)}\right\} \le \text{gap}(\ell, f, \mathcal{W}) \le \text{CE}(f, \mathcal{W}).$$

This result was only stated for the squared error, but it essentially extends to any proper loss. And what this says algorithmically is that if we can't improve our losses anymore by adding in some w, then we know we're calibrated.

The proof is not so bad – we just calculate out some derivatives – and the point is that given any proper loss, if we have a procedure that iteratively chooses $w_k = \text{argmin}\mathbb{E}[\ell(f(X) + \sum_{i=1}^{k-1} w_i(f(X)) + w(f(X)), Y)]$, then this limit is calibrated and we can develop convergence rates for the procedure.

17 November 20, 2025

We'll discuss **surrogate risk consistency** today, connecting our information-theoretic ideas with convexity. We'll define the problem and do a deep dive into the binary margin-based classification problem, and we'll also talk about classification calibration. This is a bit of a difference from what we did last time, but everything kind of provides different perspectives on prediction problems.

Example 181

Suppose we have a supervised learning problem, meaning that we have data in (X, Y) pairs in some space $\mathcal{X} \times \mathcal{Y}$ and we want to fit a predictive model $f : \mathcal{X} \to \mathbb{R}^k$ to minimize some expected loss $L(f) = \mathbb{E}[\ell(f(X), Y)]$ (where the loss maps $\mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$). The issue is that in many cases, we may be trying to minimize some loss that is not convex or smooth – a typical example to keep in mind is the margin-based binary classification, where $Y \in \{-1, 1\}$, f is real-valued, and $\ell(s, y) = 1\{sy \le 0\}$ (that is, we suffer a loss if we predict the wrong sign); this is not convex and it doesn't even tell us how to optimize locally because derivatives are always zero.

A similar story occurs for multiclass classification where $\mathcal{Y} = [k]$ and our predictions are some vector $f(x) \in \mathbb{R}^k$, and our loss is something like $\ell(s, y) = 1\{s_y \le \max_{j \ne y} s_j\}$, meaning that we lose if we assign a higher score to an incorrect label. So just optimizing over the number of mistakes doesn't give us information for differentials.

The idea is that we'll replace this non-convex loss ℓ with a **convex surrogate** which we call $\varphi : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$, and then we'll minimize this surrogate risk $L_\varphi(f) = \mathbb{E}[\varphi(f(X), Y)]$ instead. In the binary case (where we'll spend all of our time doing derivations, though things extend fairly easily), this is like using the hinge loss

$$\varphi(s, y) = (1 - sy)_+ = \max(1 - sy, 0)$$

so that we are penalized for having $sy \le 1$ at a linear rate. And similarly in multi-class classification we might use cross-entropy

$$\varphi(s, y) = \log\left(1 + \sum_{j \ne y} e^{sj - sy}\right).$$

We then want to ask whether doing this is okay (and whether we've lost anything in the process).

Definition 182

The function φ is **(surrogate-risk-)consistent** for the loss ℓ if for all distributions on (X, Y), we have that

$$L_\varphi(f_n) \to L_\varphi^* = \inf_f L_\varphi(f)$$

(a sequence of functions going to the best possible surrogate risk) implies that

$$L(f_n) \to L^* = \inf_f L(f).$$

(the same sequence also approaches the best possible loss).

Notice that this is an infimum over all functions. In other words, if we were given infinitely powerful predictors, this is asking whether we get the right thing. So this is a minimal requirement we should ask for from losses.

The approach we'll take is to proceed conditionally from here: we look at each X individually because we have so much freedom in our functions. That is, $\mathbb{E}[\varphi(f(X),X)] = \mathbb{E}[\mathbb{E}[\varphi(f(X),X)|X]]$ by the tower law of conditional expectation, and $L(f) = \mathbb{E}[\mathbb{E}[\ell(F(X),Y)|X]]$, so we can **fix a single** x and let $s = f(x)$ be the corresponding score, and we look at the gaps for the surrogate

$$\mathbb{E}[\varphi(s,Y)|X=x] - \inf_s \mathbb{E}[\varphi(s,Y)|X=x]$$

and the true loss

$$\mathbb{E}[\ell(s,Y)|X=x] - \inf_s \mathbb{E}[\ell(s,Y)|X=x].$$

Intuitively, if whenever the latter quantity is positive (so we don't do as well as possible) then the former also has to be positive, then we should have consistency. (If forcing us to be wrong implies we can't minimize the surrogate, then in order to minimize the surrogate we must be correct.)

Example 183

We'll now make this more mathematically formal, and we'll focus on the binary classification case because it already accounts for most of the subtleties in general. We will think about margin-based losses, where $f(x) \in \mathbb{R}$, $Y \in \{-1,1\}$, and our surrogate is of the form $\varphi(s,y) = \varphi(sy)$ for some convex function φ. (For example, the hinge loss has $\phi(s) = (1-s)_+$, logistic regression has $\phi(s) = \log(1+e^{-s})$, and boosting has $\phi(s) = e^{-s}$. Basically, all of these predict something with the right sign and suffer less loss the more correct we are.

Remark 184. *We won't talk much about how to actually find a surrogate φ from ℓ in this class, but there are indeed some somewhat general recipes (under some conditions).*

To do the conditioning, we'll define $\eta(x) = \mathbb{P}(Y=1|X=x)$, and so our loss with the indicator function is

$$L(f) = \mathbb{E}\left[\eta(X)1\{f(X)\geq 0\} + (1-\eta(X))1\{f(X)\geq 0\}\right],$$

and meanwhile with our surrogate function ϕ we have

$$L_\phi(f) = \mathbb{E}\left[\eta(X)\phi(f(X)) + (1-\eta(X))\phi(-f(X))\right].$$

Again, still proceeding completely conditionally, we can define the **conditional risks** for any real s and $\eta \in [0,1]$ (coming from a specific x)

$$r(s,\eta) = \eta 1\{s\leq 0\} + (1-\eta)1\{s\geq 0\}$$

and the **conditional ϕ-risk**

$$r_\phi(s,\eta) = \eta\phi(s) + (1-\eta)\phi(-s).$$

Then we have the population risk

$$L(f) = \mathbb{E}[r(f(X),\eta(X))]$$

and the population ϕ-risk

$$L_\phi(f) = \mathbb{E}[r_\phi(f(X),\eta(X))],$$

so if we can just relate these quantities we'll be in good shape. We'll try to implement our intuition about gaps in terms of these risks now: suppose that we constrained that $r_\phi(s,\eta)$ is strictly bigger than $r_\phi^*(\eta) = \inf_s r_\phi(s,\eta)$ whenever s

has the incorrect sign, meaning that $s(2\eta - 1) \leq 0$ (since if $\eta > \frac{1}{2}$, then we want to predict s positive, and otherwise we want to predict s negative). Then we would expect that the ϕ loss function is consistent.

Definition 185

Define the quantities $r_\phi^*(\eta) = \inf_s r_\phi(s, \eta)$ and $r_\phi^{\text{wrong}}(\eta) = \inf_{(2\eta-1)s \leq 0}\{\eta(\phi(s)) + (1-\eta)\phi(-s)\}$.

For example for the hinge loss $\phi(s) = (1-s)_+$, if $\eta > \frac{1}{2}$ then the ϕ-risk is

$$r_\phi(s, \eta) = \eta(1-s)_+ + (1-\eta)(1+s)_+.$$

At $s = 1$ this would be equal to $2(1-\eta)$ and for larger s it grows linearly at slope $1-\eta$, and at $s = -1$ this is equal to 2η (a little larger) and then growing linearly at a slope $-\eta$. And in between it's linear, so $r_\phi(s, \eta)$ it's piecewise linear with kinks at -1 and 1, and the smaller kink occurs at the correct value. Therefore $r_\phi^*(\eta)$ picks the smaller of the two kinks, meaning

$$r_\phi^*(\eta) = 2\min(\eta, 1-\eta).$$

But if we constrain to be wrong (meaning we force the sign of s), we can't pick the correct kink and thus the least wrong we can be is at 0, and

$$r_\phi^{\text{wrong}}(\eta) = r_\phi(0, \eta) = 1 - \eta + \eta = 1.$$

So indeed there is a strict gap unless $\eta = \frac{1}{2}$, meaning that for any x we always incur a worse loss under the constraint.

Observe also that $r_\phi^*(\eta) = r_\phi^*(1-\eta)$ and $r_\phi^{\text{wrong}}(\eta) = r_\phi^{\text{wrong}}(1-\eta)$, so we can take without loss of generality that $\eta > \frac{1}{2}$. We define the **sub-optimality gap** for any $\delta \geq 0$ by

$$\Delta_\phi(\delta) = r_\phi^{\text{wrong}}\left(\frac{1+\delta}{2}\right) - r_\phi^*\left(\frac{1+\delta}{2}\right)$$

(how much do we pay for being incorrect when the conditional probability is δ apart); for example for the hinge loss $\min(\frac{1-\delta}{2}, \frac{1+\delta}{2}) = \frac{1-\delta}{2}$ so $\Delta_\phi(\delta) = 1 - 2\left(\frac{1-\delta}{2}\right) = \delta$. So the further our probability is, the bigger penalty we have for a wrong sign.

Definition 186

The margin loss function ϕ is **classification calibrated** (this is not the same as calibration in the sense of last lecture) if $\Delta_\phi(\delta) > 0$ for all $\delta > 0$ (so making a wrong prediction is worse than making the best prediction).

What we'd really like is to develop an explicit comparison inequality, meaning that we have some function ψ such that

$$\psi(L(f) - L^*) \leq L_\phi(f) - L_\phi^*.$$

where $\psi(\delta) > 0$ if $\delta > 0$. This means that if we send the gap in our ϕ-risks to zero, then we must send the gap in our true risk to zero as well. And the point is that if ψ is **convex** then we can use Jensen and hope that ψ applied to our zero-one error gives something like the gap Δ.

We can transform a generic function into a convex one by taking its convex envelope (that is, its biconjugate), so we define

$$\psi(\delta) = \Delta_\phi^{**}(\delta);$$

this is the largest convex function which is bounded above by Δ_ϕ.

Theorem 187

Consider binary classification with the notation above. For any f and with ψ as above, we have $L_\phi(f) \geq L_\phi^* \geq \psi(L(f) - L^*)$, and the following conditions are equivalent if ϕ is nonnegative:

1. ϕ is classification calibrated (which is the same as saying $\Delta_\phi(\delta) > 0$ for $\delta > 0$),
2. $\psi(\delta)$ is strictly positive for $\delta > 0$,
3. $L_\phi(f_n) \to L_\phi^*$ implies $L(f_n) \to L^*$.

We don't quite have the cleanliness of statements beyond binary classification, but here we can make everything fairly explicit. And notice that (1) implies (2) means that Δ being positive implies $\psi = \Delta^{**}$ is positive, but in general it's not true that the biconjugate of a positive function is strictly positive.

Proof sketch. For the main statement, we expand the true zero-one error

$$\begin{aligned} r(s,\eta) - r^*(\eta) &= \eta 1\{s \leq 0\} + (1-\eta)1\{s \geq 0\} - \min(\eta, 1-\eta) \\ &= \begin{cases} 0 & \text{if } s(2\eta - 1) > 0, \\ \max(\eta, 1-\eta) - \min(\eta, 1-\eta) = |2\eta - 1| & \text{if } s(2\eta-1) \leq 0. \end{cases} \end{aligned}$$

Now by the convexity argument we described,

$$L(f) - L^* = \mathbb{E}\left[1\left\{\operatorname{sgn}(f(X)) \neq \operatorname{sgn}(2\eta(X) - 1)\right\} \cdot |2\eta(X) - 1|\right].$$

Now applying ψ to both sides and using Jensen's inequality,

$$\psi(L(f) - L^*) = \mathbb{E}\left[\psi\left(1\left\{\operatorname{sgn}(f(X)) \neq \operatorname{sgn}(2\eta(X) - 1)\right\} \cdot |2\eta(X) - 1|\right)\right].$$

Observe that ψ is nonnegative because Δ is nonnegative, and by definition of the biconjguate (remember η is a symmetric function about $\frac{1}{2}$)

$$\begin{aligned} \psi(|2\eta - 1|) &\leq \Delta_\phi(|2\eta - 1|) \\ &= \inf_{s(2\eta-1)\leq 0} \left\{r_\phi(s,\eta) - r_\phi^*(\eta)\right\} \\ &\leq r_\phi(s,\eta) - r_\phi^*(\eta) \text{ if } s \text{ is wrong.} \end{aligned}$$

So substituting this into our blue quantity above (and only caring about when the indicator fires because $\psi(0) \geq 0$), we have

$$\begin{aligned} \psi(L(f) - L^*) &\leq \mathbb{E}\left[1\{\operatorname{sgn}(f) \neq \operatorname{sgn}(2\eta - 1)\}(r_\phi(f(X), \eta(X)) - r_\phi^*(\eta(X)))\right] \\ &\leq \mathbb{E}\left[r_\phi(f(X), \eta(X)) - r_\phi^*(\eta(X))\right] \\ &= L_\phi(f) - L_\phi^*, \end{aligned}$$

as desired.

For the various equivalences, (2) implies (3) is trivial from the inequality we just proved. For (3) implies (1), since we're requiring condition (3) for **any** distribution, we can choose a point mass on $x \in \mathcal{X}$ and vary only the probability $\eta(x)$ of having 1. Then condition (3) tells us that $s_n \in \mathbb{R}$, we have $r_\phi(s_n, \eta) \to r_\phi^*(\eta)$ implies $r(s_n, \eta) \to r^*(\eta)$, which implies (by a quick contradiction argument and chasing around the definitions) that $\Delta_\phi(\delta) > 0$.

Finally, (1) implies (2) is a bit more subtle and requires some specific convexity facts. We know that

$$r_\phi^*(\eta) = \inf_s \{\eta\phi(s) + (1-\eta)\phi(s)\},$$

and since ϕ is nonnegative this is a nonnegative function. This is an infimum of affine functions, hence concave, and it's in fact closed concave so that our function is actually continuous. (One-dimensional convex and concave functions are upper semicontinuous, and the only problem is going to something big on the boundary.) Similarly we see that $r_\phi^{\text{wrong}}(\eta)$ is continuous, so their difference is also continuous, and thus $\Delta_\phi(\delta) = r_\phi^{\text{wrong}}\left(\frac{1+\delta}{2}\right) - r_\phi^*\left(\frac{1+\delta}{2}\right)$ is continuous in $\delta \geq 0$. But if $\psi = \Delta^{**}$ were zero somewhere to the left of some δ_0, then we can look at the infimum of $\Delta(\delta)$ for all $\delta \geq \frac{\delta_0}{2}$; this is strictly positive because we're taking the infimum over $\delta \in [0,1]$. So actually we could have just connected a straight line starting at $\frac{\delta_0}{2}$, which strictly increases our convex conjugate and creates a contradiction. Thus ψ is indeed strictly positive, as desired. □

Example 188

Consider the exponential (boosting) loss $\phi(s) = e^{-s}$ above. Then we can take a derivative

$$r_\phi(s,\eta) = \eta e^{-s} + (1-\eta)e^{-s}, \quad r_\phi'(s,\eta) = -\eta e^{-s} + (1-\eta)e^{s},$$

so the loss is minimized when $e^{2s} = \frac{\eta}{1-\eta} \implies s = \frac{1}{2}\log\frac{\eta}{1-\eta}$.

Substituting this back in, we can calculate r_ϕ^* and r_ϕ^{wrong}: we find that

$$r_\phi^*(\eta) = 2\sqrt{\eta(1-\eta)}, \quad r_\phi^{\text{wrong}}(\eta) = r_\phi(0,\eta) = 1$$

(the "least wrong" prediction is at $s = 0$). This means that

$$\Delta_\phi(\delta) = 1 - 2\sqrt{\frac{1-\delta}{2}\cdot\frac{1+\delta}{2}} = 1 - \sqrt{1-\delta^2},$$

and this itself is convex and bounded from below by $\frac{\delta^2}{2}$. So ψ becomes a quadratic function, and we get the following corollary:

Corollary 189

The best possible zero-one loss of any classifier satisfies

$$\frac{1}{2}(L_{0-1}(f) - L_{0-1}^*)^2 \leq L_\phi(f) - L_\phi^*.$$

We calculated previously that for the hinge loss we had $\Delta_\phi(\delta) = \delta$, and so for $\phi(s) = (1-s)_+$ we actually have $L_{0-1}(f) - L_{0-1}^* \leq L_\phi(f) - L_\phi^*$.

Fact 190

If ϕ is convex, then it is classification calibrated if and only if $\phi'(0) < 0$ (in particular it must be differentiable). So our loss has to say that predicting the correct sign is better than predicting the incorrect one, but this gives us a full characterization.

18 December 2, 2025

The last two lectures of the course will focus on a fun topic which incorporates a lot of material so far, **bandits and causal estimation**. The cool thing is that causal problems are the foundation of statistics, and we'll use some information theory to analyze procedures and get upper and lower bounds.

Example 191

Suppose we have a set of actions $\mathcal{A}$ and a set of responses $[Y(a)]_{a\in\mathcal{A}}$ for taking those actions. Our goal is to find the action that maximizes the average response (reward) $\mu_a = \mathbb{E}[Y(a)]$.

An example of such a problem is **treatment estimation**, in which we can take actions $\mathcal{A} = \{0, 1\}$ (where 0 is the control and 1 is the treatment) and each individual i has two potential outcomes $Y_i = (Y_i(0), Y_i(1))$ depending on which action we perform (but importantly, notice that we can only ever observe one of those two values). So if we have a randomized treatment like in a study which is independent of the individual, we have $\mathbb{E}[Y_i(a)|A_i = a] = \mathbb{E}[Y_i(a)] = \mu_a$, and the average treatment effect is

$$\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)],$$

and in a (balanced) randomized control trial, we can estimate this by letting $A_i \in \{0, 1\}$ be chosen uniformly such that half of them are treated, and then our estimator is

$$\hat{\tau} = \frac{1}{n/2}\left(\sum_{i:A_i=1} Y_i(1) - \sum_{i:A_i=0} Y_i(0)\right).$$

So we have a study or "two arms," and we can choose to pull one or the other and we get some feedback. (We call this a bandit problem because in the theoretical computer science literature, people thought of this as a problem of identifying the best possible slot machine.) There are two versions of these types of problems we can try to solve; one is **pure exploration** (where our goal is solely finding the best argmax) and the other is **regret-based formulation** where we also care about results along the trial and also suffer losses along the way (for example if we notice that everyone getting our treatment is getting sick): that is, we have some true regret

$$\mathrm{Reg}_n = \max_{a\in\mathcal{A}} \sum_{i=1}^{n} \mu_a - \mu_{A_i}$$

and some expected regret

$$\overline{\mathrm{Reg}}_n = \max_{A\in\mathcal{A}} \mathbb{E}\left[\sum_{i=1}^{n} \mu_a - \mu_{A_i}\right].$$

We will typically prove upper bounds (find algorithms) in the regret-based formulation, which is what we do today, and we will typically prove lower bounds in the exploration setting, since those are the more difficult problems.

Example 192

Our general setting will be that we have some losses $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ which measure the quality of actions against responses. (So for example we might have $\ell(a, y) = -y$ in a treatment effect problem.) We will have a parametric family $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ of distributions on the potential response vectors $[Y(a)]_{a \in \mathcal{A}}$, and our game proceeds as follows (for $t \in \{1, 2, \cdots, n\}$:)

- Choose a distribution ρ_t on the action set $\mathcal{A}$ based on the history we have observed up to this point $H_{t-1} = \{(A_i, Y_i(A_i)\}_{i=1}^{t-1}$.
- Draw the action A_t according to ρ_t, and observe $Y_t(A_t)$ according to some static P_θ (which we don't know); we also get to observe / measure the loss $\ell(A_t, Y_t(A_t))$. We then suffer our regret $\ell(A_t, Y_t(A_t)) - \ell(A^*, Y_t(A^*))$ for A^* the optimal action minimizing the loss, though we don't actually get to observe what this quantity is (we don't get to see what happens to the individual if they didn't take the drug).

There was an article in the New York Times about self-driving cars this morning; right now, the rate of accidents for Waymos is something like 10 percent that of human drivers (though this isn't like a randomized control trial in various ways). So this is sort of a bandit problem, since every Waymo drive is a drive that a human didn't take.

Remark 193. *Assuming that P_θ is static is natural in something like a medical trial; it's maybe less natural for something like self-driving cars where adding more cars changes people's behavior, but if we allow things to change over time we get lots of different variants of regret and more complications.*

We'll work with the **Bayesian regret**, in which we have some prior π on $\theta \in \Theta$ and define the average regret

$$\overline{\mathrm{Reg}}_n(\mathcal{A}, \ell, \pi) = \mathbb{E}_\pi \left[\sum_{t=1}^{n} \ell(A_t, Y_t(A_t)) - \ell(A^*, Y_t(A^*)) \right],$$

where here $A^* = \mathrm{argmin}_{a \in \mathcal{A}} \mathbb{E}[\ell(a, Y(a))|\theta]$ is a random action depending on $\theta \sim \pi$. The big-picture idea is that if we were to design an algorithm, then at every step of the algorithm either we should get a lot of information about the good treatments, or we should suffer low regret.

Definition 194

The **instantaneous information** of an action $a \in \mathcal{A}$ is defined by

$$I_t(a) = I(A^*; Y_t(a)|H_{t-1}).$$

However here we are specifically conditioning on the actual realization of the history, so we aren't averaging over the thing we condition on. (We should think of this as $I(A^*; Y_t(a)|A_1 = a_1, Y_1(a_1) = y_1, \cdots, A_{t-1} = a_{t-1}, Y_{t-1}(A_{t-1}) = y_{t-1})$.) The **information from sampling** $A \sim \rho$ is then

$$I_t(\rho) = \mathbb{E}_{A \sim \rho}[I_t(a)] = \sum_a \rho(a) I_t(a) = \sum_a \rho(a) I(A^*; Y_t(a)|H_{t-1}) = I(A^*; Y_t(A), A|Y_{t-1})$$

where in the last expression A is drawn from ρ.

Definition 195

The **regret ratio** at time t is

$$R_t(\rho) = \frac{\mathbb{E}_\rho[\ell(A, Y_t(A)) - \ell(A^*, Y_t(A^*))|H_{t-1}]^2}{I_t(\rho)}.$$

So the regret ratio takes the square of the gap between expected loss and the optimal action, divided by the information I get from it. This means the design principle will be to choose R_t to be small at every step, and the neurons that should be firing is that this is a lot like Donsker-Varadhan since that gives us relations between squared distances and information (in fact Pinsker is a special case of this).

So we'll try to get a regret bound that depends on this regret ratio, and then we'll show that we can actually give sufficient conditions (e.g. sub-Gaussianity) to bound that ratio with some specific algorithms.

Theorem 196

The average Bayesian regret after n steps satisfies

$$\overline{\mathrm{Reg}}_n(\mathcal{A}, \ell, \pi) \leq \sqrt{\mathbb{E}_\pi\left[\sum_{t=1}^n R_t(\rho_t)\right]} \cdot \sqrt{I(A^*; (A_i, Y_i(A_i))_{i=1}^n)}.$$

So if we have a procedure that makes sure the regret ratios are always small, then the total regret is no worse than the amount of information it takes to identify the optimal action.

Proof. This is just Cauchy-Schwarz: defining the random variable $L_t(a) = \ell(a, Y_t(a))$ for our realized losses,

$$\mathbb{E}\left[\sum_{t=1}^n L_t(A_t) - L_t(A^*)\right] = \mathbb{E}\left[\sum_{t=1}^n \frac{L_t(A_t) - L_t(A^*)}{\sqrt{I_t(\rho_t)}}\sqrt{I_t(\rho_t)}\right]$$

where remember ρ_t is the distribution of our tth action. But this quantity is the same as first conditioning on the history, and since $I_t(\rho_t)$ is a function of H_{t-1} the tower property tells us that this is the same as

$$\mathbb{E}\left[\sum_{t=1}^n \frac{\mathbb{E}[L_t(A_t) - L_t(A^*)|H_{t-1}]}{\sqrt{I_t(\rho_t)}}\sqrt{I_t(\rho_t)}\right].$$

By Cauchy-Schwarz, this is then at most $\sqrt{\sum_{t=1}^n \mathbb{E}_\pi\left[\frac{\mathbb{E}[L_t(A_t) - L_t(A^*)|H_{t-1}]^2}{I_t(\rho_t)}\right]} \cdot \sqrt{\mathbb{E}[\sum_{t=1}^n I_t(\rho_t)]}$, and now to relate this to the mutual information observe that by the definition of the usual conditional mutual information

$$\mathbb{E}_\pi[I_t(\rho_t)] = \mathbb{E}_\pi[I(A^*; Y_t(A_t), A_t|H_{t-1})] = I(A^*; Y_t(A_t), A_t|(Y_i(A_i), A_i)_{i=1}^{t-1}),$$

which yields what we want by the chain rule for mutual information. □

Corollary 197

Suppose the set of possible actions $\mathcal{A}$ is finite. Then

$$I(A^*; (Y_i(A_i), A_i)_{i=1}^n) \leq H(A^*) \leq \log |\mathcal{A}|,$$

and therefore the regret for any prior satisfies

$$\overline{\text{Reg}}_n(\mathcal{A}, \ell, \pi) \leq \sqrt{\mathbb{E}_\pi \left[\sum_{t=1}^n R_t(\rho_t) \right]} \cdot \sqrt{\log |\mathcal{A}|}.$$

Thus really we just need to control this regret ratio in a lot of problems and we're done. Let's show an example of an algorithm:

Example 198

Thompson sampling (also called posterior sampling) was one of the first bandit algorithms, but it was ignored by the TCS literature even though it was better for many decades. What we do is take π_t to be the posterior distribution on our parameter θ given H_{t-1} and then let ρ_t be the induced distribution on actions by drawing θ_t according to π_t and letting $A_t = \text{argmin}_{a \in \mathcal{A}} \mathbb{E}[\ell(a, Y(a))|\theta_t]$. (So we draw our random parameter according to the posterior and choose the known best action according to that.)

So if we are able to estimate that posterior π_t, it's pretty easy to implement. For example, suppose we have a k-**armed Bernoulli bandit**, meaning we can take k different actions / treatments and the responses are all either 0 or 1 (seeing whether those treatments worked well enough or not). So conditional on $\theta = [\theta_a]_{a=1}^k$, $Y(a)$ are independently $\text{Ber}(\theta(a))$.

Recall that the **Beta**(α, β) **distribution** is a probability distribution on $[0, 1]$ with density

$$\pi(t) \propto t^{\alpha-1}(1-t)^{\beta-1},$$

which is very nice because Bernoulli observations drawn from Beta distributions can be easily updated: the posterior is still Beta. Defining the counts of outcomes

$$N_t^0(a) = \#\text{ outcomes where } Y_t(a) = 0 \text{ when taking action } a \text{ by time } t,$$

and similarly defining N_t^1 to be the number of such outcomes when $Y_t(a) = 1$, we can do an explicit calculation and find that if our prior π is that each $\theta(a)$ is $\text{Beta}(1, 1)$ (though it can be any choice), the posterior on θ at time t is

$$\theta(a) \sim \text{Beta}(1 + N_t^1(a), 1 + N_t^0(a)).$$

So what this says is that as we iterate this procedure, we draw our parameters according to Beta distributions $(1 + N_t^1(a), 1 + N_t^0(a))$ and pick which one is largest out of all $a \in \{1, \cdots, k\}$; that is, we let $A_t = \text{argmax}_a \theta(a)$ and observe $Y_t(A_t)$ and update counts accordingly. At first any of the as has a fairly large chance of being chosen, but eventually whatever arm has the actual highest average will eventually get more and more 1s, and so it will have the highest mean and the beta distributions will be quite concentrated.

Let's try to analyze this; it'll take a bit of work but is doable. We'll make the standing assumption that the loss $\ell(a, Y(a))$ is σ^2-sub-Gaussian (there's some subtlety with what exactly this means; specifically we're saying that for **any** distribution π on θ, the induced distribution on $\ell(a, Y_t(a))$ conditional on the history H_{t-1} is always σ^2-sub-Gaussian

for any $a \in \mathcal{A}$. (For example this is just true if ℓ is bounded.)

Lemma 199 (Thompson sampling regret ratio bound)
Assume that the actual optimal action A^* and our actions A are all drawn according to ρ, where A^* is the random argmin of $\mathbb{E}[\ell(a, Y(a))|\theta]$ when $\theta \sim \pi$. (This is true in Thompson sampling **if we know the actual true prior**, because π_t is defined to be the posterior distribution, so A_t has the same distribution as A^* conditional on the history.) Then

$$\sum_{a \in \mathcal{A}} \rho(a) \Big(\mathbb{E}[\ell(a, Y(a))] - \mathbb{E}[\ell(a, Y(a))|A^* = a] \Big) \leq \sqrt{2\sigma^2 |\mathcal{A}|} \cdot \sqrt{I(A^*; A, Y(A))}.$$

Proof. Let $L(a) = \ell(a, Y(a))$ be the random realization of the loss; we know that by Cauchy-Schwarz,

$$\sum_a \rho(a) \Big(\mathbb{E}[L(a)] - \mathbb{E}[L(a)|A^* = a] \Big) \leq \Big(\sum \rho(a)^2 \left(\mathbb{E}[L(a)] - \mathbb{E}[L(a)|A^* = a] \right)^2 \Big)^{1/2} \sqrt{|\mathcal{A}|}.$$

Now noticing that

$$\rho(a)^2 \left(\mathbb{E}[L(a)] - \mathbb{E}[L(a)|A^* = a] \right)^2 \leq \rho(a) \sum_{a^* \in \mathcal{A}} \rho(a^*) \left(\mathbb{E}[L(a)] - \mathbb{E}[L(a)|A^* = a^*] \right)^2$$

since we just added in a bunch of positive numbers to our left-hand side. But then by Donsker-Varadhan (mean differences are upper bounded by KL) we know that

$$(\mathbb{E}[L(a)] - \mathbb{E}[L(a)|A^* = a^*])^2 \leq 2\sigma^2 D_{\mathrm{KL}}(P_{a|A^*=a^*} || P_a),$$

where P_a is the marginal distribution on $L(a)$ and $P_{a|A^*=a^*})$ is the distribution of $L(a)$ given that $A^* = a^*$. But now by definition of mutual information and then data processing,

$$\begin{aligned} \sum_{a^* \in \mathcal{A}} \rho(a^*) D_{\mathrm{KL}}(P_{a|A^*=a^*} || P_a) &= I(\ell(A, Y(A)); A^* | A = a) \\ &\leq I(A, Y(A); A^* | A = a). \end{aligned}$$

Substituting our bound back into the inequalities above yields exactly the result that we want. □

This yields our regret bound for finite-action sets, since we've just proven that $\mathbb{E}[R_t(\rho_t)] \leq 2\sigma^2 |\mathcal{A}|$:

Corollary 200
If $\mathcal{A}$ is finite and losses are σ^2-sub-Gaussian, then Thompson sampling satisfies

$$\overline{\mathrm{Reg}}_n(\ell, \mathcal{A}, \pi) \leq \sqrt{2\sigma^2 |\mathcal{A}|} \cdot \sqrt{H(A^*)} \cdot \sqrt{n}.$$

Corollary 201
For our k-armed Bernoulli bandit, rewards and losses are either 0 or 1, so $\sigma^2 = \frac{1}{4}$ and thus the average regret is

$$\overline{\mathrm{Reg}}_n(\ell, \mathcal{A}, \pi) \leq \sqrt{\frac{k \log k}{2}} \cdot \sqrt{n}.$$

In our last lecture, we'll see that this is actually sharp up to the log (and we can indeed remove the log factor to get something optimal too).

19 December 4, 2025

We'll finish up our discussion of bandits, in particular discussing linear bandits (a model of more sophisticated scenarios which ends up being fun and practical) and how to get lower bounds using Assouad-type methods.

Our setup is that we have some set of actions $\mathcal{A}$ and some losses $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$, where given our action the responses $Y(a)$ will be drawn from some P_θ. Our standing assumption today is that the losses ℓ are σ^2-sub-Gaussian no matter the history or how we choose our actions. We saw last time that the average regret has an information-theoretic bound

$$\overline{\mathrm{Reg}}_n(\pi, \ell\mathcal{A}) = \mathbb{E}_{\theta\sim\pi}\left[\sum_{t=1}^n \ell(A_t, Y_t(A_t)) - \ell(A^*, Y(A^*))\right] \le \sqrt{\mathbb{E}_\pi\left[\sum_{t=1}^n R_t(\rho_t)\right] I(A^*; (A_t, Y_t(A_t))_{t=1}^n)}$$

for R the regret ratio

$$R_t(\rho) = \frac{\mathbb{E}_\rho[\ell(A, Y_t(A)) - \ell(A^*, Y_t(A^*))|H_{t-1}]^2}{I_t(\rho)}.$$

and I the mutual information between A^* and $Y_t(A)$, A given H_{t-1} when A is sampled according to the action distribution ρ. So the game is that the mutual information has some upper bound, and so as long as the losses are sub-Gaussian we can typically relate the expected gap in losses to the mutual information with Donsker-Varadhan.

We'll do that again today but in a slightly more elaborate scenario where we have a different feedback mechanism for the responses we get:

Example 202

Assume we have a model where our response is of the form $Y(a) = \langle\theta, a\rangle + \xi$ for some noise ξ (where $a, \theta \in \mathbb{R}^d$). This is a fairly stylized model but is kind of a stand-in for the general principle that knowing about responses $Y(a)$ under one action a might tell us about similar actions under a' (for example, dosage responses might be correlated or similar if we give 100 or 101 milligrams of the drug), rather than just having a bunch of independent arms. (Everything we did last time was valid under any dependence conditions; it just couldn't leverage the fact that two arms might be very similar.) We will use the losses $\ell(a, y) = -y$.

One semi-explicit example which motivates this is **routing in networks**: perhaps we have nodes $\{1, \cdots, d\}$ in a network and some edges connecting them, and we want to send data from some source node to a target node. An action in this case is a matrix $a \in \{0, 1\}^{d\times d}$, where $a_{ij} = 1$ if we send traffic across the edge $i \to j$ and 0 otherwise. The total cost of sending information is then $\langle\theta, a\rangle = \sum_{i,j}\theta_{ij}a_{ij}$, and our goal is to find the shortest path.

To analyze regret bounds in this case, we'll have to relate the regret ratio to a ratio of variances. The basic observation (which follows from Donsker-Varadhan) is the following:

Lemma 203

If ℓ is σ^2-sub-Gaussian, then for any fixed action a (and the randomness here is over Y and A^*)

$$\mathrm{Var}(\mathbb{E}[\ell(a, Y(a))|A^*]) \le 2\sigma^2 I(A^*; Y(A)|A = a).$$

Proof. We have by data processing that

$$2\sigma^2 I(A^*; Y(A), A|A = a) \ge 2\sigma^2 I(A^*; \ell(A, Y(A))|A = a),$$

and now information is the same as KL-divergence so we can write this as

$$2\sigma^2\mathbb{E}\left[D_{\mathrm{KL}}(P_{a|A^*}||P_a)\right] \geq \int \Big(\mathbb{E}[\ell(a,Y(a)] - \mathbb{E}[\ell(a,Y(a))|A^*=a^*]\Big)^2 \rho(a^*)da^*,$$

where we've used Donsker-Varadhan in the last step. And this is exactly the conditional variance we're looking for. □

Intuitively, the left-hand side measures the variance in θ on the left side, and the right-hand side compensates for the noise in ξ if we think about having KL divergence between Gaussians. Writing out the Donsker-Varadhan step in more detail, recall that for any distributions P, Q, we have $D_{\mathrm{KL}}(P||Q) = \sup_g\{\mathbb{E}_P[g] - \log \mathbb{E}_q[e^g]\}$, meaning that if X is σ^2-sub-Gaussian under Q and we set $g = \lambda(X - \mathbb{E}_Q[X])$, then

$$\begin{aligned} D_{\mathrm{KL}}(P||Q) &\geq \mathbb{E}_P[g] - \log \mathbb{E}_q[e^g] \\ &\geq \lambda(\mathbb{E}_P[X] - \mathbb{E}_Q[X]) - \frac{\lambda^2\sigma^2}{2}, \end{aligned}$$

and choosing $\lambda = \frac{\mathbb{E}_P[X]-\mathbb{E}_Q[X]}{\sigma^2}$ we get that

$$D_{\mathrm{KL}}(P||Q) \geq \frac{(\mathbb{E}_P[X] - \mathbb{E}_Q[X])^2}{2}.$$

So if we average this over any average of distributions P, we get

$$\int D_{\mathrm{KL}}(P_\theta||Q)d\pi(\theta) \geq \frac{1}{2\sigma^2}\int (\mathbb{E}_{P_\theta}[X] - \mathbb{E}_Q[X])^2 d\pi(\theta),$$

and now setting Q to be exactly the average $\overline{P} = \int P_\theta \pi(\theta)d\theta$, the left-hand side becomes the information $I(\theta; \text{observations})$ and the right-hand side is $\frac{1}{2\sigma^2}\mathrm{Var}(\mathbb{E}[X|\theta])$ since $\mathbb{E}_{P_\theta}[X]$ is just $\mathbb{E}[X|\theta]$. So the main lesson here is that for a random variable X which is σ^2-sub-Gaussian conditioned on all θ, we get $\mathrm{Var}(\mathbb{E}[X|\theta]) \leq 2\sigma^2 I(X;\theta)$.

So now we can replace these information-theoretic quantities with ratios of covariances, which will let us do linear algebra:

Proposition 204

Define the ratio (replacing the denominator of the regret ratio with what we just derived)

$$V_t(\rho) = \frac{\mathbb{E}_\rho[\ell(A, Y_t(A)) - \ell(A^*, Y_t(A^*))|H_{t-1}]^2}{\int \mathrm{Var}(\mathbb{E}[\ell(a,Y(a))|A^*])\rho(a)da}.$$

Then $R_t(\rho) \leq 2\sigma^2 V_t(\rho)$, and furthermore posterior sampling in the linear bandit problem (that is, ρ is the distribution of A^* given the history) satisfies

$$V_t(\rho_t) \leq d.$$

Proof. We already proved the first inequality above, and now define the matrix $M = \mathrm{Cov}(\mathbb{E}[\theta|A^*])$. If A is distributed according to the action distribution ρ and remembering that $\ell(a, Y(a)) = -Y(a)$, we have

$$\mathbb{E}[\ell(a,Y(a))|\theta] = -\langle\theta, a\rangle.$$

Thus all we need to do is compute some variances: we have

$$\mathbb{E}_\rho[\mathrm{Var}(\langle\mathbb{E}[\theta|A^*], A\rangle|A)] = \langle \mathrm{Cov}(\mathbb{E}[\theta|A^*], \mathbb{E}_\rho[AA^T]\rangle$$

and our regret satisfies

$$\begin{aligned}\mathbb{E}[\ell(A, Y(A))] - \mathbb{E}[\ell(A^*, Y(A^*))] &= \mathbb{E}[\langle \theta, A^* - A\rangle] \\ &= \mathbb{E}[\langle \mathbb{E}[\theta|A^*], A^*\rangle] - \mathbb{E}[\langle \theta, A\rangle] \\ &= \mathbb{E}[\langle \mathbb{E}[\theta|A^*] - \mathbb{E}[\theta], A^*\rangle]\end{aligned}$$

because A and A^* are independent and sampled from ρ under posterior sampling, and θ is independent of A so we have $\mathbb{E}[\langle \theta, A\rangle] = \mathbb{E}[\langle \mathbb{E}[\theta], A\rangle]$ and then we can replace A with A^*. So now by Cauchy-Schwarz,

$$\begin{aligned}\mathbb{E}[\langle \mathbb{E}[\theta|A^*] - \mathbb{E}[\theta], A^*\rangle]^2 &= \mathbb{E}[\langle M^{-1/2}\mathbb{E}[\theta|A^*] - \mathbb{E}[\theta], M^{1/2}A^*\rangle]^2 \\ &= \mathbb{E}\left[||M^{-1/2}(\mathbb{E}[\theta|A^*] - \mathbb{E}[\theta])||^2\right] \cdot \mathbb{E}[||M^{1/2}A^*||^2] \\ &= \langle M^{-1}, \mathrm{Cov}(\mathbb{E}[\theta|A^*])\rangle \cdot \langle M, \mathbb{E}_\rho[AA^T]\rangle \\ &= \mathrm{tr}(M^{-1}M)\langle M, \mathbb{E}[AA^T]\rangle \\ &= d\langle M, \mathbb{E}[AA^T]\rangle.\end{aligned}$$

So plugging this and the other variance calculation in and going back to the ratio bounds and definitions, the ratio $V_t(\rho)$ in posterior sampling satisfies

$$V_t(\rho) \leq \frac{d\langle M, \mathbb{E}_\rho[AA^T]}{\langle M, \mathbb{E}_\rho[AA^T]\rangle} = d,$$

as desired. □

So in linear bandits, the ratio between our expected gap and the information we receive is at most the dimension, and that means we've proved the following basic theorem:

Theorem 205 (Linear bandits)
If we are just trying to maximize our inner products and have $\ell(a, y) = -y$, then

$$\overline{\mathrm{Reg}}_n(\ell, \mathcal{A}, \pi) \leq \sqrt{2n\sigma^2 d} \cdot \sqrt{I(A^*; (A_t, Y_t(A_t))_{t=1}^n)}.$$

Corollary 206
Suppose $\mathcal{A}$ is finite. Then $\sqrt{I(A^*; (A_t, Y_t(A_t))_{t=1}^n)} \leq H(A^*) \leq \log|\mathcal{A}|$, so in fact

$$\overline{\mathrm{Reg}}_n(\ell, \mathcal{A}, \pi) \leq \sqrt{2n\sigma^2 d}\sqrt{\log|\mathcal{A}|}.$$

For example, if we want our action set $\mathcal{A}$ to be the ℓ_2-ball, the log-covering-number of that by ε-balls is $d\log\frac{1}{\varepsilon}$, so we heuristically expect that our regret bound looks like

$$\overline{\mathrm{Reg}}_n(\ell, \mathcal{A}, \pi) \lesssim \sqrt{2n}\sigma d.$$

And via some matrix convexity facts, if we assume that π is something like $N(0, \Sigma_0)$ and our noise is $N(0, \sigma^2)$, then at every step our posterior $\theta|(a_i, y_i)_{i=1}^t$ will still be Gaussian, and we can prove just by working out the recursion of mutual informations that

$$I(A^*; (A_t, Y_t(A_t))_{t=1}^n) \lesssim d + \log n + \frac{\max_{a\in\mathcal{A}}||a||_2^2 \mathrm{tr}(\Sigma_0)}{n}.$$

So it's easy to run Thompson sampling and we will get a reasonable regret bound as well which doesn't need to depend on the size of the action set – it's something like $d\sigma\sqrt{n} + \sigma\sqrt{n\log n} + \mathrm{diam}(\mathcal{A})\sqrt{\mathrm{tr}(\Sigma_0)}$. Intuitively, what's happening

is that we have a burn-in period to nail the variance of our prior, and then everything is a bit of wiggle room in sampling the direction.

In our remaining time, we'll try to prove some lower bounds: we'll do a particular case where we are interested in the gap between actions we play and the best possible action. We'll switch notation and let $y(x) = \langle \theta, x \rangle + \xi$ and define

$$\text{gap}_{P_\theta}(a) = \sup_{x \in \mathcal{A}} \langle \theta, x \rangle - \langle \theta, a \rangle$$

(that is, how suboptimal was our action compared to the best we could have played). The minimax risk we will consider now is

$$\mathcal{M}_n = \inf_{\rho_1, \cdots, \rho_n, \rho_{n+1}} \sup_P \mathbb{E}_P \left[\text{gap}_P(A_{n+1}) \right],$$

meaning that we can do whatever we want for the first n steps and only measure the loss on the last step, which is a much easier quantity to upper bound but harder to lower bound – in particular this is always at most $\frac{1}{n}\overline{\text{Reg}}_n$ since we could have just taken the best action out of our first n attempts.

Theorem 207

If $P = \{P_\theta\}_{||\theta||_2 \le 1}$ is the set of linear bandits and the action set satisfies $\{-1, 1\}^d \subset \mathcal{A} \subset [-1, 1]^d$, then

$$\mathcal{M}_n \ge \frac{1}{8} \min\left(\frac{d\sigma}{\sqrt{n}}, \sqrt{d} \right).$$

Proof. Recall that Assouad's lemma embeds a problem into identifying corners of a hypercube, showing that if we play an action which identifies the wrong corners, then we must lose something significant. So we need to show a Hamming separation, which we do via the following separation lemma (the reason for our assumption on $\mathcal{A}$ is just to make this part notationally easier):

Lemma 208

Let $a^*(\theta) = \text{argmax}_{a \in \mathcal{A}} \langle a, \theta \rangle$ be the optimal action. Then for any $\theta \in \mathbb{R}^d$, the gap satisfies

$$\text{gap}_{P_\theta}(a) = \langle \theta, a^*(\theta) - a \rangle \ge \sum_{j=1}^d |\theta_j| 1\{\text{sign}(a_j) \ne \text{sign}(a_j^*)\}.$$

(Indeed, the optimal action $a^*(\theta)$ is exactly going to be $\text{sgn}(\theta)$, and so the gap is $\sum_{j=1}^d |\theta_j| - a_j \theta_j$. Then if a_j is of the wrong sign then the corresponding term in the summand is at least $|a_j|$, and otherwise it is at least zero.) So once we have a per-coordinate penalty, we can get our bounds: we pick our set of parameters and identify everything with the hypercube, so we will work with

$$\theta_v = \frac{\delta v}{\sqrt{d}}, \quad v \in \{\pm 1\}^d$$

for some $\delta \in [0, 1]$ to be chosen later. Assouad's method then says that if $P_{v, \pm j}$ are the distributions of responses $Y_1, \cdots, Y_n$ observed under θ_v with coordinate v_j fixed to be ± 1, then

$$\mathcal{M}_n \ge \frac{d}{2} \cdot \frac{\delta}{\sqrt{d}} \cdot \left(1 - \sqrt{\frac{1}{d} \frac{1}{2^d} \sum_{j=1}^d \sum_{v \in \{\pm 1\}^d} ||P_{v+j} - P_{v-j}||_{\text{TV}}^2} \right)$$

(since $\frac{\delta}{\sqrt{d}}$ is the per-coordinate separation). But now all we need to do is upper bound each variation distance, and we see how powerful things are when dealing with extremely interactive settings (even though everything depends on our

history): letting P_i be the distribution of the ith observation given the whole history $Y_i|H_{i-1}$ under P_{v+j}, and similarly let Q_i be the distribution of $Y_i|H_{i-1}$ under P_{v-j}.

Lemma 209

Suppose all actions a lie in a ball $||a||_2 \leq D$. Then

$$\frac{1}{d}\frac{1}{2^d}\sum_j \sum_v ||P_{v+j} - P_{v-j}||_{\text{TV}}^2 \leq \frac{n\delta^2}{\sigma^2 d^2} D^2.$$

Proof of lemma. By Pinsker we can just work with KL-divergences instead; we have

$$D_{\text{KL}}(P_{v+j}||P_{v-j}) = \sum_{i=1}^n \mathbb{E}_{P_{v+j}}\left[D_{\text{KL}}(P_i||Q_i)\right]$$

(where P_i, Q_i are random distributions depending on the history). But now the distribution of the ith action is a function of the past history, and so A_i actually has the same distribution under P_i and Q_i. Then Y_i is normal $N(\langle\theta_+, A_i\rangle, \sigma^2)$ under P_i and similarly $N(\langle\theta_-, A_i\rangle, \sigma^2)$ under Q_i, where θ_+ is θ with the jth coordinate set to be positive and similarly θ_- the same with it set to be negative. So we can just calculate the KL divergence exactly, getting that

$$D_{\text{KL}}(P_{v+j}||P_{v-j}) = \sum_{i=1}^n \frac{1}{2\sigma^2}\mathbb{E}_P\left[(\langle A_i, \theta_+\rangle - \langle A_i, \theta_-\rangle)^2\right] = \frac{2}{\sigma^2}\sum_{i=1}^n \frac{\delta^2}{d}\mathbb{E}[(A_i)_j^2].$$

Summing over all j coordinates we thus get a bound of $\frac{2\delta^2}{\sigma^2 d}\sum_{i=1}^n \mathbb{E}[||A_i||_2^2]$. □

So from here we just plug things in to get our result: the minimax bound is

$$\mathcal{M}_n \geq \frac{\sqrt{d}\delta}{2}\left(1 - \sqrt{\frac{n\delta^2 D^2}{\sigma^2 d^2}}\right)$$

and choose $\delta^2 = \min\left(1, \frac{\sigma d}{2D\sqrt{n}}\right)$ and use that $D^2 = d$, which yields our result. □

So as we've seen many times, solving the problem is like solving coordinate-wise testing problems, and we choose our separation parameters so that we have constant probability of error on average under testing.