Chapter 2

Minimax lower bounds: the Fano and Le Cam methods

Understanding the fundamental limits of estimation and optimization procedures is important for a multitude of reasons. Indeed, developing bounds on the performance of procedures can give complementary insights. By exhibiting fundamental limits of performance (perhaps over restricted classes of estimators), it is possible to guarantee that an algorithm we have developed is optimal, so that searching for estimators with better statistical performance will have limited returns, though searching for estimators with better performance in other metrics may be interesting. Moreover, exhibiting refined lower bounds on the performance of estimators can also suggest avenues for developing alternative, new optimal estimators; lower bounds need not be a fully pessimistic exercise.

In this set of notes, we define and then discuss techniques for lower-bounding the minimax risk, giving three standard techniques for deriving minimax lower bounds that have proven fruitful in a variety of estimation problems [21]. In addition to reviewing these standard techniques—the Le Cam, Fano, and Assouad methods—we present a few simplifications and extensions that may make them more “user friendly.”

2.1 Basic framework and minimax risk

Our first step here is to establish the minimax framework we use. When we study classical estimation problems, we use a standard version of minimax risk; we will also show how minimax bounds can be used to study optimization problems, in which case we use a specialization of the general minimax risk that we call minimax excess risk (while minimax risk handles this case, it is important enough that we define additional notation).

Let us begin by defining the standard minimax risk, deferring temporarily our discussion of minimax excess risk. Throughout, we let $\mathcal{P}$ denote a class of distributions on a sample space $\mathcal{X}$, and let $\theta: \mathcal{P} \to \Theta$ denote a function defined on $\mathcal{P}$, that is, a mapping $P \mapsto \theta(P)$. The goal is to estimate the parameter $\theta(P)$ based on observations $X_i$ drawn from the (unknown) distribution $P$. In certain cases, the parameter $\theta(P)$ uniquely determines the underlying distribution; for example, if we attempt to estimate a normal mean $\theta$ from the family $\mathcal{P} = \{N(\theta, \sigma^2) : \theta \in \mathbb{R}\}$ with known variance $\sigma^2$, then $\theta(P) = E_P[X]$ uniquely determines distributions in $\mathcal{P}$. In other scenarios, however, $\theta$ does not uniquely determine the distribution: for instance, we may be given a class of densities $\mathcal{P}$ on the unit interval $[0, 1]$, and we wish to estimate $\theta(P) = \int_0^1 (p'(t))^2 dt$, where $p$ is the
density of $P$. In this case, $\theta$ does not parameterize $P$, so we take a slightly broader viewpoint of estimating functions of distributions in these notes.

The space $\Theta$ in which the parameter $\theta(P)$ takes values depends on the underlying statistical problem; as an example, if the goal is to estimate the univariate mean $\theta(P) = \mathbb{E}_P[X]$, we have $\Theta \subset \mathbb{R}$. To evaluate the quality of an estimator $\hat{\theta}$, we let $\rho : \Theta \times \Theta \to \mathbb{R}_+$ denote a (semi)metric on the space $\Theta$, which we use to measure the error of an estimator for the parameter $\theta$, and let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function with $\Phi(0) = 0$ (for example, $\Phi(t) = t^2$).

For a distribution $P \in \mathcal{P}$, we assume we receive i.i.d. observations $X_i$ drawn according to some $P$, and based on these $\{X_i\}$, the goal is to estimate the unknown parameter $\theta(P) \in \Theta$. For a given estimator $\hat{\theta}$—a measurable function $\hat{\theta} : \mathcal{X}^n \to \Theta$—we assess the quality of the estimate $\hat{\theta}(X_1, \ldots, X_n)$ in terms of the risk

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

For instance, for a univariate mean problem with $\rho(\theta, \theta') = |\theta - \theta'|$ and $\Phi(t) = t^2$, this risk is the mean-squared error. As the distribution $P$ is varied, we obtain the risk functional for the problem, which gives the risk of any estimator $\hat{\theta}$ for the family $\mathcal{P}$.

For any fixed distribution $P$, there is always a trivial estimator of $\theta(P)$: simply return $\theta(P)$, which will have minimal risk. Of course, this “estimator” is unlikely to be good in any real sense, and it is thus important to consider the risk functional not in a pointwise sense (as a function of individual $P$) but to take a more global view. One approach to this is Bayesian: we place a prior $\pi$ on the set of possible distributions $\mathcal{P}$, viewing $\theta(P)$ as a random variable, and evaluate the risk of an estimator $\hat{\theta}$ taken in expectation with respect to this prior on $P$. Another approach, first suggested by Wald [19], which is to choose the estimator $\hat{\theta}$ minimizing the maximum risk

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \Phi(\rho(\hat{\theta}(X_1, \ldots, X_n), \theta(P))) \right].$$

An optimal estimator for this metric then gives the minimax risk, which is defined as

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

(2.1.1)

where we take the supremum (worst-case) over distributions $P \in \mathcal{P}$, and the infimum is taken over all estimators $\hat{\theta}$. Here the notation $\theta(\mathcal{P})$ indicates that we consider parameters $\theta(P)$ for $P \in \mathcal{P}$ and distributions in $\mathcal{P}$.

In some scenarios, we study a specialized notion of risk appropriate for optimization problems (and statistical problems in which all we care about is prediction). In these settings, we assume there exists some loss function $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$, where for an observation $x \in \mathcal{X}$, the value $\ell(\theta; x)$ measures the instantaneous loss associated with using $\theta$ as a predictor. In this case, we define the risk

$$R_P(\theta) := \mathbb{E}_P[\ell(\theta; X)] = \int_{\mathcal{X}} \ell(\theta; x) dP(x)$$

(2.1.2)

as the expected loss of the vector $\theta$. (See, e.g., Chapter 5 of the lectures by Shapiro, Dentcheva, and Ruszczyński [17], or work on stochastic approximation by Nemirovski et al. [15].)

Such problems arise, for example, in estimating the uniformity of the distribution of a species over an area (large $\theta(P)$ indicates an irregular distribution).
Example 2.1 (Support vector machines): In linear classification problems, we observe pairs \( z = (x, y) \), where \( y \in \{1, -1\} \) and \( x \in \mathbb{R}^d \), and the goal is to find a parameter \( \theta \in \mathbb{R}^d \) so that \( \text{sign}(\langle \theta, x \rangle) = y \). A convex loss surrogate for this problem is the hinge loss \( \ell(\theta; z) = \max(1 - y \langle \theta, x \rangle, 0) \); minimizing the associated risk functional (2.1.2) over a set \( \Theta = \{ \theta \in \mathbb{R}^d : \|\theta\|_2 \leq r \} \) gives the support vector machine [3]. ♣

Example 2.2 (Two-stage stochastic programming): In operations research, one often wishes to allocate resources to a set of locations \( \{1, \ldots, m\} \) before seeing demand for the resources. Suppose that the (unobserved) sample \( x \) consists of the pair \( x = (C, v) \), where \( C \in \mathbb{R}^{m \times m} \) corresponds to the prices of shipping a unit of material, so \( c_{ij} \geq 0 \) gives the cost of shipping from location \( i \) to \( j \), and \( v \in \mathbb{R}^m \) denotes the value (price paid for the good) at each location. Letting \( \theta \in \mathbb{R}^m \) denote the amount of resources allocated to each location, we formulate the loss as

\[
\ell(\theta; x) := \inf_{r \in \mathbb{R}^m, T \in \mathbb{R}^{m \times m}} \left\{ \sum_{i,j} c_{ij} T_{ij} - \sum_{i=1}^m v_i r_i \mid r_i = \theta_i + \sum_{j=1}^m T_{ji} - \sum_{j=1}^m T_{ij}, T_{ij} \geq 0, \sum_{j=1}^m T_{ij} \leq \theta_i \right\}.
\]

Here the variables \( T \) correspond to the goods transported to and from each location (so \( T_{ij} \) is goods shipped from \( i \) to \( j \)), and we wish to minimize the cost of our shipping and maximize the profit. By minimizing the risk (2.1.2) over a set \( \Theta = \{ \theta \in \mathbb{R}^m : \sum_i \theta_i \leq b \} \), we maximize our expected reward given a budget constraint \( b \) on the amount of allocated resources. ♣

For a (potentially random) estimator \( \hat{\theta} : \mathcal{X}^n \to \Theta \) given access to a sample \( X_1, \ldots, X_n \), we may define the associated maximum excess risk for the family \( \mathcal{P} \) by

\[
\mathcal{M}_n(\Theta, \mathcal{P}, \ell) := \inf_{\hat{\theta} \in \Theta} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[ R_P(\hat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} R_P(\theta) \right\},
\]

where the expectation is taken over \( X_i \) and any randomness in the procedure \( \hat{\theta} \). This expression captures the difference between the (expected) risk performance of the procedure \( \hat{\theta} \) and the best possible risk, available if the distribution \( P \) were known ahead of time. The minimax excess risk, defined with respect to the loss \( \ell \), domain \( \Theta \), and family \( \mathcal{P} \) of distributions, is then defined by the best possible maximum excess risk,

\[
\mathcal{M}_n(\Theta, \mathcal{P}, \ell) := \inf_{\hat{\theta} \in \Theta} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[ R_P(\hat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} R_P(\theta) \right\}.
\]

2.2 Preliminaries on methods for lower bounds

There are a variety of techniques for providing lower bounds on the minimax risk (2.1.1). Each of them transforms the maximum risk by lower bounding it via a Bayesian problem (e.g. [11, 13, 14]), then proving a lower bound on the performance of all possible estimators for the Bayesian problem (it is often the case that the worst case Bayesian problem is equivalent to the original minimax
problem \[13\]. In particular, let \( \{P_v\} \subset \mathcal{P} \) be a collection of distributions in \( \mathcal{P} \) indexed by \( v \) and \( \pi \) be any probability mass function over \( v \). Then for any estimator \( \hat{\theta} \), the maximum risk has lower bound

\[
\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \Phi(\rho(\hat{\theta}(X^n), \theta(P))) \right] \geq \sum_v \pi(v) \mathbb{E}_{P_v} \left[ \Phi(\rho(\hat{\theta}(X^n), \theta(P_v))) \right].
\]

While trivial, this lower bound serves as the departure point for each of the subsequent techniques for lower bounding the minimax risk.

2.2.1 From estimation to testing

A standard first step in proving minimax bounds is to “reduce” the estimation problem to a testing problem \[21, 20, 18\]. The idea is to show that estimation risk can be lower bounded by the probability of error in testing problems, which we can develop tools for. We use two types of testing problems: one a multiple hypothesis test, the second based on multiple binary hypothesis tests, though we defer discussion of the second.

Given an index set \( V \) of finite cardinality, consider a family of distributions \( \{P_v\}_{v \in V} \) contained within \( \mathcal{P} \). This family induces a collection of parameters \( \{\theta(P_v)\}_{v \in V} \); we call the family a \( 2\delta \)-packing in the \( \rho \)-semimetric if

\[
\rho(\theta(P_v), \theta(P_{v'})) \geq 2\delta \quad \text{for all } v \neq v'.
\]

We use this family to define the canonical hypothesis testing problem:

- first, nature chooses \( V \) according to the uniform distribution over \( V \);
- second, conditioned on the choice \( V = v \), the random sample \( X = X^n = (X_1, \ldots, X_n) \) is drawn from the \( n \)-fold product distribution \( P^n_v \).

Given the observed sample \( X \), the goal is to determine the value of the underlying index \( v \). We refer to any measurable mapping \( \Psi : \mathcal{X}^n \to V \) as a test function. Its associated error probability is \( P(\Psi(X^n) \neq V) \), where \( P \) denotes the joint distribution over the random index \( V \) and \( X \). In particular, if we set \( \mathcal{P} = \frac{1}{|V|} \sum_{v \in V} P_v \) to be the mixture distribution, then the sample \( X \) is drawn (marginally) from \( \mathcal{P} \), and our hypothesis testing problem is to determine the randomly chosen index \( V \) given a sample from this mixture \( \mathcal{P} \).

With this setup, we obtain the classical reduction from estimation to testing.

**Proposition 2.3.** The minimax error (2.1.1) has lower bound

\[
\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} P(\Psi(X_1, \ldots, X_n) \neq V),
\]

where the infimum ranges over all testing functions.

**Proof** To see this result, fix an arbitrary estimator \( \hat{\theta} \). Suppressing dependence on \( X \) throughout the derivation, first note that it is clear that for any fixed \( \theta \), we have

\[
\mathbb{E}[\Phi(\rho(\hat{\theta}, \theta))] \geq \mathbb{E} \left[ \Phi(\delta) 1 \left\{ \rho(\hat{\theta}, \theta) \geq \delta \right\} \right] = \Phi(\delta) P(\rho(\hat{\theta}, \theta) \geq \delta),
\]

where the final inequality follows because \( \Phi \) is non-decreasing. Now, let us define \( \theta_v = \theta(P_v) \), so that \( \rho(\theta_v, \theta_{v'}) \geq 2\delta \) for \( v \neq v' \). By defining the testing function

\[
\Psi(\hat{\theta}) := \arg\min_{v \in V} \{\rho(\hat{\theta}, \theta_v)\},
\]

we have

\[
\mathbb{E}[\Phi(\rho(\hat{\theta}, \theta))] \geq \mathbb{E} \left[ \Phi(\delta) 1 \left\{ \rho(\hat{\theta}, \theta_v) \geq \delta \right\} \right] = \Phi(\delta) P(\rho(\hat{\theta}, \theta_v) \geq \delta),
\]

for each \( v \) and

\[
\mathbb{E}[\Phi(\rho(\hat{\theta}, \theta))] \geq \sum_v \pi(v) \mathbb{E}_{P_v} \left[ \Phi(\delta) 1 \left\{ \rho(\hat{\theta}, \theta) \geq \delta \right\} \right] = \Phi(\delta) \mathbb{E}[\Phi(\rho(\hat{\theta}, \theta))].
\]
Figure 2.1. Example of a 2\(\delta\)-packing of a set. The estimate \(\hat{\theta}\) is contained in at most one of the \(\delta\)-balls around the points \(\theta_v\).

breaking ties arbitrarily, we have that \(\rho(\hat{\theta}, \theta_v) < \delta\) implies that \(\Psi(\hat{\theta}) = v\) because of the triangle inequality and 2\(\delta\)-separation of the set \(\{\theta_v\}_{v \in V}\). Indeed, assume that \(\rho(\hat{\theta}, \theta_v) < \delta\); then for any \(v' \neq v\), we have

\[
\rho(\hat{\theta}, \theta_{v'}) \geq \rho(\theta_v, \theta_{v'}) - \rho(\hat{\theta}, \theta_v) > 2\delta - \delta = \delta.
\]

The test must thus return \(v\) as claimed. Equivalently, for \(v \in V\), the inequality \(\Psi(\hat{\theta}) \neq v\) implies \(\rho(\hat{\theta}, \theta_v) \geq \delta\). (See Figure 2.1) By averaging over \(V\), we find that

\[
\sup_F \mathbb{P}(\rho(\hat{\theta}, \theta(P)) \geq \delta) \geq \frac{1}{|V|} \sum_{v \in V} \mathbb{P}(\rho(\hat{\theta}, \theta(P_v)) \geq \delta \mid V = v) \geq \frac{1}{|V|} \sum_{v \in V} \mathbb{P}(\Psi(\hat{\theta}) \neq v \mid V = v).
\]

Taking an infimum over all tests \(\Psi : \mathcal{X}^n \to V\) gives inequality (2.2.1).

The remaining challenge is to lower bound the probability of error in the underlying multi-way hypothesis testing problem, which we do by choosing the separation \(\delta\) to trade off between the loss \(\Phi(\delta)\) (large \(\delta\) increases the loss) and the probability of error (small \(\delta\), and hence separation, makes the hypothesis test harder). Usually, one attempts to choose the largest separation \(\delta\) that guarantees a constant probability of error. There are a variety of techniques for this, and we present three: Le Cam’s method, Fano’s method, and Assouad’s method, including extensions of the latter two to enhance their applicability. Before continuing, however, we review some inequalities between divergence measures defined on probabilities, which will be essential for our development, and concepts related to packing sets (metric entropy, covering numbers, and packing).

### 2.2.2 Inequalities between divergences and product distributions

We now present a few inequalities, and their consequences when applied to product distributions, that will be quite useful for proving our lower bounds. The three divergences we relate are the total variation distance, Kullback-Leibler divergence, and Hellinger distance, all of which are instances
of $f$-divergences. We first recall the definitions of the three when applied to distributions $P, Q$ on a set $\mathcal{X}$, which we assume have densities $p, q$ with respect to a base measure $\mu$. Then we recall the total variation distance \([1.2.3]\) is
\[
\|P - Q\|_{TV} := \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)| = \frac{1}{2} \int |p(x) - q(x)| d\mu(x),
\]
which is the $f$-divergence $D_f (P \| Q)$ generated by $f(t) = \frac{1}{2} |t - 1|$. The Hellinger distance is
\[
d_{\text{hel}}(P, Q)^2 := \int (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x),
\]
which is the $f$-divergence $D_f (P \| Q)$ generated by $f(t) = (\sqrt{t} - 1)^2$. We also recall the Kullback-Leibler (KL) divergence
\[
D_{\text{kl}} (P \| Q) := \int p(x) \log \frac{p(x)}{q(x)} d\mu(x),
\]
which is the $f$-divergence $D_f (P \| Q)$ generated by $f(t) = t \log t$.

We then have the following proposition, which relates the total variation distance to each of the other two divergences.

**Proposition 2.4.** The total variation distance satisfies the following relationships:

(a) For the Hellinger distance,
\[
\frac{1}{2} d_{\text{hel}}(P, Q)^2 \leq \|P - Q\|_{TV} \leq d_{\text{hel}}(P, Q) \sqrt{1 - d_{\text{hel}}(P, Q)^2/4}.
\]

(b) Pinsker’s inequality: for any distributions $P, Q$,
\[
\|P - Q\|_{TV}^2 \leq \frac{1}{2} D_{\text{kl}} (P \| Q).
\]

We defer the proof of Proposition [2.4] to Section [2.6.1], showing how it is useful because KL-divergence and Hellinger distance both are easier to manipulate on product distributions than is total variation. Specifically, consider the product distributions $P = P_1 \times \cdots \times P_n$ and $Q = Q_1 \times \cdots \times Q_n$. Then the KL-divergence satisfies the decoupling equality
\[
D_{\text{kl}} (P \| Q) = \sum_{i=1}^n D_{\text{kl}} (P_i \| Q_i),
\]
while the Hellinger distance satisfies
\[
d_{\text{hel}}(P, Q)^2 = \int \left( \sqrt{p_1(x_1) \cdots p_n(x_n)} - \sqrt{q_1(x_1) \cdots q_n(x_n)} \right)^2 d\mu(x_n)
\]
\[
= \int \left( \prod_{i=1}^n p_i(x_i) + \prod_{i=1}^n q_i(x_i) - 2 \sqrt{p_1(x_1) \cdots p_n(x_n) q_1(x_n) \cdots q_n(x_n)} \right) d\mu(x_n)
\]
\[
= 2 - 2 \prod_{i=1}^n \int \sqrt{p_i(x) q_i(x)} d\mu(x) = 2 - 2 \prod_{i=1}^n \left( 1 - \frac{1}{2} d_{\text{hel}}(P_i, Q_i)^2 \right).
\]
In particular, we see that for product distributions $P^n$ and $Q^n$, Proposition 2.4 implies that
\[
\|P^n - Q^n\|_{TV}^2 \leq \frac{1}{2} D_{KL}(P^n || Q^n) = \frac{n}{2} D_{KL}(P || Q)
\]
and
\[
\|P^n - Q^n\|_{TV} \leq d_{hel}(P^n, Q^n) \leq \sqrt{2 - 2(1 - d_{hel}(P, Q)^2)^n}.
\]
As a consequence, if we can guarantee that $D_{KL}(P || Q) \leq 1/n$ or $d_{hel}(P, Q) \leq 1/\sqrt{n}$, then we guarantee the strict inequality $\|P^n - Q^n\|_{TV} \leq 1 - c$ for a fixed constant $c > 0$, for any $n$. We will see how this type of guarantee can be used to prove minimax lower bounds in the following sections.

### 2.2.3 Metric entropy and packing numbers

The second part of proving our lower bounds involves the construction of the packing set in Section 2.2.1. The size of the space $\Theta$ of parameters associated with our estimation problem—and consequently, how many parameters we can pack into it—is strongly coupled with the difficulty of estimation. Given a non-empty set $\Theta$ with associated (semi)metric $\rho$, a natural way to measure the size of the set is via the number of balls of a fixed radius $\delta > 0$ required to cover it.

**Definition 2.1** (Covering number). Let $\Theta$ be a set with (semi)metric $\rho$. A $\delta$-cover of the set $\Theta$ with respect to $\rho$ is a set $\{\theta_1, \ldots, \theta_N\}$ such that for any point $\theta \in \Theta$, there exists some $v \in \{1, \ldots, N\}$ such that $\rho(\theta, \theta_v) \leq \delta$. The $\delta$-covering number of $\Theta$ is
\[
N(\delta, \Theta, \rho) := \inf \{N \in \mathbb{N} : \text{there exists a $\delta$-cover $\theta_1, \ldots, \theta_N$ of $\Theta$}\}.
\]

The **metric entropy**\[^{[12]}\] of the set $\Theta$ is simply the logarithm of its covering number $\log N(\delta, \Theta, \rho)$. We can define a related measure—more useful for constructing our lower bounds—of size that relates to the number of disjoint balls of radius $\delta > 0$ that can be placed into the set $\Theta$.

**Definition 2.2** (Packing number). A $\delta$-packing of the set $\Theta$ with respect to $\rho$ is a set $\{\theta_1, \ldots, \theta_M\}$ such that for all distinct $v, v' \in \{1, \ldots, M\}$, we have $\rho(\theta_v, \theta_{v'}) \geq \delta$. The $\delta$-packing number of $\Theta$ is
\[
M(\delta, \Theta, \rho) := \sup \{M \in \mathbb{N} : \text{there exists a $\delta$-packing $\theta_1, \ldots, \theta_M$ of $\Theta$}\}.
\]

An exercise in proof by contradiction shows that the packing and covering numbers of a set are in fact closely related:

**Lemma 2.5.** The packing and covering numbers satisfy the following inequalities:
\[
M(2\delta, \Theta, \rho) \leq N(\delta, \Theta, \rho) \leq M(\delta, \Theta, \rho).
\]

We leave derivation of this lemma to the reader, noting that it shows that (up to constant factors) packing and covering numbers have the same scaling in the radius $\delta$. As a simple example, we see for any interval $[a, b]$ on the real line that in the usual absolute distance metric, $N(\delta, [a, b], | \cdot |) \propto (b - a)/\delta$.

We can now provide a few more complex examples of packing and covering numbers, presenting two standard results that will be useful for constructing the packing sets used in our lower bounds to come. We remark in passing that these constructions are essentially identical to those used to construct well-separated code-books in communication; in showing our lower bounds, we show that even if a code-book is well-separated, it may still be hard to estimate. Our first bound shows that there are (exponentially) large packings of the $d$-dimensional hypercube of points that are $O(d)$-separated in the Hamming metric.
Lemma 2.6 (Gilbert-Varshamov bound). Let \( d \geq 1 \). There is a subset \( V \) of the \( d \)-dimensional hypercube \( \mathcal{H}_d = \{-1, 1\}^d \) of size \( |V| \geq \exp(d/8) \) such that the \( \ell_1 \)-distance

\[
\|v - v'\|_1 = 2 \sum_{j=1}^{d} \mathbf{1}\{v_j \neq v'_j\} \geq \frac{d}{2}
\]

for all \( v \neq v' \) with \( v, v' \in V \).

**Proof** We use the proof of Guntuboyina [9]. Consider a maximal subset \( V \) of \( \mathcal{H}_d = \{-1, 1\}^d \) satisfying

\[
\|v - v'\|_1 \geq d/2 \quad \text{for all distinct } v, v' \in V. \tag{2.2.6}
\]

That is, the addition of any vector \( w \in \mathcal{H}_d, w \notin V \) to \( V \) will break the constraint (2.2.6). This means that if we construct the closed balls \( B(v, d/2) := \{w \in \mathcal{H}_d : \|v - w\|_1 \leq d/2\} \), we must have

\[
\bigcup_{v \in V} B(v, d/2) = \mathcal{H}_d \quad \text{so} \quad |V||B(0, d/2)| = \sum_{v \in V} |B(v, d/2)| \geq 2^d. \tag{2.2.7}
\]

We now upper bound the cardinality of \( B(v, d/2) \) using the probabilistic method, which will imply the desired result. Let \( S_i, i = 1, \ldots, d \), be i.i.d. Bernoulli \( \{0, 1\} \)-valued random variables. Then by their uniformity, for any \( v \in \mathcal{H}_d \),

\[
2^{-d}|B(v, d/2)| = \mathbb{P}(S_1 + S_2 + \ldots + S_d \leq d/4) = \mathbb{P}(S_1 + S_2 + \ldots + S_d \geq 3d/4) \leq \mathbb{E}[\exp(\lambda S_1 + \ldots + \lambda S_d)] \exp(-3\lambda d/4)
\]

for any \( \lambda > 0 \), by Markov’s inequality (or the Chernoff bound). Since \( \mathbb{E}[\exp(\lambda S_1)] = \frac{1}{2}(1 + e^\lambda) \), we obtain

\[
2^{-d}|B(v, d/2)| \leq \inf_{\lambda \geq 0} \left\{ 2^{-d}(1 + e^\lambda)^d \exp(-3\lambda d/4) \right\}
\]

Choosing \( \lambda = \log 3 \), we have

\[
|B(v, d/2)| \leq 4^d \exp(-(3/4)d \log 3) = 3^{-3d/4}4^d.
\]

Recalling inequality (2.2.7), we have

\[
|V|3^{-3d/4}4^d \geq |V||B(v, d/2)| \geq 2^d, \quad \text{or} \quad |V| \geq \frac{3^{3d/4}2^d}{2^d} = \exp\left(d \left[ \frac{3}{4} \log 3 - \log 2 \right] \right) \geq \exp(d/8),
\]

as claimed. \( \square \)

Given the relationships between packing, covering, and size of sets \( \Theta \), we would expect there to be relationships between volume, packing, and covering numbers. This is indeed the case, as we now demonstrate for arbitrary norm balls in finite dimensions.

**Lemma 2.7.** Let \( B \) denote the unit \( \|\cdot\| \)-ball in \( \mathbb{R}^d \). Then

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

17
As a consequence of Lemma 2.7, we see that for any \( \delta < 1 \), there is a packing \( \mathcal{V} \) of \( \mathbb{B} \) such that
\[
\|v - v'\| \geq \delta \text{ for all distinct } v, v' \in \mathcal{V} \text{ and } |\mathcal{V}| \geq (1/\delta)^d,
\]
because we know \( M(\delta, \mathbb{B}, \|\cdot\|) \geq N(\delta, \mathbb{B}, \|\cdot\|) \) as in Lemma 2.5. In particular, the lemma shows that any norm ball has a \( \frac{1}{2} \)-packing in its own norm with cardinality at least \( 2^d \). We can also construct exponentially large packings of arbitrary norm-balls (in finite dimensions) where points are of constant distance apart.

**Proof** We prove the lemma via a volumetric argument. For the lower bound, note that if the points \( v_1, \ldots, v_N \) are a \( \delta \)-cover of \( \mathbb{B} \), then
\[
\text{Vol}(\mathbb{B}) \leq \sum_{i=1}^{N} \text{Vol}(\delta \mathbb{B} + v_i) = N \text{Vol}(\delta \mathbb{B}) = N \text{Vol}(\mathbb{B}) \delta^d.
\]
In particular, \( N \geq \delta^{-d} \). For the upper bound on \( N(\delta, \mathbb{B}, \|\cdot\|) \), let \( \mathcal{V} \) be a \( \delta \)-packing of \( \mathbb{B} \) with maximal cardinality, so that \( |\mathcal{V}| = M(\delta, \mathbb{B}, \|\cdot\|) \geq N(\delta, \mathbb{B}, \|\cdot\|) \) (recall Lemma 2.5). Notably, the collection of \( \delta \)-balls \( \{\delta \mathbb{B} + v_i\}_{i=1}^{M} \) cover the ball \( \mathbb{B} \) (as otherwise, we could put an additional element in the packing \( \mathcal{V} \)), and moreover, the balls \( \{\frac{1}{2} \mathbb{B} + v_i\} \) are all disjoint by definition of a packing. Consequently, we find that
\[
M\left(\frac{\delta}{2}\right)^d \text{Vol}(\mathbb{B}) = M \text{Vol}\left(\frac{\delta}{2} \mathbb{B}\right) \leq \text{Vol}\left(\mathbb{B} + \frac{\delta}{2} \mathbb{B}\right) = \left(1 + \frac{\delta}{2}\right)^d \text{Vol}(\mathbb{B}).
\]
Rewriting, we obtain
\[
M(\delta, \mathbb{B}, \|\cdot\|) \leq \left(\frac{2}{\delta}\right)^d \left(1 + \frac{\delta}{2}\right)^d \frac{\text{Vol}(\mathbb{B})}{\text{Vol}(\mathbb{B})} = \left(1 + \frac{2}{\delta}\right)^d,
\]
completing the proof.

### 2.3 Le Cam’s method

Le Cam’s method, in its simplest form, provides lower bounds on the error in simple binary hypothesis testing problems. In this section, we explore this connection, showing the connection between hypothesis testing and total variation distance, and then show how this can yield lower bounds on minimax error (or the optimal Bayes’ risk) for simple—often one-dimensional—estimation problems.

In the first homework, we considered several representations of the total variation distance, including a question showing its relation to optimal testing. We begin again with this strand of thought, recalling the general testing problem discussed in Section 2.2.1. Suppose that we have a Bayesian hypothesis testing problem where \( V \) is chosen with equal probability to be 1 or 2, and given \( V = v \), the sample \( X \) is drawn from the distribution \( P_v \). Denoting by \( \mathbb{P} \) the joint distribution of \( V \) and \( X \), we have for any test \( \Psi : \mathcal{X} \to \{1, 2\} \) that the probability of error is
\[
\mathbb{P}(\Psi(X) \neq V) = \frac{1}{2} P_1(\Psi(X) \neq 1) + \frac{1}{2} P_2(\Psi(X) \neq 2).
\]

We can give an exact expression for the minimal possible error in the above hypothesis test. Indeed, a standard result of Le Cam (see [13, 21, Lemma 1]) is the following variational representation of the total variation distance as a function of testing error.
Proposition 2.8. For any distributions \( P_1 \) and \( P_2 \) on \( \mathcal{X} \), we have

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

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

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

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

Taking an infimum over such acceptance regions, we have

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

which yields the total variation distance as desired. \( \square \)

Returning to the setting in which we receive \( n \) i.i.d. observations \( X_1, \ldots, X_n \sim P \), when \( V = 1 \) with probability \( \frac{1}{2} \) and 2 with probability \( \frac{1}{2} \), we have

\[
\inf_{\Psi} \mathbb{P}(\Psi(X_1, \ldots, X_n) \neq V) = 1 - \frac{1}{2} \| P^n_1 - P^n_2 \|_{TV}.
\]

The representations (2.3.1) and (2.3.2), in conjunction with our reduction of estimation to testing in Proposition 2.3 imply the following lower bound on minimax risk. For any family \( \mathcal{P} \) of distributions for which there exists a pair \( P_1, P_2 \in \mathcal{P} \) satisfying \( \rho(\theta(P_1), \theta(P_2)) \geq 2\delta \), then the minimax risk after \( n \) observations has lower bound

\[
\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \left[ \frac{1}{2} - \frac{1}{2} \| P^n_1 - P^n_2 \|_{TV} \right].
\]

The lower bound (2.3.3) suggests the following strategy: we find distributions \( P_1 \) and \( P_2 \), which we choose as a function of \( \delta \), that guarantee \( \| P^n_1 - P^n_2 \|_{TV} \leq \frac{1}{2} \). In this case, so long as \( \rho(\theta(P_1), \theta(P_2)) \geq 2\delta \), we have the lower bound

\[
\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \left[ \frac{1}{2} - \frac{1}{2} \cdot \frac{1}{4} \right] = \frac{1}{4} \Phi(\delta).
\]

We now give an example illustrating this idea.

**Example 2.9** (Bernoulli mean estimation): Consider the problem of estimating the mean \( \theta \in [-1, 1] \) of a \( \{\pm 1\} \)-valued Bernoulli distribution under the squared error loss \( (\theta - \hat{\theta})^2 \), where \( X_i \in \{-1, 1\} \). In this case, by fixing some \( \delta > 0 \), we set \( V = \{-1, 1\} \), and we define \( P_v \) so that

\[
P_v(X = 1) = \frac{1 + v\delta}{2} \quad \text{and} \quad P_v(X = -1) = \frac{1 - v\delta}{2},
\]

whence we see that the mean \( \theta(P_v) = \delta v \). Using the metric \( \rho(\theta, \theta') = |\theta - \theta'| \) and loss \( \Phi(\delta) = \delta^2 \), we have separation \( 2\delta \) of \( \theta(P_{-1}) \) and \( \theta(P_1) \). Thus, via Le Cam’s method (2.3.3), we have that

\[
\mathcal{M}_n(\text{Bernoulli}([-1, 1]), \cdot^2) \geq \frac{1}{2} \delta^2 \left( 1 - \| P^n_{-1} - P^n_1 \|_{TV} \right).
\]
We would thus like to upper bound \( \|P_n^1 - P_n^2\|_{TV} \) as a function of the separation \( \delta \) and sample size \( n \); here we use Pinsker’s inequality (Proposition 2.3(b)) and the tensorization identity (2.2.4) that makes KL-divergence so useful. Indeed, we have

\[
\|P_n^1 - P_n^2\|_{TV}^2 \leq \frac{1}{2} D_{kl}(P_n^1 | P_n^2) = \frac{n}{2} D_{kl}(P_1 | P_1) = \frac{n}{2} \delta \log \frac{1 + \delta}{1 - \delta}.
\]

Noting that \( \delta \log \frac{1 + \delta}{1 - \delta} \leq 3\delta^2 \) for \( \delta \in [0, 1/2] \), we obtain that \( \|P_n^1 - P_n^2\|_{TV} \leq \delta \sqrt{3n/2} \) for \( \delta \leq 1/2 \). In particular, we can guarantee a high probability of error in the associated hypothesis testing problem (recall inequality (2.3.2)) by taking \( \delta = 1/\sqrt{6n} \); this guarantees \( \|P_n^1 - P_n^2\|_{TV} \leq 1/2 \). We thus have the minimax lower bound

\[
\mathcal{M}_n(\text{Bernoulli}([-1, 1]), (\cdot)^2) \geq \frac{1}{2} \delta^2 \left(1 - \frac{1}{2}\right) = \frac{1}{24n}.
\]

While the factor 1/24 is smaller than necessary, this bound is optimal to within constant factors; the sample mean \((1/n) \sum_{i=1}^{n} X_i \) achieves mean-squared error \((1 - \theta^2)/n\).

As an alternative proof, we may use the Hellinger distance and its associated decoupling identity (2.2.5). We sketch the idea, ignoring lower order terms when convenient. In this case, Proposition 2.4(a) implies

\[
\|P_n^1 - P_n^2\|_{TV} \leq d_{hel}(P_n^1, P_n^2) = \sqrt{2 - 2(1 - d_{hel}(P_1, P_2)^2)^n}.
\]

Noting that

\[
d_{hel}(P_1, P_2)^2 = \left(\sqrt{\frac{1 + \delta}{2}} - \sqrt{\frac{1 - \delta}{2}}\right)^2 = 1 - 2\sqrt{\frac{1 - \delta^2}{4}} = 1 - \sqrt{1 - \delta^2} \approx \frac{1}{2} \delta^2,
\]

and noting that \((1 - \delta^2) \approx e^{-\delta^2}\), we have (up to lower order terms in \( \delta \)) that \( \|P_n^1 - P_n^2\|_{TV} \leq \sqrt{2 - 2\exp(-\delta^2n/2)} \). Choosing \( \delta^2 = 1/(4n) \), we have \( \sqrt{2 - 2\exp(-\delta^2n/2)} \leq 1/2 \), thus giving the lower bound

\[
\mathcal{M}_n(\text{Bernoulli}([-1, 1]), (\cdot)^2) \geq \frac{1}{2} \delta^2 \left(1 - \frac{1}{2}\right) = \frac{1}{16n},
\]

where the quotations indicate we have been fast and loose in the derivation. ♣

This example shows the “usual” rate of convergence in parametric estimation problems, that is, that we can estimate a parameter \( \theta \) at a rate (in squared error) scaling as \( 1/n \). The mean estimator above is, in some sense, the prototypical example of such regular problems. In some “irregular” scenarios—including estimating the support of a uniform random variable, which we study in the homework—faster rates are possible.

We also note in passing that there are substantially more complex versions of Le Cam’s method that can yield sharp results for a wider variety of problems, including some in nonparametric estimation [13, 21]. For our purposes, the simpler two-point perspective provided in this section will be sufficient.
2.4 Fano’s method

Fano’s method, originally proposed by Has’minskii [10] for providing lower bounds in nonparametric estimation problems, gives a somewhat more general technique than Le Cam’s method, and it applies when the packing set \( V \) has cardinality larger than two. The method has played a central role in minimax theory, beginning with the pioneering work of Has’minskii and Ibragimov [10, 11]. More recent work following this initial push continues to the present day (e.g. [1, 21, 20, 2, 16, 9, 4]).

2.4.1 The classical (local) Fano method

We begin by stating Fano’s inequality, which provides a lower bound on the error in a multi-way hypothesis testing problem. Let \( V \) be a random variable taking values in a finite set \( V \) with cardinality \(|V| \geq 2\). If we let the function \( h_2(p) = -p \log p - (1 - p) \log(1 - p) \) denote the entropy of the Bernoulli random variable with parameter \( p \), Fano’s inequality takes the following form [e.g. 6, Chapter 2]:

\[
\text{Proposition 2.10 (Fano inequality). For any Markov chain } V \rightarrow X \rightarrow \hat{V}, \text{ we have}
\]

\[
h_2(\mathbb{P}(\hat{V} \neq V)) + \mathbb{P}(\hat{V} \neq V) \log(|V| - 1) \geq H(V | \hat{V}).
\]

(2.4.1)

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

\[
H(V, E | \hat{V}) = H(V | E, \hat{V}) + H(E | \hat{V}) = \mathbb{P}(E = 1)H(V | E = 1, \hat{V}) + \mathbb{P}(E = 0) \underbrace{H(V | E = 0, \hat{V}) + H(E | \hat{V})}_{=0},
\]

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

\[
H(V, E | \hat{V}) = H(V | \hat{V}) + H(E | \hat{V}, V),
\]

because \( E \) is perfectly predicted by \( \hat{V} \) and \( V \). Combining these equalities, we have

\[
H(V | \hat{V}) = H(V, E | \hat{V}) = \mathbb{P}(E = 1)H(V | E = 1, \hat{V}) + H(E | \hat{V}).
\]

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

We can rewrite Proposition 2.10 in a convenient way when \( V \) is uniform in \( V \). Indeed, by definition of the mutual information, we have \( I(V; \hat{V}) = H(V) - H(V | \hat{V}) \), so Proposition 2.10 implies that in the canonical hypothesis testing problem from Section 2.2.1 we have

**Corollary 2.11.** Assume that \( V \) is uniform on \( V \). For any Markov chain \( V \rightarrow X \rightarrow \hat{V} \),

\[
\mathbb{P}(\hat{V} \neq V) \geq 1 - \frac{I(V; X) + \log 2}{\log(|V|)}.
\]

(2.4.2)
Proof Let \( P_{\text{error}} = \mathbb{P}(V \neq \hat{V}) \) denote the probability of error. Noting that \( h_2(p) \leq \log 2 \) for any \( p \in [0, 1] \), then using Proposition 2.10 we have

\[
\log 2 + P_{\text{error}} \log(|V|) \geq h_2(P_{\text{error}}) + P_{\text{error}} \log(|V| - 1) \overset{(i)}{=} H(V | \hat{V}) = H(V) - I(V; \hat{V}),
\]

where step (i) uses Proposition 2.10 and step (ii) uses the definition of mutual information, that \( I(V; \hat{V}) = H(V) - H(V | \hat{V}) \). The data processing inequality implies that \( I(V; \hat{V}) \leq I(V; X) \), and using \( H(V) = \log(|V|) \) completes the proof.

In particular, Corollary 2.11 shows that we have

\[
\inf_{\Psi} \mathbb{P}(\Psi(X) \neq V) \geq 1 - \frac{I(V; X) + \log 2}{\log |V|},
\]

where the infimum is taken over all testing procedures \( \Psi \). By combining Corollary 2.11 with the reduction from estimation to testing in Proposition 2.3 we obtain the following result.

**Proposition 2.12.** Let \( \{\theta(P_v)\}_{v \in V} \) be a \( 2\delta \)-packing in the \( \rho \)-semimetric. Assume that \( V \) is uniform on the set \( V \), and conditional on \( V = v \), we draw a sample \( X \sim P_v \). Then the minimax risk has lower bound

\[
\mathfrak{M}(\theta(P); \Phi \circ \rho) \geq \Phi(\delta) \left( 1 - \frac{I(V; X) + \log 2}{\log |V|} \right).
\]

To gain some intuition for Proposition 2.12 we think of the lower bound as a function of the separation \( \delta > 0 \). Roughly, as \( \delta \downarrow 0 \), the separation condition between the distributions \( P_v \) is relaxed and we expect the distributions \( P_v \) to be closer to one another. In this case—as will be made more explicitly presently—the hypothesis testing problem of distinguishing the \( P_v \) becomes more challenging, and the information \( I(V; X) \) shrinks. Thus, what we roughly attempt to do is to choose our packing \( \theta(P_v) \) as a function of \( \delta \), and find the largest \( \delta > 0 \) making the mutual information small enough that

\[
\frac{I(V; X) + \log 2}{\log |V|} \leq \frac{1}{2}. \quad (2.4.3)
\]

In this case, the minimax lower bound is at least \( \Phi(\delta)/2 \). We now explore techniques for achieving such results.

**Mutual information and KL-divergence**

Many techniques for upper bounding mutual information rely on its representation as the KL-divergence between multiple distributions. Indeed, given random variables \( V \) and \( X \) as in the preceding sections, if we let \( \mathbb{P}_{V,X} \) denote their joint distribution and \( \mathbb{P}_V \) and \( \mathbb{P}_X \) their marginals, then

\[
I(V; X) = D_{\text{kl}}(\mathbb{P}_{X,V} \| \mathbb{P}_X \mathbb{P}_V),
\]

where \( \mathbb{P}_X \mathbb{P}_V \) denotes the distribution of \( (X, V) \) when the random variables are independent. By manipulating this definition, we can rewrite it in a way that is a bit more convenient for our purposes.

Indeed, focusing on our setting of testing, let us assume that \( V \) is drawn from a prior distribution \( \pi \) (this may be a discrete or arbitrary distribution, though for simplicity we focus on the case when
\( \pi \) is discrete). Let \( P_v \) denote the distribution of \( X \) conditional on \( V = v \), as in Proposition 2.12. Then marginally, we know that \( X \) is drawn from the mixture distribution
\[
P := \sum_v \pi(v) P_v.
\]
With this definition of the mixture distribution, via algebraic manipulations, we have
\[
I(V; X) = \sum_v \pi(v) D_{\text{KL}}(P_v \| P),
\]
(2.4.4)
a representation that plays an important role in our subsequent derivations. To see equality (2.4.4), let \( \mu \) be a base measure over \( X \) (assume w.l.o.g. that \( X \) has density \( p(\cdot | v) = p_v(\cdot) \) conditional on \( V = v \)), and note that
\[
I(V; X) = \sum_v \int_X p(x | v) \pi(v) \log \frac{p(x | v)}{\overline{p}(x)} d\mu(x) = \sum_v \pi(v) \int_X p(x | v) \log \frac{p(x | v)}{p(x)} d\mu(x).
\]
Representation (2.4.4) makes it clear that if the distributions of the sample \( X \) conditional on \( V \) are all similar, then there is little information content. Returning to the discussion after Proposition 2.12, we have in this uniform setting that
\[
P = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \quad \text{and} \quad I(V; X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\text{KL}}(P_v \| P).
\]
The mutual information is small if the typical conditional distribution \( P_v \) is difficult to distinguish—has small KL-divergence—from \( P \).

**The local Fano method**

The local Fano method is based on a weakening of the mixture representation of mutual information (2.4.4), then giving a uniform upper bound on divergences between all pairs of the conditional distributions \( P_v \) and \( P_{v'} \). (This method is known in the statistics literature as the “generalized Fano method,” a poor name, as it is based on a weak upper bound on mutual information.) In particular (focusing on the case when \( V \) is uniform), the convexity of \(- \log\) implies that
\[
I(V; X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\text{KL}}(P_v \| P) \leq \frac{1}{|\mathcal{V}|^2} \sum_{v, v'} D_{\text{KL}}(P_v \| P_{v'}). 
\]
(2.4.5)
(In fact, the KL-divergence is *jointly* convex in its arguments; see Appendix 2.7 for a proof of this fact generalized to all \( f \)-divergences.)

In the local Fano method approach, we construct a local packing. This local packing approach is based on constructing a family of distributions \( P_v \) for \( v \in \mathcal{V} \) defining a \( 2\delta \)-packing (recall Section 2.2.1), meaning that \( \rho(\theta(P_v), \theta(P_{v'})) \geq 2\delta \) for all \( v \neq v' \), but which additionally satisfy the uniform upper bound
\[
D_{\text{KL}}(P_v \| P_{v'}) \leq \kappa^2 \delta^2 \quad \text{for all} \ v, v' \in \mathcal{V},
\]
(2.4.6)
where \( \kappa > 0 \) is a fixed problem-dependent constant. If we have the inequality (2.4.6), then so long as we can find a local packing \( \mathcal{V} \) such that
\[
\log |\mathcal{V}| \geq 2(\kappa^2 \delta^2 + \log 2),
\]

23
we are guaranteed the testing error condition (2.4.3), and hence the minimax lower bound
\[ \mathcal{M}(\theta(P), \Phi \circ \rho) \geq \frac{1}{2} \Phi(\delta). \]

The difficulty in this approach is constructing the packing set \( \mathcal{V} \) that allows \( \delta \) to be chosen to obtain sharp lower bounds, and we often require careful choices of the packing sets \( \mathcal{V} \). (We will see how to reduce such difficulties in subsequent sections.)

**Constructing local packings**  As mentioned above, the main difficulty in using Fano’s method is in the construction of so-called “local” packings. In these problems, the idea is to construct a packing \( \mathcal{V} \) of a fixed set (in a vector space, say \( \mathbb{R}^d \)) with constant radius and constant distance. Then we scale elements of the packing by \( \delta > 0 \), which leaves the cardinality \( |\mathcal{V}| \) identical, but allows us to scale \( \delta \) in the separation in the packing and the uniform divergence bound (2.4.6). In particular, Lemmas 2.6 and 2.7 show that we can construct exponentially large packings of certain sets with balls of a fixed radius.

We now illustrate these techniques via two examples.

**Example 2.13** (Normal mean estimation): Consider the \( d \)-dimensional normal location family \( \mathcal{N}_d = \{ N(\theta, \sigma^2 I_{d \times d}) \mid \theta \in \mathbb{R}^d \} \); we wish to estimate the mean \( \theta = \theta(P) \) of a given distribution \( P \in \mathcal{N}_d \) in mean-squared error, that is, with loss \( \| \hat{\theta} - \theta \|_2^2 \). Let \( \mathcal{V} \) be a 1/2-packing of the unit \( \ell_2 \)-ball with cardinality at least \( 2^d \), as guaranteed by Lemma 2.7. (We assume for simplicity that \( d \geq 2 \).)

Now we construct our local packing. Fix \( \delta > 0 \), and for each \( v \in \mathcal{V} \), set \( \theta_v = \delta v \in \mathbb{R}^d \). Then we have
\[ \| \theta_v - \theta_{v'} \|_2 = \delta \| v - v' \|_2 \geq \frac{\delta}{2} \]
for each distinct pair \( v, v' \in \mathcal{V} \), and moreover, we note that \( \| \theta_v - \theta_{v'} \|_2 \leq \delta \) for such pairs as well. By applying the Fano minimax bound of Proposition 2.12, we see that (given \( n \) normal observations \( X_i \overset{i.i.d.}{\sim} P \))
\[ \mathcal{M}_n(\theta(\mathcal{N}_d), \| \cdot \|_2^2) \geq \left( \frac{1}{2} \cdot \frac{\delta}{2} \right)^2 \left( 1 - \frac{I(V; X^n) + \log 2}{\log |\mathcal{V}|} \right) = \frac{\delta^2}{16} \left( 1 - \frac{I(V; X^n) + \log 2}{d \log 2} \right). \]

Now note that for any pair \( v, v' \), if \( P_v \) is the normal distribution \( N(\theta_v, \sigma^2 I_{d \times d}) \) we have
\[ D_{KL}(P_v \| P_{v'}) = n \cdot D_{KL}(N(\delta v, \sigma^2 I_{d \times d}) \| N(\delta v', \sigma^2 I_{d \times d})) = n \cdot \frac{\delta^2}{2\sigma^2} \| v - v' \|_2^2, \]
as the KL-divergence between two normal distributions with identical covariance is
\[ D_{KL}(N(\theta_1, \Sigma) \| N(\theta_2, \Sigma)) = \frac{1}{2} (\theta_1 - \theta_2)^\top \Sigma^{-1} (\theta_1 - \theta_2). \]  
(2.4.7)

(The equality (2.4.7) is left as an exercise for the reader.) As \( \| v - v' \|_2 \leq 1 \), we have the KL-divergence bound (2.4.6) with \( \kappa^2 = n/2\sigma^2 \).

Combining our derivations, we have the minimax lower bound
\[ \mathcal{M}_n(\theta(\mathcal{N}_d), \| \cdot \|_2^2) \geq \frac{\delta^2}{16} \left( 1 - \frac{n\delta^2/2\sigma^2 + \log 2}{d \log 2} \right). \]  
(2.4.8)
Then by taking \( \delta^2 = d\sigma^2 \log 2/(2n) \), we see that
\[
1 - \frac{n\delta^2/2\sigma^2 + \log 2}{d \log 2} = 1 - \frac{d}{d} - \frac{1}{4} \geq \frac{1}{4}
\]
by assumption that \( d \geq 2 \), and inequality (2.4.8) implies the minimax lower bound
\[
\mathfrak{M}_n(\theta(N_d), \|\cdot\|_2^2) \geq \frac{d\sigma^2 \log 2}{32n} \cdot \frac{1}{4} \geq \frac{1}{185} \cdot \frac{d\sigma^2}{n}.
\]
While the constant \( 1/185 \) is not sharp, we do obtain the right scaling in \( d, n \), and the variance \( \sigma^2 \); the sample mean attains the same risk. ♤

Example 2.14 (Linear regression): In this example, we show how local packings can give (up to some constant factors) sharp minimax rates for standard linear regression problems. In particular, for fixed matrix \( X \in \mathbb{R}^{n \times d} \), we observe
\[
Y = X\theta + \varepsilon,
\]
where \( \varepsilon \in \mathbb{R}^n \) consists of independent random variables \( \varepsilon_i \) with variance bounded by \( \text{Var}(\varepsilon_i) \leq \sigma^2 \), and \( \theta \in \mathbb{R}^d \) is allowed to vary over \( \mathbb{R}^d \). For the purposes of our lower bound, we may assume that \( \varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n \times n}) \). Let \( \mathcal{P} \) denote the family of such normally distributed linear regression problems, and assume for simplicity that \( d \leq 32 \).

In this case, we use the Gilbert-Varshamov bound (Lemma 2.6) to construct a local packing and attain minimax rates. Indeed, let \( \mathcal{V} \) be a packing of \( \{-1, 1\}^d \) such that \( \|v - v'\|_1 \geq d/2 \) for distinct elements of \( \mathcal{V} \), and let \( |\mathcal{V}| \geq \exp(d/8) \) as guaranteed by the Gilbert-Varshamov bound. For fixed \( \delta > 0 \), if we set \( \theta_v = \delta v \), then we have the packing guarantee for distinct elements \( v, v' \) that
\[
\|\theta_v - \theta_{v'}\|_2^2 = \delta^2 \sum_{j=1}^d (v_j - v'_j)^2 = 4\delta^2 \|v - v'\|_1 \geq 2d\delta^2.
\]
Moreover, we have the upper bound
\[
D_{\text{kl}}(\mathcal{N}(X\theta_v, \sigma^2 I_{n \times n}) || \mathcal{N}(X\theta_{v'}, \sigma^2 I_{n \times n})) = \frac{1}{2\sigma^2} \|X(\theta_v - \theta_{v'})\|_2^2 \leq \frac{\delta^2}{2\sigma^2} \gamma_{\text{max}}^2(X) \|v - v'\|_2^2 \leq \frac{2d\gamma_{\text{max}}^2(X)\delta^2}{\sigma^2},
\]
where \( \gamma_{\text{max}}(X) \) denotes the maximum singular value of \( X \). Consequently, the bound (2.3.6) holds with \( \kappa^2 \leq 2d\gamma_{\text{max}}^2(X)/\sigma^2 \), and we have the minimax lower bound
\[
\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|_2^2) \geq \frac{d\sigma^2}{2} \left( 1 - \frac{I(V; Y) + \log 2}{\log |\mathcal{V}|} \right) \geq \frac{d\sigma^2}{2} \left( 1 - \frac{2d\gamma_{\text{max}}^2(X)\delta^2 + \log 2}{d/8} \right).
\]
Now, if we choose
\[
\delta^2 = \frac{\sigma^2}{64\gamma_{\text{max}}^2(X)}, \quad \text{then} \quad 1 - \frac{8\log 2}{d} - \frac{16d\gamma_{\text{max}}^2(X)\delta^2}{d} \geq 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2},
\]
by assumption that \( d \geq 32 \). In particular, we obtain the lower bound
\[
\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|_2^2) \geq \frac{1}{256} \frac{\sigma^2 d}{\gamma_{\text{max}}^2(X)} = \frac{1}{256} \frac{\sigma^2 d}{n} \frac{1}{\gamma_{\text{max}}(\mathcal{U}/n)}
\]

25
for a convergence rate (roughly) of $\sigma^2 d/n$ after rescaling the singular values of $X$ by $1/\sqrt{n}$.

This bound is sharp in terms of the dimension, dependence on $n$, and the variance $\sigma^2$, but it does not fully capture the dependence on $X$, as it depends only on the maximum singular value. Indeed, in this case, an exact calculation (cf. [14]) shows that the minimax value of the problem is exactly $\sigma^2 \text{tr}((X^T X)^{-1})$. Letting $\lambda_j(A)$ be the $j$th eigenvalue of a matrix $A$, we have

$$
\sigma^2 \text{tr}((X^T X)^{-1}) = \frac{\sigma^2}{n} \text{tr}((n^{-1}X^T X)^{-1}) = \frac{\sigma^2}{n} \sum_{j=1}^{d} \frac{1}{\lambda_j(\frac{1}{n}X^T X)} 
\geq \frac{\sigma^2 d}{n} \min_j \frac{1}{\lambda_j(\frac{1}{n}X^T X)} = \frac{\sigma^2 d}{n} \frac{1}{\gamma_{\max}(\frac{1}{n}X^T X)}.
$$

Thus, the local Fano method captures most—but not all—of the difficulty of the problem. ♦

### 2.4.2 A distance-based Fano method

While the testing lower bound (2.4.2) is sufficient for proving lower bounds for many estimation problems, for the sharpest results it sometimes requires a somewhat delicate construction of a well-separated packing (e.g. [4, 8]). To that end, we also provide extensions of inequalities (2.4.1) and (2.4.2) that more directly yield bounds on estimation error, allowing more direct and simpler proofs of a variety of minimax lower bounds (see also reference [7]).

More specifically, suppose that the distance function $\rho_V$ is defined on $V$, and we are interested in bounding the estimation error $\rho_V(\hat{V}, V)$. We begin by providing analogues of the lower bounds (2.4.1) and (2.4.2) that replace the testing error with the tail probability $\mathbb{P}(\rho_V(\hat{V}, V) > t)$. By Markov’s inequality, such control directly yields bounds on the expectation $\mathbb{E}[\rho_V(\hat{V}, V)]$. As we show in the sequel and in chapters to come, these distance-based Fano inequalities allow more direct proofs of a variety of minimax lower bounds without the need for careful construction of packing sets or metric entropy calculations as in other arguments.

We begin with the distance-based analogue of the usual discrete Fano inequality in Proposition 2.10. Let $V$ be a random variable supported on a finite set $V$ with cardinality $|V| \geq 2$, and let $\rho : V \times V \to \mathbb{R}$ be a function defined on $V \times V$. In the usual setting, the function $\rho$ is a metric on the space $V$, but our theory applies to general functions. For a given scalar $t \geq 0$, the maximum and minimum neighborhood sizes at radius $t$ are given by

$$
N_t^{\max} := \max_{v \in V} \{ \text{card}\{ v' \in V \mid \rho(v, v') \leq t \} \} \quad \text{and} \quad N_t^{\min} := \min_{v \in V} \{ \text{card}\{ v' \in V \mid \rho(v, v') \leq t \} \}.
$$

Defining the error probability $P_t = \mathbb{P}(\rho_V(\hat{V}, V) > t)$, we then have the following generalization of Fano’s inequality:

**Proposition 2.15.** For any Markov chain $V \to X \to \hat{V}$, we have

$$
h_2(P_t) + P_t \log \frac{|V| - N_t^{\min}}{N_t^{\max}} + \log N_t^{\max} \geq H(V \mid \hat{V}).
$$

(2.4.10)

Before proving the proposition, which we do in Section 2.5.1, it is informative to note that it reduces to the standard form of Fano’s inequality (2.4.1) in a special case. Suppose that we take $\rho_V$ to be the 0-1 metric, meaning that $\rho_V(v, v') = 0$ if $v = v'$ and 1 otherwise. Setting $t = 0$ in Proposition 2.15 we have $P_0 = \mathbb{P}[\hat{V} \neq V]$ and $N_0^{\min} = N_0^{\max} = 1$, whence inequality (2.4.10) reduces
to inequality (2.4.1). Other weakenings allow somewhat clearer statements (see Section 2.5.2 for a proof):

**Corollary 2.16.** If \( V \) is uniform on \( V \) and \( (|V| - N^\text{min}_t) > N^\text{max}_t \), then

\[
\mathbb{P}(\rho_V(\hat{V}, V) > t) \geq 1 - \frac{I(V; X) + \log 2}{\log \frac{|V|}{N^\text{max}_t}}. \tag{2.4.11}
\]

Inequality (2.4.11) is the natural analogue of the classical mutual-information based form of Fano’s inequality (2.4.2), and it provides a qualitatively similar bound. The main difference is that the usual cardinality \(|V|\) is replaced by the ratio \(|V|/N^\text{max}_t\). This quantity serves as a rough measure of the number of possible “regions” in the space \( V \) that are distinguishable—that is, the number of subsets of \( V \) for which \( \rho_V(v, v') > t \) when \( v \) and \( v' \) belong to different regions. While this construction is similar in spirit to the usual construction of packing sets in the standard reduction from testing to estimation (cf. Section 2.2.1), our bound allows us to skip the packing set construction. We can directly compute \( I(V; X) \) where \( V \) takes values over the full space, as opposed to computing the mutual information \( I(V'; X) \) for a random variable \( V' \) uniformly distributed over a packing set contained within \( V \). In some cases, the former calculation can be much simpler, as illustrated in examples and chapters to follow.

We now turn to providing a few consequences of Proposition 2.15 and Corollary 2.16, showing how they can be used to derive lower bounds on the minimax risk. Proposition 2.15 is a generalization of the classical Fano inequality (2.4.1), so it leads naturally to a generalization of the classical Fano lower bound on minimax risk, which we describe here. This reduction from estimation to testing is somewhat more general than the classical reductions, since we do not map the original estimation problem to a strict test, but rather a test that allows errors. Consider as in the standard approach to minimax lower bounds. On the other hand, allowing for \( t > 0 \) lends greater flexibility to the construction, since only certain pairs \( \theta_v \) and \( \theta_{v'} \) are required to be well-separated.

Given a set \( V \) and associated separation function (2.4.12), we assume the canonical estimation setting: nature chooses \( V \in V \) uniformly at random, and conditioned on this choice \( V = v \), a sample \( X \) is drawn from the distribution \( P_v \). We then have the following corollary of Proposition 2.15, whose argument is completely identical to that for inequality (2.2.1):

**Corollary 2.17.** Given \( V \) uniformly distributed over \( V \) with separation function \( \delta(t) \), we have

\[
\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi\left( \frac{\delta(t)}{2} \right) \left[ 1 - \frac{I(X; V) + \log 2}{\log \frac{|V|}{N^\text{max}_t}} \right] \quad \text{for all } t. \tag{2.4.13}
\]

Notably, using the discrete metric \( \rho_V(v, v') = 1 \{ v \neq v' \} \) and taking \( t = 0 \) in the lower bound (2.4.13) gives the classical Fano lower bound on the minimax risk based on constructing a packing (11, 21, 20). We now turn to an example illustrating the use of Corollary 2.17 in providing a minimax lower bound on the performance of regression estimators.
We then obtain the following minimax lower bound on the minimax error in squared \( \ell_2 \)-norm: there is a universal (numerical) constant \( c > 0 \) such that

\[
\mathcal{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \geq \frac{c \sigma^2 d^2}{\|X\|^2_F} \geq \frac{c}{\gamma_{\max}(X/\sqrt{n})^2} \cdot \frac{\sigma^2 d}{n},
\]

where \( \gamma_{\max} \) denotes the maximum singular value. Notably, this inequality is nearly the sharpest known bound proved via Fano inequality-based methods [4], but our technique is essentially direct and straightforward.

To see inequality (2.4.14), let the set \( \mathcal{V} = \{-1, 1\}^d \) be the \( d \)-dimensional hypercube, and define \( \theta_v = \delta v \) for some fixed \( \delta > 0 \). Then letting \( \rho_v \) be the Hamming metric on \( \mathcal{V} \) and \( \rho \) be the usual \( \ell_2 \)-norm, the associated separation function (2.4.12) satisfies \( \delta(t) > \max\{\sqrt{t}, 1\} \delta \). Now, for any \( t \leq \lfloor d/3 \rfloor \), the neighborhood size satisfies

\[
N_{t}^{\max} = \sum_{\tau=0}^{t} \left( \frac{d}{t} \right)^{d/3} \leq 2 \left( \frac{d}{t} \right)^{d/3}.
\]

Consequently, for \( t \leq d/6 \), the ratio \( \frac{\|\mathcal{V}\|}{N_{t}^{\max}} \) satisfies

\[
\log \frac{\|\mathcal{V}\|}{N_{t}^{\max}} \geq d \log 2 - \log 2 \left( \frac{d}{t} \right) \geq d \log 2 - \frac{d}{6} \log(6e) - \log 2 = d \log \frac{2}{2^{1/d} \cdot \sqrt[6]{6e}} > \max\left\{ \frac{d}{6}, \log 4 \right\}
\]

for \( d \geq 12 \). (The case \( 2 \leq d < 12 \) can be checked directly). In particular, by taking \( t = \lfloor d/6 \rfloor \) we obtain via Corollary 2.17 that

\[
\mathcal{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \geq \frac{\max\{\lfloor d/6 \rfloor, 2\} \delta^2}{4} \left( 1 - \frac{\log \frac{X}{\mathcal{V}}}{} \right).
\]

But of course, for \( V \) uniform on \( \mathcal{V} \), we have \( \mathbb{E}[VV^\top] = I_{d \times d} \), and thus for \( V, V' \) independent and uniform on \( \mathcal{V} \),

\[
\begin{align*}
I(Y; V) &\leq n \frac{1}{|\mathcal{V}|^2} \sum_{v \in \mathcal{V}} \sum_{v' \in \mathcal{V}} D_{KL}(N(X\theta_v, \sigma^2 I_{n \times n}) || N(X\theta_{v'}, \sigma^2 I_{n \times n})) \\
&= \frac{\delta^2}{2 \sigma^2} \mathbb{E} \left[ \|XV - XV'\|_2^2 \right] = \frac{\delta^2}{\sigma^2} \|X\|_{F}^2.
\end{align*}
\]

Substituting this into the preceding minimax bound, we obtain

\[
\mathcal{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \geq \frac{\max\{\lfloor d/6 \rfloor, 2\} \delta^2}{4} \left( 1 - \frac{\delta^2 \|X\|_F^2 / \sigma^2 + \log 2}{\max\{d/6, 2 \log 2\}} \right).
\]

Choosing \( \delta^2 \approx d \sigma^2 / \|X\|_{F}^2 \) gives the result (2.4.14).
2.5 Proofs of results

2.5.1 Proof of Proposition 2.15

Our argument for proving the proposition parallels that of the classical Fano inequality by Cover and Thomas [6]. Letting $E$ be a $\{0, 1\}$-valued indicator variable for the event $\rho(\hat{V}, V) \leq t$, we compute the entropy $H(E, V | \hat{V})$ in two different ways. On one hand, by the chain rule for entropy, we have

$$H(E, V | \hat{V}) = H(V | \hat{V}) + H(E | V, \hat{V}),$$

where the final term vanishes since $E$ is $(V, \hat{V})$-measurable. On the other hand, we also have

$$H(E, V | \hat{V}) = H(E | \hat{V}) + H(V | E, \hat{V}) \leq H(E) + H(V | E, \hat{V}),$$

using the fact that conditioning reduces entropy. Applying the definition of conditional entropy yields

$$H(V | E, \hat{V}) = P(E = 0)H(V | E = 0, \hat{V}) + P(E = 1)H(V | E = 1, \hat{V}),$$

and we upper bound each of these terms separately. For the first term, we have

$$H(V | E = 0, \hat{V}) \leq \log(|V| - N_t^{min}),$$

since conditioned on the event $E = 0$, the random variable $V$ may take values in a set of size at most $|V| - N_t^{min}$. For the second, we have

$$H(V | E = 1, \hat{V}) \leq \log N_t^{max},$$

since conditioned on $E = 1$, or equivalently on the event that $\rho(\hat{V}, V) \leq t$, we are guaranteed that $V$ belongs to a set of cardinality at most $N_t^{max}$.

Combining the pieces and and noting $P(E = 0) = P_t$, we have proved that

$$H(E, V | \hat{V}) \leq H(E) + P_t \log(|V| - N_t^{min}) + (1 - P_t) \log N_t^{max}.$$  

Combining this inequality with our earlier equality (2.5.1), we see that

$$H(V | \hat{V}) \leq H(E) + P_t \log(|V| - N_t^{min}) + (1 - P_t) \log N_t^{max}.$$  

Since $H(E) = h_2(P_t)$, the claim (2.4.10) follows.

2.5.2 Proof of Corollary 2.16

First, by the information-processing inequality [e.g. 6, Chapter 2], we have $I(V; \hat{V}) \leq I(V; X)$, and hence $H(V | X) \leq H(V | \hat{V})$. Since $h_2(P_t) \leq \log 2$, inequality (2.4.10) implies that

$$H(V | X) - \log N_t^{max} \leq H(V | \hat{V}) - \log N_t^{max} \leq P(\rho(\hat{V}, V) > t) \log \frac{|V| - N_t^{min}}{N_t^{max}} + \log 2.$$  

Rearranging the preceding equations yields

$$P(\rho(\hat{V}, V) > t) \geq \frac{H(V | X) - \log N_t^{max} - \log 2}{\log \frac{|V| - N_t^{min}}{N_t^{max}}}. \quad (2.5.2)$$
Note that his bound holds without any assumptions on the distribution of $V$.

By definition, we have $I(V; X) = H(V) - H(V \mid X)$. When $V$ is uniform on $\mathcal{V}$, we have $H(V) = \log |\mathcal{V}|$, and hence $H(V \mid X) = \log |\mathcal{V}| - I(V; X)$. Substituting this relation into the bound \[2.5.2\] yields the inequality

$$P(\rho(\hat{V}, V) > t) \geq \frac{\log |\mathcal{V}|}{\log |\mathcal{V}| - N_{\text{max}} - I(V; X) + \log 2} - 1 - \frac{\log |\mathcal{V}| - N_{\text{max}} - I(V; X) + \log 2}{\log |\mathcal{V}| - N_{\text{max}}}.$$

\section{Deferred proofs}

\subsection{Proof of Proposition 2.4}

For part (a), we begin with the upper bound. We have by Hölder’s inequality that

$$\int |p(x) - q(x)|d\mu(x) = \int |\sqrt{p(x)} - \sqrt{q(x)}| \cdot |\sqrt{p(x)} + \sqrt{q(x)}|d\mu(x) \leq \left( \int (\sqrt{p(x)} - \sqrt{q(x)})^2d\mu(x) \right)^{\frac{1}{2}} \left( \int (\sqrt{p(x)} + \sqrt{q(x)})^2d\mu(x) \right)^{\frac{1}{2}} = d_{\text{hel}}(P, Q) \left( 2 + \int \sqrt{p(x)q(x)}d\mu(x) \right)^{\frac{1}{2}}.$$

But of course, we have $d_{\text{hel}}(P, Q)^2 = 2 - \int \sqrt{p(x)q(x)}d\mu(x)$, so this implies

$$\int |p(x) - q(x)|d\mu(x) \leq d_{\text{hel}}(P, Q)(4 - d_{\text{hel}}(P, Q)^2)^{\frac{1}{2}}.$$

Dividing both sides by 2 gives the upper bound on $\|P - Q\|_{\text{TV}}$. For the lower bound on total variation, note that for any $a, b \in \mathbb{R}_+$, we have $a + b - 2\sqrt{ab} \leq |a - b|$ (check the cases $a > b$ and $a < b$ separately); thus

$$d_{\text{hel}}(P, Q)^2 = \int [p(x) + q(x) - 2\sqrt{p(x)q(x)}]d\mu(x) \leq \int |p(x) - q(x)|d\mu(x).$$

For part (b) we present a proof based on the Cauchy-Schwarz inequality, which differs from standard arguments \[6, 18\]. From the notes on KL-divergence and information theory and Question 5 of homework 1, we may assume without loss of generality that $P$ and $Q$ are finitely supported, say with p.m.f.s $p_1, \ldots, p_m$ and $q_1, \ldots, q_m$. Define the function $h(p) = \sum_{i=1}^{m} p_i \log p_i$. Then showing that $D_{\text{kl}}(P \mid Q) \geq 2\|P - Q\|_{\text{TV}}^2 = \frac{1}{2}\|p - q\|_1^2$ is equivalent to showing that

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

because by inspection $h(p) - h(q) - \langle \nabla h(q), p - q \rangle = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i}$. We do this via a Taylor expansion: we have

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

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

$$h(p) = h(q) + \langle \nabla h(q), p - q \rangle + \frac{1}{2}\langle p - q, \nabla^2 h(\tilde{p})(p - q) \rangle.$$
But looking at the final quadratic, we have for any vector \( v \) and any \( p \geq 0 \) satisfying \( \sum_i p_i = 1 \),

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

where the inequality follows from Cauchy-Schwarz applied to the vectors \( \sqrt{p} \) and \( \frac{|v|}{\sqrt{p}} \).

Thus inequality (2.6.1) holds.

2.7 \( f \)-divergences are jointly convex in their arguments

In this appendix, we prove that \( f \)-divergences are jointly convex in their arguments. To do so, we recall the fact that if a function \( f : \mathbb{R}^d \to \mathbb{R} \) is convex, then its perspective, defined by \( g(x, t) = tf(x/t) \) for \( t > 0 \) and \( (x, t) \) such that \( x/t \in \text{dom} f \), is jointly convex in the arguments \( x, t \) (see Chapter 3.2.6 of Boyd and Vandenberghe [3]). Then we have

**Proposition 2.18.** Let \( P_1, P_2, Q_1, Q_2 \) be distributions on a set \( \mathcal{X} \) and \( f : \mathbb{R}_+ \to \mathbb{R} \) be convex. Then for any \( \lambda \in [0, 1] \),

\[
D_f (\lambda P_1 + (1 - \lambda) P_2 \| \lambda Q_1 + (1 - \lambda) Q_2) \leq \lambda D_f (P_1 \| Q_1) + (1 - \lambda) D_f (P_2 \| Q_2).
\]

**Proof** Assume w.l.o.g. that \( P_i \) and \( Q_i \) have densities \( p_i \) and \( q_i \) w.r.t. the base measure \( \mu \). Define the perspective \( g(x, t) = tf(x/t) \). Then

\[
D_f (\lambda P_1 + (1 - \lambda) P_2 \| \lambda Q_1 + (1 - \lambda) Q_2) = \int (\lambda p_1 + (1 - \lambda) p_2) f \left( \frac{\lambda p_1 + (1 - \lambda) p_2}{\lambda q_1 + (1 - \lambda) q_2} \right) d\mu
\]

\[
= \int g(\lambda p_1 + (1 - \lambda) p_2, \lambda q_1 + (1 - \lambda) q_2) d\mu
\]

\[
\leq \lambda \int g(p_1, q_1) d\mu + (1 - \lambda) \int g(p_2, q_2) d\mu
\]

\[
= \lambda D_f (P_1 \| Q_1) + (1 - \lambda) D_f (P_2 \| Q_2),
\]

where we have used the joint convexity of the perspective function.

\( \square \)
Bibliography


