Lecture Notes for Statistics 311/Electrical Engineering 377

John Duchi

February 23, 2016

Contents

Chapter 1

Introduction and setting

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

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

1.1 Information theory

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

1. How much information does a signal contain?

2. How much information can a noisy channel reliably transmit?

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

Example 1.1 (Source coding)**:** The source coding, or data compression problem, is to take information from a source, compress it, decompress it, and recover the original message. Graphically, we have

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

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

Example 1.2: The channel coding, or data transmission problem, is the same as the source coding problem of Example 1.1, except that between the compressor and decompressor is a source of noise, a *channel*. In this case, the graphical representation is

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

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

1.2 Moving to statistics

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

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

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

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

In our analogy with channel coding, we make a connection with estimation and inference. Roughly, the major problem in statistics we consider is as follows: there exists some unknown function f on a space $\mathcal{X}$ that we wish to estimate, and we are able to observe a noisy version of $f(X_i)$ for a series of X_i drawn from a distribution P. Recalling the graphical description of Example 1.2, we now have a channel $P(Y \mid f(X))$ that gives us noisy observations of $f(X)$ for each X_i, but we may (generally) now longer choose the encoder/compressor. That is, we have

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

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

Example 1.3: A classical example of the statistical paradigm in this lens is the usual linear regression problem. Here the data X_i belong to $\mathbb{R}^d$, and the compression function $f(x) = \theta^\top x$ for some vector $\theta \in \mathbb{R}^d$. Then the channel is often of the form

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

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2)$ are independent mean zero normal perturbations. The goal is, given a sequence of pairs (X_i, Y_i), to recover the true θ in the linear model.

In *active learning* or *active sensing* scenarios, also known as (sequential) experimental design, we may choose the sequence X_i so as to better explore properties of θ. Later in the course we will investigate whether it is possible to improve estimation by these strategies. As one concrete idea, if we allow infinite *power*, which in this context corresponds to letting $\|X_i\| \to \infty$—choosing very "large" vectors x_i—then the signal of $\theta^\top X_i$ should swamp any noise and make estimation easier. ♣

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

Chapter 2

Review of basic (and not so basic) concepts in information theory

Readings covering the material in this set of notes: Chapter 2 of Cover and Thomas [4], which covers all of the results for the discrete case. For the more advanced (measure-theoretic) version, see Chapter 5 of Gray [6] (available on Bob Gray's webpage), or Chapter 7 of the second edition of the same book.

2.1 Basics of Information Theory

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

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

2.1.1 Definitions

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

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

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

While we do not explore it in this class, there is an operational interpretation of entropy via Shannon's source-coding theorem (see, for example, Chapter 5 of Cover and Thomas [4]). In particular, Shannon's source coding theorem states that if we wish to encode a random variable X, distributed according to P, with a k-ary string (i.e. each entry of the string takes on one of k values), then the minimal expected length of the encoding is given by $H(X) = -\sum_x p(x) \log_k p(x)$. Moreover, this is achievable (to within a length of at most 1 symbol) by using Huffman codes (among many other types of codes). As an example of this interpretation, we may consider encoding a random variable X with equi-probable distribution on m items, which has $H(X) = \log(m)$. In base-2, this makes sense: we simply assign an integer to each item and encode each integer with the natural (binary) integer encoding of length $\lceil \log m \rceil$.

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

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

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

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

Example 2.1 (Uniform random variables)**:** As we noted earlier, if a random variable X is uniform on a set of size m, then $H(X) = \log m$. ♣

Example 2.2 (Bernoulli random variables)**:** Let $h_2(p) = -p \log p - (1-p)\log(1-p)$ denote the binary entropy, which is the entropy of a $\mathsf{Bernoulli}(p)$ random variable. ♣

Example 2.3 (Geometric random variables)**:** A random variable X is $\mathsf{Geometric}(p)$, for some $p \in [0, 1]$, if it is supported on $\{1, 2, \ldots\}$, and $P(X = k) = (1-p)^{k-1}p$; this is the probability distribution of the number X of $\mathsf{Bernoulli}(p)$ trials until a single success. The entropy of such a random variable is

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

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

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

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

Example 2.4 (A random variable with infinite entropy)**:** While most "reasonable" discrete random variables have finite entropy, it is possible to construct distributions with infinite entropy. Indeed, let X have p.m.f. on $\{2, 3, \ldots\}$ defined by

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

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

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

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

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

where C is a numerical constant. ♣

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

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

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

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

Proposition 2.5 (Jensen's inequality)**.** *Let f be convex. Then for any random variable X,*

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

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

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

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

We observe immediately that $D_{\rm kl}\left(P \| Q\right) \ge 0$. To see this, we apply Jensen's inequality (Proposition 2.5) to the function $-\log$ and the random variable $q(X)/p(X)$, where X is distributed according to P:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and similarly, we have the analogues of mutual information as

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

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

Example 2.6 (Entropy of normal random variables)**:** The differential entropy (2.1.4) of a normal random variable is straightforward to compute. Indeed, for $X \sim \mathsf{N}(\mu, \sigma^2)$ we have $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$, so that

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

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

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

♣

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

Example 2.7 (Divergence between Gaussian distributions)**:** Let P be the multivariate normal $\mathsf{N}(\mu_1, \Sigma)$, and Q be the multivariate normal distribution with mean μ_2 and identical covariance $\Sigma \succ 0$. Then we have that

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

We leave the computation of the identity (2.1.5) to the reader. ♣

An interesting consequence of Example 2.7 is that if a random vector X has a given covariance $\Sigma \in \mathbb{R}^{n \times n}$, then the multivariate Gaussian with identical covariance has larger differential entropy. Put another way, differential entropy for random variables with second moments is always maximized by the Gaussian distribution.

Proposition 2.8. *Let X be a random vector on $\mathbb{R}^n$ with a density, and assume that $\operatorname{Cov}(X) = \Sigma$. Then for $Z \sim \mathsf{N}(0, \Sigma)$, we have*

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

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

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

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

We remark in passing that the fact that Gaussian random variables have the largest entropy has been used to prove stronger variants of the central limit theorem; see the original results of Barron [3], as well as later quantitative results on the increase of entropy of normalized sums by Artstein et al. [2] and Madiman and Barron [9].

2.1.2 Properties and data processing

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

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

For the entropy, we have the simplest chain rule:

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

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

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

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

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

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

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

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

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

Indeed, we have

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

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

Expanding upon this, we give several *tensorization* identities, showing how to transform questions about the joint distribution of many random variables to simpler questions about their marginals. As a first example, we see that as a consequence of the fact that conditioning decreases entropy, we see that for any sequence of (discrete or continuous, as appropriate) random variables, we have

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

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

We return to information and divergence now. Suppose that random variables Y_i are independent conditional on X, meaning that $P(Y_1 = y_1, \ldots, Y_n = y_n \mid X = x) = P(Y_1 = y_1 \mid X = x) \cdots P(Y_n = y_n \mid X = x)$. Such scenarios are common—as we shall see—when we make multiple observations from a fixed distribution parameterized by some X. Then we have the inequality

$$\begin{aligned} I(X;Y_1, \ldots, Y_n) &= \sum_{i=1}^n [H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid X, Y_1^{i-1})] \\ &= \sum_{i=1}^n [H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid X)] \le \sum_{i=1}^n [H(Y_i) - H(Y_i \mid X)] = \sum_{i=1}^n I(X;Y_i), \end{aligned} \tag{2.1.6}$$

where the inequality follows because conditioning reduces entropy.

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

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

We remark in passing that these two identities hold for arbitrary distributions P_i and Q_i or random variables X, Y.

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

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

Then we obtain the classical *data processing inequality*.

Proposition 2.9. *With the above Markov chain, we have $I(X; Z) \le I(X; Y)$.*

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

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

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

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

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

2.2 General divergence measures and definitions

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

2.2.1 Partitions, algebras, and quantizers

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

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

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

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

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

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

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

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

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

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

2.2.2 KL-divergence

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

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

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

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

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

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

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

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

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

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

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

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

precisely as in the discrete case.

We remark in passing that if the set $\mathcal{X}$ is a product space, meaning that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ for some $n < \infty$ (this is the case for mutual information, for example), then we may assume our quantizer *always* quantizes sets of the form $A = A_1 \times A_2 \times \cdots \times A_n$, that is, Cartesian products. Written differently, when we consider algebras on $\mathcal{X}$, the atoms of the algebra may be assumed to be Cartesian products of sets, and our partitions of $\mathcal{X}$ can always be taken as Cartesian products. (See Gray [6, Chapter 5].) Written slightly differently, if P and Q are distributions on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and q^i is a quantizer for the set $\mathcal{X}_i$ (inducing the partition $A_1^i, \ldots, A_{m_i}^i$ of $\mathcal{X}_i$) we may define

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

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

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

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

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

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

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

The proof of this fact is somewhat involved, requiring the technology of Lebesgue integration. (See Gray [6, Chapter 5].)

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

2.2.3 f-divergences

A more general notion of divergence is the so-called f-divergence, or Ali-Silvey divergence [1, 5] (see also the alternate interpretations in the article by Liese and Vajda [8]). Here, the definition is as follows. Let P and Q be probability distributions on the set $\mathcal{X}$, and let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function satisfying $f(1) = 0$. Assume w.l.o.g. that P and Q are absolutely continuous with respect to the base measure μ. The f divergence between P and Q is

$$D_f\left(P\|Q\right) := \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) d\mu(x). \tag{2.2.3}$$

In the first homework set, you will explore several properties of f-divergences, including a quantized representation equivalent to that for the KL-divergence (2.2.1). Broadly, f-divergences satisfy essentially the same properties as KL-divergence, such as data-processing inequalities, and they provide a generalization of mutual information.

Examples We give three examples of f-divergences here; in Section 13.2.2 we provide a few examples of their uses as well as providing a few natural inequalities between them.

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

2. KL-divergence, reversed: by taking $f(t) = -\log t$, we obtain $D_f\left(P\|Q\right) = D_{\rm kl}\left(Q\|P\right)$.

3. The *total variation distance* between probability distributions P and Q defined on a set $\mathcal{X}$ is defined as the maximum difference between probabilities they assign on subsets of $\mathcal{X}$:

$$\|P - Q\|_{\rm TV} := \sup_{A \subset \mathcal{X}} |P(A) - Q(A)|. \tag{2.2.4}$$

Note that (by considering compliments $P(A^c) = 1 - P(A)$) the absolute value on the right hand side is unnecessary. The total variation distance, as we shall see later in the course, is very important for verifying the optimality of different tests, and appears in the measurement of difficulty of solving hypothesis testing problems. An important inequality, known as *Pinsker's inequality*, is that

$$\|P - Q\|_{\rm TV}^2 \le \frac{1}{2} D_{\rm kl}\left(P\|Q\right). \tag{2.2.5}$$

By taking $f(t) = \frac{1}{2}|t-1|$, we obtain the total variation distance. Indeed, we have

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

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

$$d_{\rm hel}(P, Q)^2 := \int (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x). \tag{2.2.6}$$

5. The χ^2-*divergence* is generated by taking $f(t) = \frac{1}{2}(t-1)^2$, and between distributions P and Q is given by

$$D_{\chi^2}\left(P\|Q\right) = \frac{1}{2}\int \left(\frac{p(x)}{q(x)} - 1\right)^2 q(x) d\mu(x). \tag{2.2.7}$$

There are a variety of inequalities relating different f-divergences, which are often convenient for analyzing the properties of product distributions (as will become apparent in Chapter 13. We enumerate a few of the most important inequalities here, which provide inequalities relating variation distance to the others.

Proposition 2.10. *The total variation distance satisfies the following relationships:*

(a) For the Hellinger distance,

$$\frac{1}{2}d_{\rm hel}(P,Q)^2 \le \|P-Q\|_{\rm TV} \le d_{\rm hel}(P,Q)\sqrt{1-d_{\rm hel}(P,Q)^2/4}.$$

(b) Pinsker's inequality: for any distributions P, Q,

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

We provide the proof of Proposition 2.10 in Section 2.A.1.

2.3 First steps into optimal procedures: testing inequalities

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

2.3.1 Le Cam's inequality and binary hypothesis testing

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

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

We can give an exact expression for the minimal possible error in the above hypothesis test. Indeed, a standard result of Le Cam (see [7, 11, Lemma 1]) is the following variational representation of the total variation distance (2.2.4), which is the f-divergence associated with $f(t) = \frac{1}{2}|t-1|$, as a function of testing error.

Proposition 2.11. *Let $\mathcal{X}$ be an arbitrary set. For any distributions P_1 and P_2 on $\mathcal{X}$, we have*

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

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

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

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

Taking an infimum over such acceptance regions, we have

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

which yields the total variation distance as desired. □

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

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

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

We give one example application of Proposition 2.11 to the problem of testing a normal mean.

Example 2.12 (Testing a normal mean): Suppose we observe $X_1, \ldots, X_n \stackrel{\rm i.i.d.}{\sim} P$ for $P = P_1$ or $P = P_2$, where P_v is the normal distribution $\mathsf{N}(\mu_v, \sigma^2)$, where $\mu_1 \neq \mu_2$. We would like to understand the sample size n necessary to guarantee that no test can have small error, that is, say, that

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

By Proposition 2.11, we have that

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

where P_v^n denotes the n-fold product of P_v, that is, the distribution of $X_1, \ldots, X_n \stackrel{\rm i.i.d.}{\sim} P_v$. The interaction between total variation distance and product distributions is somewhat subtle, so it is often advisable to use a divergence measure more attuned to the i.i.d. nature of the sampling scheme. Two such measures are the KL-divergence and Hellinger distance, both of which we explore in the coming chapters. With that in mind, we apply Pinsker's inequality (2.2.5) to see that $\|P_1^n - P_2^n\|_{\rm TV}^2 \leq \frac{1}{2} D_{\rm kl}(P_1^n \| P_2^n) = \frac{n}{2} D_{\rm kl}(P_1 \| P_2)$, which implies that

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

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

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

2.3.2 Fano's inequality and multiple hypothesis testing

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

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

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

and we wish to provide lower bounds on the probability of error—that is, that $\widehat{X} \neq X$. If we let the function $h_2(p) = -p \log p - (1-p)\log(1-p)$ denote the binary entropy (entropy of a Bernoulli random variable with parameter p), Fano's inequality takes the following form [e.g. 4, Chapter 2]:

Proposition 2.13 (Fano inequality)**.** *For any Markov chain $X \to Y \to \widehat{X}$, we have*

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

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

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

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

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

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

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

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

We can rewrite Proposition 2.13 in a convenient way when X is uniform in $\mathcal{X}$. Indeed, by definition of the mutual information, we have $I(X; \widehat{X}) = H(X) - H(X \mid \widehat{X})$, so Proposition 13.8 implies that in the canonical hypothesis testing problem from Section 13.2.1, we have

Corollary 2.14. *Assume that X is uniform on $\mathcal{X}$. For any Markov chain $X \to Y \to \widehat{X}$,*

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

Proof Let $P_{\mathsf{error}} = \mathbb{P}(X \neq \widehat{X})$ denote the probability of error. Noting that $h_2(p) \le \log 2$ for any $p \in [0, 1]$ (recall inequality (2.1.2), that is, that uniform random variables maximize entropy), then using Proposition 13.8, we have

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

Here step (i) uses Proposition 2.13 and step (ii) uses the definition of mutual information, that $I(X; \widehat{X}) = H(X) - H(X \mid \widehat{X})$. The data processing inequality implies that $I(X; \widehat{X}) \le I(X; Y)$, and using $H(X) = \log(|\mathcal{X}|)$ completes the proof. □

In particular, Corollary 2.14 shows that when X is chosen uniformly at random and we observe Y, we have

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

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

Example 2.15 (20 questions game)**:** In the 20 questions game—a standard children's game—there are two players, the "chooser" and the "guesser," and an agreed upon universe $\mathcal{X}$. The chooser picks an element $x \in \mathcal{X}$, and the guesser's goal is to find x by using a series of yes/no questions about x. We consider optimal strategies for each player in this game, assuming that $\mathcal{X}$ is finite and letting $m = |\mathcal{X}|$ be the universe size for shorthand.

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

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

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

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

By the chain rule for mutual information, we have

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

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

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

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

2.A Deferred proofs

2.A.1 Proof of Proposition 2.10

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

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

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

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

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

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

For part (b) we present a proof based on the Cauchy-Schwarz inequality, which differs from standard arguments [4, 10]. 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 $D_{\rm kl}(P \| Q) \ge 2 \|P - Q\|_{\rm TV}^2 = \frac{1}{2} \|p - q\|_1^2$ is equivalent to showing that

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

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

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

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

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

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

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

where the inequality follows from Cauchy-Schwarz applied to the vectors $[\sqrt{p_i}]_i$ and $[|v_i|/\sqrt{p_i}]_i$. Thus inequality (2.A.1) holds. □

Bibliography

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

[2] S. Artstein, K. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004.

[3] A. Barron. Entropy and the central limit theorem. *Annals of Probability*, 14(1):336–342, 1986.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition.* Wiley, 2006.

[5] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientifica Mathematica Hungary*, 2:299–318, 1967.

[6] R. M. Gray. *Entropy and Information Theory.* Springer, 1990.

[7] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, 1986.

[8] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[9] M. Madiman and A. Barron. Generalized entropy power inequalities and monotonicity properties of information. *IEEE Transactions on Information Theory*, 53(7):2317–2329, 2007.

[10] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2009.

[11] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

Chapter 3

Concentration Inequalities

In many scenarios, it is useful to understand how a random variable X behaves by giving bounds on the probability that it deviates far from its mean or median. This can allow us to give prove that estimation and learning procedures will have certain performance, that different decoding and encoding schemes work with high probability, among other results. In this chapter, we give several tools for proving bounds on the probability that random variables are far from their typical values, using more advanced information-theoretic ideas to give stronger results.

A few references on concentration, random matrices, and entropies include Vershynin's extraordinarily readable lecture notes [10], the comprehensive book of Boucheron, Lugosi, and Massart [4], and the more advanced material in Buldygin and Kozachenko [5]. Many of our arguments are based off of those of Vershynin and Boucheron et al.

3.1 Basic tail inequalities

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

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

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

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

Proposition 3.1 (Markov's inequality)**.** *Let X be a nonnegative random variable, meaning that $X \geq 0$ with probability 1. Then*

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

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

When we know more about a random variable than that its expectation is finite, we can give somewhat more powerful bounds on the probability that the random variable deviates from its typical values. The first step in this direction, Chebyshev's inequality, requires two moments, and when we have exponential moments, we can give even stronger results. As we shall see, each of these results is but an application of Proposition 3.1.

Proposition 3.2 (Chebyshev's inequality)**.** *Let X be a random variable with* $\mathrm{Var}(X) < \infty$*. Then*

$$\mathbb{P}(X - \mathbb{E}[X] \ge t) \le \frac{\mathrm{Var}(X)}{t^2} \quad \textit{and} \quad \mathbb{P}(X - \mathbb{E}[X] \le -t) \le \frac{\mathrm{Var}(X)}{t^2}$$

for all $t \ge 0$*.*

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

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

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

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

Proposition 3.3. *For any random variable X, we have*

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

for all $\lambda \ge 0$*.*

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

In particular, taking the infimum over all $\lambda \ge 0$ in Proposition 3.3 gives the more standard Chernoff (large deviation) bound

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

Example 3.4 (Gaussian random variables)**:** When X is a mean-zero Gaussian variable with variance σ^2, we have

$$\varphi_X(\lambda) = \mathbb{E}[\exp(\lambda X)] = \exp\left(\frac{\lambda^2 \sigma^2}{2}\right). \tag{3.1.1}$$

To see this, we compute the integral; we have

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

because this is simply the integral of the Gaussian density.
As a consequence of the equality (3.1.1) and the Chernoff bound technique (Proposition 3.3), we see that for X Gaussian with variance σ^2, we have

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

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

3.1.1 Sub-Gaussian random variables

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

Definition 3.1. *A random variable X is* sub-Gaussian with parameter σ^2 *if*

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

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

Of course, Gaussian random variables satisfy Definition 3.1 with equality. This would be uninteresting if only Gaussian random variables satisfied this property; happily, that is not the case, and we detail several examples.

Example 3.5 (Random signs (Rademacher variables))**:** The random variable X taking values $\{-1, 1\}$ with equal property is 1-sub-Gaussian. Indeed, we have

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

as claimed. ♣

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

Example 3.6 (Bounded random variables)**:** Suppose that X is bounded, say $X \in [a, b]$. Then Hoeffding's lemma states that

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

so that X is $(b-a)^2/4$-sub-Gaussian.

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

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

where the inequality follows from Jensen's inequality and the last equality is a conseqence of the fact that $X - X'$ is symmetric about 0. Using the result of Example 3.5,

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

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

Chernoff bounds for sub-Gaussian random variables are immediate; indeed, they have the same concentration properties as Gaussian random variables, a consequence of the nice analytical properties of their moment generating functions (that their logarithms are at most quadratic). Thus, using the technique of Example 3.4, we obtain the following proposition.

Proposition 3.7. *Let X be a σ^2-sub-Gaussian. Then for all $t \ge 0$ we have*

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

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

Proposition 3.8. *Let $X_1, X_2, \ldots, X_n$ be independent σ_i^2-sub-Gaussian random variables. Then*

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

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

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

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

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

Two immediate corollary to Propositions 3.7 and 3.8 show that sums of sub-Gaussian random variables concentrate around their expectations. We begin with a general concentration inequality.

Corollary 3.9. *Let X_i be independent σ_i^2-sub-Gaussian random variables. Then for all $t \ge 0$*

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

Additionally, the classical Hoeffding bound, follows when we couple Example 3.6 with Corollary 3.9: if $X_i \in [a_i, b_i]$, then

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

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

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

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

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

There are a variety of other conditions equivalent to sub-Gaussianity, which we relate by defining the sub-Gaussian norm of a random variable. In particular, we define the sub-Gaussian norm (sometimes known as the ψ_2-Orlicz norm in the literature) as

$$\|X\|_{\psi_2} := \sup_{k \ge 1} \frac{1}{\sqrt{k}} \mathbb{E}[|X|^k]^{1/k} \tag{3.1.2}$$

Then we have the following various characterizations of sub-Gaussianity.

Theorem 3.10. *Let X be a mean-zero random variable and $\sigma^2 \ge 0$ be a constant. The following statements are all equivalent, meaning that there are numerical constant factors K_j such that if one statement (i) holds with parameter K_i, then statement (j) holds with parameter $K_j \le CK_i$, where C is a numerical constant.*

(1) Sub-gaussian tails: $\mathbb{P}(|X| \ge t) \le 2\exp(-\frac{t^2}{K_1\sigma^2})$ for all $t \ge 0$.

(2) Sub-gaussian moments: $\mathbb{E}[|X|^k]^{1/k} \le K_2\sigma\sqrt{k}$ for all k.

(3) Super-exponential moment: $\mathbb{E}[\exp(X^2/(K_3\sigma^2))] \le e$.

(4) Sub-gaussian moment generating function: $\mathbb{E}[\exp(\lambda X)] \le \exp(K_4\lambda^2\sigma^2)$ for all $\lambda \in \mathbb{R}$.

Particularly, (1) implies (2) with $K_1 = 1$ and $K_2 \le e^{1/e}$; (2) implies (3) with $K_2 = 1$ and $K_3 = e\sqrt{\frac{2}{e-1}} < 3$; (3) implies (4) with $K_3 = 1$ and $K_4 \le \frac{3}{4}$; and (4) implies (1) with $K_4 = \frac{1}{2}$ and $K_1 \le 2$.

This result is standard in the literature on concentration and random variables; our proof is based on Vershynin [10]. See Appendix 3.A.1 for a proof of this theorem. We note in passing that in each of the statements of Theorem 3.10, we may take $\sigma = \|X\|_{\psi_2}$, and (in general) these are the sharpest possible results except for numerical constants.

For completeness, we can give a tighter result than part (3) of the preceding theorem, giving a concrete upper bound on squares of sub-Gaussian random variables. The technique used in the example, to introduce an independent random variable for auxiliary randomization, is a common and useful technique in probabilistic arguments (similar to our use of symmetrization in Example 3.6).

Example 3.11 (Sub-Gaussian squares)**:** Let X be a mean-zero σ^2-sub-Gaussian random variable. Then

$$\mathbb{E}[\exp(\lambda X^2)] \le \frac{1}{[1 - 2\sigma^2\lambda]_+^{\frac{1}{2}}}, \tag{3.1.3}$$

and expression (3.1.3) holds with equality for $X \sim \mathsf{N}(0, \sigma^2)$.

To see this result, we focus on the Gaussian case first and assume (for this case) without loss of generality (by scaling) that $\sigma^2 = 1$. Assuming that $\lambda < \frac{1}{2}$, we have

$$\mathbb{E}[\exp(\lambda Z^2)] = \int \frac{1}{\sqrt{2\pi}} e^{-(\frac{1}{2}-\lambda)z^2} dz = \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1-2\lambda}{2}z^2} dz = \frac{\sqrt{2\pi}}{\sqrt{1-2\lambda}} \frac{1}{\sqrt{2\pi}},$$

the final equality a consequence of the fact that (as we know for normal random variables) $\int e^{-\frac{1}{2\sigma^2}z^2} dz = \sqrt{2\pi\sigma^2}$. When $\lambda \ge \frac{1}{2}$, the above integrals are all infinite, giving the equality in expression (3.1.3).

For the more general inequality, we recall that if Z is an independent $\mathsf{N}(0,1)$ random variable, then $\mathbb{E}[\exp(tZ)] = \exp(\frac{t^2}{2})$, and so

$$\mathbb{E}[\exp(\lambda X^2)] = \mathbb{E}[\exp(\sqrt{2\lambda} XZ)] \overset{(i)}{\le} \mathbb{E}\left[\exp(\lambda\sigma^2 Z^2)\right] \overset{(ii)}{=} \frac{1}{[1-2\sigma^2\lambda]_+^{\frac{1}{2}}},$$

where inequality (i) follows because X is sub-Gaussian, and inequality (ii) because $Z \sim \mathsf{N}(0,1)$. ♣

3.1.2 Sub-exponential random variables

A slightly weaker condition than sub-Gaussianity is for a random variable to be *sub-exponential*, which—for a mean-zero random variable—means that its moment generating function exists in a neighborhood of zero.

Definition 3.2. *A random variable X is* sub-exponential *with parameters (τ^2, b) if for all λ such that $|\lambda| \le 1/b$,*

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

It is clear from Definition 3.2 that a σ^2-sub-Gaussian random variable is $(\sigma^2, 0)$-sub-exponential.

A variety of random variables are sub-exponential. As a first example, χ^2-random variables are sub-exponential with constant values for τ and b:

Example 3.12: Let $X = Z^2$, where $Z \sim \mathsf{N}(0,1)$. We claim that

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \le \exp(2\lambda^2) \ \text{ for } \lambda \le \frac{1}{4}. \tag{3.1.4}$$

Indeed, for $\lambda < \frac{1}{2}$ we have that

$$\mathbb{E}[\exp(\lambda(Z^2 - \mathbb{E}[Z^2]))] = \exp\left(-\frac{1}{2}\log(1-2\lambda) - \lambda\right) \overset{(i)}{\le} \exp\left(\lambda + 2\lambda^2 - \lambda\right)$$

where inequality (i) holds for $\lambda \le \frac{1}{4}$, because $-\log(1-2\lambda) \le 2\lambda + 4\lambda^2$ for $\lambda \le \frac{1}{4}$. ♣

As a second example, we can show that bounded random variables are sub-exponential. It is clear that this is the case as they are also sub-Gaussian; however, in many cases, it is possible to show that their parameters yield much tighter control over deviations than is possible using only sub-Gaussian techniques.

Example 3.13 (Bounded random variables are sub-exponential)**:** Suppose that X is a mean zero random variable taking values in $[-b, b]$ with variance $\sigma^2 = \mathbb{E}[X^2]$ (note that we are guaranteed that $\sigma^2 \le b^2$ in this case). We claim that

$$\mathbb{E}[\exp(\lambda X)] \le \exp\left(\frac{3\lambda^2\sigma^2}{5}\right) \quad \text{for } |\lambda| \le \frac{1}{2b}. \tag{3.1.5}$$

To see this, note first that for $k \ge 2$ we have $\mathbb{E}[|X|^k] \le \mathbb{E}[X^2 b^{k-2}] = \sigma^2 b^{k-2}$. Then by an expansion of the exponential, we find

$$\begin{aligned} \mathbb{E}[\exp(\lambda X)] &= 1 + \mathbb{E}[\lambda X] + \frac{\lambda^2 \mathbb{E}[X^2]}{2} + \sum_{k=3}^{\infty} \frac{\lambda^k \mathbb{E}[X^k]}{k!} \le 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \frac{\lambda^k \sigma^2 b^{k-2}}{k!} \\ &= 1 + \frac{\lambda^2\sigma^2}{2} + \lambda^2\sigma^2 \sum_{k=1}^{\infty} \frac{(\lambda b)^k}{(k+2)!} \overset{(i)}{\le} 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{10}, \end{aligned}$$

inequality (i) holding for $\lambda \le \frac{1}{2b}$. Using that $1 + x \le e^x$ gives the result. ♣

In particular, if the variance $\sigma^2 \ll b^2$, the absolute bound on X, inequality (3.1.5) gives much tighter control on the moment generating function of X than typical sub-Gaussian bounds based only on the fact that $X \in [-b, b]$ allow.

We can give a broader characterization, as with sub-Gaussian random variables in Theorem 3.10. First, we define the sub-exponential norm (in the literature, there is an equivalent norm often called the Orlicz ψ_1-norm)

$$\|X\|_{\psi_1} := \sup_{k \ge 1} \frac{1}{k} \mathbb{E}[|X|^k]^{1/k}.$$

For any sub-Gaussian random variable—whether it has mean-zero or not—we have that sub-exponential is sub-Gaussian squared:

$$\|X\|_{\psi_2}^2 \le \left\|X^2\right\|_{\psi_1} \le 2\|X\|_{\psi_2}^2, \tag{3.1.6}$$

which is immediate from the definitions. More broadly, we can show a result similar to Theorem 3.10.

Theorem 3.14. *Let X be a random variable and $\sigma \ge 0$. Then—in the sense of Theorem 3.10—the following statements are all equivalent for suitable numerical constants $K_1, \ldots, K_4$.*

(1) Sub-exponential tails: $\mathbb{P}(|X| \ge t) \le 2\exp(-\frac{t}{K_1\sigma})$ for all $t \ge 0$

(2) Sub-exponential moments: $\mathbb{E}[|X|^k]^{1/k} \le K_2\sigma k$ for all $k \ge 1$.

(3) Existence of moment generating function: $\mathbb{E}[\exp(X/(K_3\sigma))] \le e$.

(4) If, in addition, $\mathbb{E}[X] = 0$, then $\mathbb{E}[\exp(\lambda X)] \le \exp(K_4\lambda^2\sigma^2)$ for all $|\lambda| \le K_4'/\sigma$.

In particular, if (2) holds with $K_2 = 1$, then (4) holds with $K_4 = 2e^2$ and $K_4' = \frac{1}{2e}$.

The proof, which is similar to that for Theorem 3.10, is presented in Section 3.A.2.

While the concentration properties of sub-exponential random variables are not quite so nice as those for sub-Gaussian random variables (recall Hoeffding's inequality, Corollary 3.9), we can give sharp tail bounds for sub-exponential random variables. We first give a simple bound on deviation probabilities.

Proposition 3.15. *Let X be a mean-zero (τ^2, b)-sub-exponential random variable. Then for all $t \ge 0$,*

$$\mathbb{P}(X \ge t) \vee \mathbb{P}(X \le -t) \le \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{\tau^2}, \frac{t}{b}\right\}\right).$$

Proof The proof is an application of the Chernoff bound technique; we prove only the upper tail as the lower tail is similar. We have

$$\mathbb{P}(X \ge t) \le \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}} \overset{(i)}{\le} \exp\left(\frac{\lambda^2\tau^2}{2} - \lambda t\right),$$

inequality (i) holding for $|\lambda| \le 1/b$. To minimize the last term in λ, we take $\lambda = \min\{\frac{t}{\tau^2}, 1/b\}$, which gives the result. □

Comparing with sub-Gaussian random variables, which have $b = 0$, we see that Proposition 3.15 gives a similar result for small t—essentially the same concentration sub-Gaussian random variables—while for large t, the tails decrease only exponentially in t.

We can also give a tensorization identity similar to Proposition 3.8.

Proposition 3.16. *Let $X_1, \ldots, X_n$ be independent mean-zero sub-exponential random variables, where X_i is (σ_i^2, b_i)-sub-exponential. Then for any vector $a_i \in \mathbb{R}^n$, we have*

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n X_i\right)\right] \le \exp\left(\frac{\lambda^2\sum_{i=1}^n a_i^2\sigma_i^2}{2}\right) \quad \textit{for} \quad |\lambda| \le \frac{1}{b_*},$$

where $b_* = \max_i b_i|a_i|$. That is, $\langle a, X\rangle$ is $(\sum_{i=1}^n a_i^2\sigma_i^2, \min_i \frac{1}{b_i|a_i|})$-sub-exponential.

Proof We apply an inductive technique similar to that used in the proof of Proposition 3.8. First, for any fixed i, we know that if $|\lambda| \le \frac{1}{b_i|a_i|}$, then $|a_i\lambda| \le \frac{1}{b_i}$ and so

$$\mathbb{E}[\exp(\lambda a_i X_i)] \le \exp\left(\frac{\lambda^2 a_i^2\sigma_i^2}{2}\right).$$

Now, we inductively apply the preceding inequality, which applies so long as $|\lambda| \le \frac{1}{b_i|a_i|}$ for all i. We have

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n a_i X_i\right)\right] = \prod_{i=1}^n \mathbb{E}[\exp(\lambda a_i X_i)] \le \prod_{i=1}^n \exp\left(\frac{\lambda^2 a_i^2\sigma_i^2}{2}\right),$$

which is our desired result. □

As in the case of sub-Gaussian random variables, a combination of the tensorization property—that the moment generating functions of sums of sub-exponential random variables are well-behaved—of Proposition 3.16 and the concentration inequality (3.15) immediately yields the following Bernstein-type inequality. (See also Vershynin [10].)

Corollary 3.17. *Let $X_1, \ldots, X_n$ be independent mean-zero (σ_i^2, b_i)-sub-exponential random variables (Definition 3.2). Define $b_* := \max_i b_i$. Then for all $t \ge 0$ and all vectors $a \in \mathbb{R}^n$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \ge t\right) \vee \mathbb{P}\left(\sum_{i=1}^n a_i X_i \le -t\right) \le \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{\sum_{i=1}^n a_i^2\sigma_i^2}, \frac{t}{b_* \|a\|_\infty}\right\}\right).$$

It is instructive to study the structure of the bound of Corollary 3.17. Notably, the bound is similar to the Hoeffding-type bound of Corollary 3.9 (holding for σ^2-sub-Gaussian random variables) that

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \ge t\right) \le \exp\left(-\frac{t^2}{2\left\|a\right\|_2^2 \sigma^2}\right),$$

so that for small t, Corollary 3.17 gives sub-Gaussian tail behavior. For large t, the bound is weaker. However, in many cases, Corollary 3.17 can give finer control than naive sub-Gaussian bounds. Indeed, suppose that the random variables X_i are i.i.d., mean zero, and satisfy $X_i \in [-b, b]$ with probability 1, but have variance $\sigma^2 = \mathbb{E}[X_i^2] \le b^2$ as in Example 3.13. Then Corollary 3.17 implies that

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \ge t\right) \le \exp\left(-\frac{1}{2}\min\left\{\frac{5}{6}\frac{t^2}{\sigma^2\left\|a\right\|_2^2}, \frac{t}{2b\left\|a\right\|_\infty}\right\}\right). \tag{3.1.7}$$

When applied to a standard mean (and with a minor simplification that $5/12 < 1/3$) with $a_i = \frac{1}{n}$, we obtain the bound that $\frac{1}{n}\sum_{i=1}^n X_i \le t$ with probability at least $1 - \exp(-n\min\{\frac{t^2}{3\sigma^2}, \frac{t}{4b}\})$. Written differently, we take $t = \max\{\sigma\sqrt{\frac{3\log\frac{1}{\delta}}{n}}, \frac{4b\log\frac{1}{\delta}}{n}\}$ to obtain

$$\frac{1}{n}\sum_{i=1}^n X_i \le \max\left\{\sigma\frac{\sqrt{3\log\frac{1}{\delta}}}{\sqrt{n}}, \frac{4b\log\frac{1}{\delta}}{n}\right\} \quad \text{with probability } 1-\delta.$$

The sharpest such bound possible via more naive Hoeffding-type bounds is $b\sqrt{2\log\frac{1}{\delta}}/\sqrt{n}$, which has substantially worse scaling.

3.1.3 First applications of concentration: random projections

In this section, we investigate the use of concentration inequalities in random projections. As motivation, consider nearest-neighbor (or k-nearest-neighbor) classification schemes. We have a sequence of data points as pairs (u_i, y_i), where the vectors $u_i \in \mathbb{R}^d$ have labels $y_i \in \{1, \ldots, L\}$, where L is the number of possible labels. Given a new point $u \in \mathbb{R}^d$ that we wish to label, we find the k-nearest neighbors to u in the sample $\{(u_i, y_i)\}_{i=1}^n$, then assign u the majority label of these k-nearest neighbors (ties are broken randomly). Unfortunately, it can be prohibitively expensive to store high-dimensional vectors and search over large datasets to find near vectors; this has motivated a line of work in computer science on fast methods for nearest neighbors based on reducing the dimension while preserving essential aspects of the dataset. This line of research begins with Indyk and Motwani [8], and continuing through a variety of other works, including Indyk [7] and work on locality-sensitive hashing by Andoni et al. [3], among others. The original approach is due to Johnson and Lindenstrauss, who used the results in the study of Banach spaces [9]; our proof follows a standard argument.

The most specific variant of this problem is as follows: we have n points $u_1, \ldots, u_n$, and we could like to construct a mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^m$, where $m \ll d$, such that

$$\left\|\Phi u_i - \Phi u_j\right\|^2 \in (1 \pm \epsilon)\left\|u_i - u_j\right\|^2.$$

Depending on the norm chosen, this task may be impossible; for the Euclidean (ℓ_2) norm, however, such an embedding is easy to construct using Gaussian random variables and with $m = O(\frac{1}{\epsilon^2}\log n)$. This embedding is known as the Johnson-Lindenstrauss embedding. Note that this size m is *independent* of the dimension d, only depending on the number of points n.

Example 3.18 (Johnson-Lindenstrauss)**:** Let the matrix $\Phi \in \mathbb{R}^{m \times d}$ be defined as follows:

$$\Phi_{ij} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1/m),$$

and let $\Phi_i \in \mathbb{R}^d$ denote the ith row of this matrix. We claim that

$$m \ge \frac{8}{\epsilon^2}\left[2 \log n + \log \frac{1}{\delta}\right] \quad \text{implies} \quad \|\Phi u_i - \Phi u_j\|_2^2 \in (1 \pm \epsilon) \|u_i - u_j\|_2^2$$

for all pairs u_i, u_j with probability at least $1 - \delta$. In particular, $m \gtrsim \frac{\log n}{\epsilon^2}$ is sufficient to achieve accurate dimension reduction with high probability.
To see this, note that for any fixed vector u,

$$\frac{\langle \Phi_i, u \rangle}{\|u\|_2} \sim \mathsf{N}(0, 1/m), \quad \text{and} \quad \frac{\|\Phi u\|_2^2}{\|u\|_2^2} = \sum_{i=1}^m \langle \Phi_i, u / \|u\|_2 \rangle^2$$

is a sum of independent scaled χ^2-random variables. In particular, we have $\mathbb{E}[\|\Phi u / \|u\|_2\|_2^2] = 1$, and using the χ^2-concentration result of Example 3.12 yields

$$\begin{aligned}\mathbb{P}\left(\left|\|\Phi u\|_2^2 / \|u\|_2^2 - 1\right| \ge \epsilon\right) &= \mathbb{P}\left(m \left|\|\Phi u\|_2^2 / \|u\|_2^2 - 1\right| \ge m\epsilon\right) \\ &\le 2 \inf_{|\lambda| \le \frac{1}{4}} \exp\left(2m\lambda^2 - \lambda m \epsilon\right) = 2 \exp\left(-\frac{m\epsilon^2}{8}\right),\end{aligned}$$

the last inequality holding for $\epsilon \in [0, 1]$. Now, using the union bound applied to each of the pairs (u_i, u_j) in the sample, we have

$$\mathbb{P}\left(\text{there exist } i \ne j \text{ s.t. } \left|\|\Phi(u_i - u_j)\|_2^2 - \|u_i - u_j\|_2^2\right| \ge \epsilon \|u_i - u_j\|_2^2\right) \le 2\binom{n}{2} \exp\left(-\frac{m\epsilon^2}{8}\right).$$

Taking $m \ge \frac{8}{\epsilon^2} \log \frac{n^2}{\delta} = \frac{16}{\epsilon^2} \log n + \frac{8}{\epsilon^2} \log \frac{1}{\delta}$ yields that with probability at least $1 - \delta$, we have $\|\Phi u_i - \Phi u_j\|_2^2 \in (1 \pm \epsilon) \|u_i - u_j\|_2^2$. ♣

Computing low-dimensional embeddings of high-dimensional data is an area of active research, and more recent work has shown how to achieve sharper constants [6] and how to use more structured matrices to allow substantially faster computation of the embeddings Φu (see, for example, Achlioptas [1] for early work in this direction, and Ailon and Chazelle [2] for the so-called "Fast Johnson-Lindenstrauss transform").

3.1.4 A second application of concentration: codebook generation

We now consider a (very simplified and essentially un-implementable) view of encoding a signal for transmission and generation of a codebook for transmitting said signal. Suppose that we have a set of words, or signals, that we wish to transmit; let us index them by $i \in \{1, \ldots, m\}$, so that there are m total signals we wish to communicate across a *binary symmetric channel* Q, meaning that given an input bit $x \in \{0, 1\}$, Q outputs a $z \in \{0, 1\}$ with $Q(Z = x \mid x) = 1 - \epsilon$ and $Q(Z = 1 - x \mid x) = \epsilon$, for some $\epsilon < \frac{1}{2}$. (For simplicity, we assume Q is *memoryless*, meaning that when the channel is used multiple times on a sequence $x_1, \ldots, x_n$, its outputs $Z_1, \ldots, Z_n$ are conditionally independent: $Q(Z_{1:n} = z_{1:n} \mid x_{1:n}) = Q(Z_1 = z_1 \mid x_1) \cdots Q(Z_n = z_n \mid x_n)$.)

We consider a simplified block coding scheme, where we for each i we associate a codeword $x_i \in \{0,1\}^d$, where d is a dimension (block length) to be chosen. Upon sending the codeword over the channel, and receiving some $z^{\rm rec} \in \{0,1\}^d$, we decode by choosing

$$i^* \in \mathop{\rm argmax}_{i \in [m]} Q(Z = z^{\rm rec} \mid x_i) = \mathop{\rm argmin}_{i \in [m]} \left\| z^{\rm rec} x_i \right\|_1, \tag{3.1.8}$$

the maximum likelihood decoder. We now investigate how to choose a collection $\{x_1, \ldots, x_m\}$ of such codewords and give finite sample bounds on its probability of error. In fact, by using concentration inequalities, we can show that a randomly drawn codebook of fairly small dimension is likely to enjoy good performance.

Intuitively, if our codebook $\{x_1, \ldots, x_m\} \subset \{0,1\}^d$ is *well-separated*, meaning that each pair of words x_i, x_k satisfies $\left\|x_i - x_k\right\|_1 \ge cd$ for some numerical constant $c > 0$, we should be unlikely to make a mistake. Let us make this precise. We mistake word i for word k only if the received signal Z satisfies $\left\|Z - x_i\right\|_1 \ge \left\|Z - x_k\right\|_1$, and letting $J = \{j \in [d] : x_{ij} \neq x_{kj}\}$ denote the set of at least $c \cdot d$ indices where x_i and x_k differ, we have

$$\left\|Z - x_i\right\|_1 \ge \left\|Z - x_k\right\|_1 \quad \text{if and only if} \quad \sum_{j \in J} |Z_j - x_{ij}| - |Z_j - x_{kj}| \ge 0.$$

If x_i is the word being sent and x_i and x_k differ in position j, then $|Z_j - x_{ij}| - |Z_j - x_{kj}| \in \{-1, 1\}$, and is equal to -1 with probability $(1 - \epsilon)$ and 1 with probability ϵ. That is, we have $\left\|Z - x_i\right\|_1 \ge \left\|Z - x_k\right\|_1$ if and only if

$$\sum_{j \in J} |Z_j - x_{ij}| - |Z_j - x_{kj}| + |J|(1 - 2\epsilon) \ge |J|(1 - 2\epsilon) \ge cd(1 - 2\epsilon),$$

and the expectation $\mathbb{E}_Q[|Z_j - x_{ij}| - |Z_j - x_{kj}| \mid x_i] = -(1 - 2\epsilon)$ when $x_{ij} \neq x_{kj}$. Using the Hoeffding bound, then, we have

$$Q(\left\|Z - x_i\right\|_1 \ge \left\|Z - x_k\right\|_1 \mid x_i) \le \exp\left(-\frac{|J|(1 - 2\epsilon)^2}{2}\right) \le \exp\left(-\frac{cd(1 - 2\epsilon)^2}{2}\right),$$

where we have used that there are at least $|J| \ge cd$ indices differing between x_i and x_k. The probability of making a mistake at all is thus at most $m \exp(-\frac{1}{2} cd(1 - 2\epsilon)^2)$ if our codebook has separation $c \cdot d$.

For low error decoding to occur with extremely high probability, it is thus sufficient to choose a set of code words $\{x_1, \ldots, x_m\}$ that is well separated. To that end, we state a simple lemma.

Lemma 3.19. *Let X_i, $i = 1, \ldots, m$ be drawn independently and uniformly on the d-dimensional hypercube $\mathcal{H}_d := \{0,1\}^d$. Then for any $t \ge 0$,*

$$\mathbb{P}\left(\exists\, i, j \text{ s.t. } \left\|X_i - X_j\right\|_1 < \frac{d}{2} - dt\right) \le \binom{m}{2} \exp\left(-2dt^2\right) \le \frac{m^2}{2} \exp\left(-2dt^2\right).$$

Proof First, let us consider two independent draws X and X' uniformly on the hypercube. Let $Z = \sum_{j=1}^d \mathbf{1}\left\{X_j \neq X_j'\right\} = d_{\rm ham}(X, X') = \|X - X'\|_1$. Then $\mathbb{E}[Z] = \frac{d}{2}$. Moreover, Z is an i.i.d. sum of Bernoulli $\frac{1}{2}$ random variables, so that by our concentration bounds of Corollary 3.9, we have

$$\mathbb{P}\left(\left\|X - X'\right\|_1 \le \frac{d}{2} - t\right) \le \exp\left(-\frac{2t^2}{d}\right).$$

Using a union bound gives the remainder of the result. □

Rewriting the lemma slightly, we may take $\delta \in (0, 1)$. Then

$$\mathbb{P}\left(\exists\, i, j \text{ s.t. } \|X_i - X_j\|_1 < \frac{d}{2} - \sqrt{d \log \frac{1}{\delta} + d \log m}\right) \le \delta.$$

As a consequence of this lemma, we see two things:

(i) If $m \le \exp(d/16)$, or $d \ge 16 \log m$, then taking $\delta \uparrow 1$, there at least exists a codebook $\{x_1, \ldots, x_m\}$ of words that are all separated by at least $d/4$, that is, $\|x_i - x_j\|_1 \ge \frac{d}{4}$ for all i, j.

(ii) By taking $m \le \exp(d/32)$, or $d \ge 32 \log m$, and $\delta = e^{-d/32}$, then with probability at least $1 - e^{-d/32}$—exponentially large in d—a randomly drawn codebook has all its entries separated by at least $\|x_i - x_j\|_1 \ge \frac{d}{4}$.

Summarizing, we have the following result: choose a codebook of m codewords $x_1, \ldots, x_m$ uniformly at random from the hypercube $\mathcal{H}_d = \{0, 1\}^d$ with

$$d \ge \max\left\{32 \log m, \frac{8 \log \frac{m}{\delta}}{(1 - 2\epsilon)^2}\right\}.$$

Then with probability at least $1 - 1/m$ over the draw of the codebook, the probability we make a mistake in transmission of any given symbol i over the channel Q is at most δ.

3.2 Martingale methods

3.3 Entropy and concentration inequalities

In the previous sections, we saw how moment generating functions and related techniques could be used to give bounds on the probability of deviation for fairly simple quantities, such as sums of random variables. In many situations, however, it is desirable to give guarantees for more complex functions. As one example, suppose that we draw a matrix $X \in \mathbb{R}^{m \times n}$, where the entries of X are bounded independent random variables. The operator norm of X, $|||X||| := \sup_{u,v}\{u^\top X v : \|u\|_2 = \|v\|_2 = 1\}$, is one measure of the size of X. We would like to give upper bounds on the probability that $|||X||| \ge \mathbb{E}[|||X|||] + t$ for $t \ge 0$, which the tools of the preceding sections do not address well because of the complicated dependencies on $|||X|||$.

In this section, we will develop techniques to give control over such complex functions. In particular, throughout we let $Z = f(X_1, \ldots, X_n)$ be some function of a sample of independent random variables X_i; we would like to know if Z is concentrated around its mean. We will use deep connections between information theoretic quantities and deviation probabilities to investigate these connections.

First, we give a definition.

Definition 3.3. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. The* ϕ-entropy *of a random variable X is*

$$\mathbb{H}_\phi(X) := \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]), \tag{3.3.1}$$

assuming the relevant expectations exist.

A first example of the ϕ-entropy is the variance:

> **Example 3.20** (Variance as ϕ-entropy)**:** Let $\phi(t) = t^2$. Then $\mathbb{H}_\phi(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathrm{Var}(X)$. ♣

This example is suggestive of the fact that ϕ-entropies may help us to control deviations of random variables from their means. More generally, we have by Jensen's inequality that $\mathbb{H}_\phi(X) \ge 0$ for any convex ϕ; moreover, if ϕ is strictly convex and X is non-constant, then $\mathbb{H}_\phi(X) > 0$. The rough intuition we consider throughout this section is as follows: if a random variable X is tightly concentrated around its mean, then we should have $X \approx \mathbb{E}[X]$ "most" of the time, and so $\mathbb{H}_\phi(X)$ should be small. The goal of this section is to make this claim rigorous.

3.3.1 The Herbst argument

Perhaps unsurprisingly given the focus of these lecture notes, we focus on a specific ϕ, using $\phi(t) = t \log t$, which gives the entropy on which we focus:

$$\mathbb{H}(Z) := \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z], \tag{3.3.2}$$

defined whenever $Z \ge 0$ with probability 1. As our particular focus throughout this chapter, we consider the moment generating function and associated transformation $X \mapsto e^{\lambda X}$. If we know the moment generating function $\varphi_X(\lambda) := \mathbb{E}[e^{\lambda X}]$, then $\varphi_X'(\lambda) = \mathbb{E}[X e^{\lambda X}]$, and so

$$\mathbb{H}(e^{\lambda X}) = \lambda \varphi_X'(\lambda) - \varphi_X(\lambda) \log \varphi_X(\lambda).$$

This suggests—in a somewhat roundabout way we make precise—that control of the entropy $\mathbb{H}(e^{\lambda X})$ should be sufficient for controlling the moment generating function of X.

The Herbst argument makes this rigorous.

Proposition 3.21. *Let X be a random variable and assume that there exists a constant $\sigma^2 < \infty$ such that*

$$\mathbb{H}(e^{\lambda X}) \le \frac{\lambda^2 \sigma^2}{2} \varphi_X(\lambda). \tag{3.3.3}$$

for all $\lambda \in \mathbb{R}$ (respectively, $\lambda \in \mathbb{R}_+$) where $\varphi_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ denotes the moment generating function of X. Then

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

for all $\lambda \in \mathbb{R}$ (respectively, $\lambda \in \mathbb{R}_+$).

Proof Let $\varphi = \varphi_X$ for shorthand. The proof procedes by an integration argument, where we show that $\log \varphi(\lambda) \le \frac{\lambda^2 \sigma^2}{2}$. First, note that

$$\varphi'(\lambda) = \mathbb{E}[X e^{\lambda X}],$$

so that inequality (3.3.3) is equivalent to

$$\lambda \varphi'(\lambda) - \varphi(\lambda) \log \varphi(\lambda) = \mathbb{H}(e^{\lambda X}) \le \frac{\lambda^2 \sigma^2}{2} \varphi(\lambda),$$

and dividing both sides by $\lambda^2\varphi(\lambda)$ yields the equivalent statement

$$\frac{\varphi'(\lambda)}{\lambda\varphi(\lambda)} - \frac{1}{\lambda^2}\log\varphi(\lambda) \le \frac{\sigma^2}{2}.$$

But by inspection, we have

$$\frac{\partial}{\partial\lambda}\frac{1}{\lambda}\log\varphi(\lambda) = \frac{\varphi'(\lambda)}{\lambda\varphi(\lambda)} - \frac{1}{\lambda^2}\log\varphi(\lambda).$$

Moreover, we have that

$$\lim_{\lambda\to 0}\frac{\log\varphi(\lambda)}{\lambda} = \lim_{\lambda\to 0}\frac{\log\varphi(\lambda)-\log\varphi(0)}{\lambda} = \frac{\varphi'(0)}{\varphi(0)} = \mathbb{E}[X].$$

Integrating from 0 to any λ_0, we thus obtain

$$\frac{1}{\lambda_0}\log\varphi(\lambda_0) - \mathbb{E}[X] = \int_0^{\lambda_0}\left[\frac{\partial}{\partial\lambda}\frac{1}{\lambda}\log\varphi(\lambda)\right]d\lambda \le \int_0^{\lambda_0}\frac{\sigma^2}{2}d\lambda = \frac{\sigma^2\lambda_0}{2}.$$

Multiplying each side by λ_0 gives

$$\log\mathbb{E}[e^{\lambda_0(X-\mathbb{E}[X])}] = \log\mathbb{E}[e^{\lambda_0 X}] - \lambda_0\mathbb{E}[X] \le \frac{\sigma^2\lambda_0^2}{2},$$

as desired. □

It is possible to give a similar argument for sub-exponential random variables, which allows us to derive Bernstein-type bounds, of the form of Corollary 3.17, but using the entropy method. In particular, in the exercises, we show the following result.

Proposition 3.22. *Assume that there exist positive constants b and σ such that*

$$\mathbb{H}(e^{\lambda X}) \le \lambda^2\left[b\varphi_X'(\lambda) + \varphi_X(\lambda)(\sigma^2 - b\mathbb{E}[X])\right] \tag{3.3.4a}$$

for all $\lambda \in [0, 1/b)$. Then X satisfies the sub-exponential bound

$$\log\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le \frac{\sigma^2\lambda^2}{[1-b\lambda]_+} \tag{3.3.4b}$$

for all $\lambda \ge 0$.

An immediate consequence of this proposition is that any random variable satisfying the entropy bound (3.3.4a) is $(2\sigma^2, 2b)$-sub-exponential. As another immediate consequence, we obtain the concentration guarantee

$$\mathbb{P}(X \ge \mathbb{E}[X] + t) \le \exp\left(-\frac{1}{4}\min\left\{\frac{t^2}{\sigma^2}, \frac{t}{b}\right\}\right)$$

as in Proposition 3.15.

3.3.2 Tensorizing the entropy

A benefit of the moment generating function approach we took in the prequel is the excellent behavior of the moment generating function for sums. In particular, the fact that $\varphi_{X_1+\cdots+X_n}(\lambda) = \prod_{i=1}^n \varphi_{X_i}(\lambda)$ allowed us to derive sharper concentration inequalities, and we were only required to work with *marginal* distributions of the X_i, computing only the moment generating functions of individual random variables rather than characteristics of the entire sum. One advantage of the entropy-based tools we develop is that they allow similar tensorization—based on the chain rule identities of Chapter 2 for entropy, mutual information, and KL-divergence—for substantially more complex functions. Our approach here mirrors that of Boucheron, Lugosi, and Massart [4].

With that in mind, we now present a series of inequalities that will allow us to take this approach. For shorthand throughout this section, we let

$$X_{\setminus i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

be the collection of all variables except X_i. Our first result is a consequence of the chain rule for entropy and is known as Han's inequality.

Proposition 3.23 (Han's inequality)**.** *Let $X_1, \ldots, X_n$ be discrete random variables. Then*

$$H(X_1^n) \le \frac{1}{n-1} \sum_{i=1}^n H(X_{\setminus i}).$$

Proof The proof is a consequence of the chain rule for entropy and that conditioning reduces entropy. We have

$$H(X_1^n) = H(X_i \mid X_{\setminus i}) + H(X_{\setminus i}) \le H(X_i \mid X_1^{i-1}) + H(X_{\setminus i}).$$

Writing this inequality for each $i = 1, \ldots, n$, we obtain

$$nH(X_1^n) \le \sum_{i=1}^n H(X_{\setminus i}) + \sum_{i=1}^n H(X_i \mid X_1^{i-1}) = \sum_{i=1}^n H(X_{\setminus i}) + H(X_1^n),$$

and subtracting $H(X_1^n)$ from both sides gives the result. □

We also require a divergence version of Han's inequality, which will allow us to relate the entropy $\mathbb{H}$ of a random variable to divergences and other information-theoretic quantities. Let $\mathcal{X}$ be an arbitrary space, and let Q be a distribution over $\mathcal{X}^n$ and $P = P_1 \times \cdots \times P_n$ be a product distribution on the same space. For $A \subset \mathcal{X}^{n-1}$, define the marginal densities

$$Q^{(i)}(A) := Q(X_{\setminus i} \in A) \quad \text{and} \quad P^{(i)}(A) = P(X_{\setminus i} \in A).$$

We then obtain the tensorization-type Han's inequality for relative entropies.

Proposition 3.24. *With the above definitions,*

$$D_{\rm kl}\left(Q \| P\right) \le \sum_{i=1}^n \left[D_{\rm kl}\left(Q \| P\right) - D_{\rm kl}\left(Q^{(i)} \| P^{(i)}\right)\right].$$

Proof We have seen earlier in the notes (recall the definition (2.2.1) of the KL divergence as a supremum over all quantizers and the surrounding discussion) that it is no loss of generality to assume that $\mathcal{X}$ is discrete. Thus, noting that the probability mass functions

$$q^{(i)}(x_{\setminus i}) = \sum_x q(x_1^{i-1}, x, x_{i+1}^n) \ \text{ and } \ p^{(i)}(x_{\setminus i}) = \prod_{j \neq i} p_j(x_j),$$

we have that Han's inequality (Proposition 3.23) is equivalent to

$$(n-1) \sum_{x_1^n} q(x_1^n) \log q(x_1^n) \ge \sum_{i=1}^n \sum_{x_{\setminus i}} q^{(i)}(x_{\setminus i}) \log q^{(i)}(x_{\setminus i}).$$

Now, by subtracting $q(x_1^n) \log p(x_1^n)$ from both sides of the preceding display, we obtain

$$\begin{aligned} (n-1) D_{\rm kl}\left(Q \| P\right) &= (n-1) \sum_{x_1^n} q(x_1^n) \log q(x_1^n) - (n-1) \sum_{x_1^n} q(x_1^n) \log p(x_1^n) \\ &\ge \sum_{i=1}^n \sum_{x_{\setminus i}} q^{(i)}(x_{\setminus i}) \log q^{(i)}(x_{\setminus i}) - (n-1) \sum_{x_1^n} q(x_1^n) \log p(x_1^n). \end{aligned}$$

We expand the final term. Indeed, by the product nature of the distributions p, we have

$$\begin{aligned} (n-1) \sum_{x_1^n} q(x_1^n) \log p(x_1^n) &= (n-1) \sum_{x_1^n} q(x_1^n) \sum_{i=1}^n \log p_i(x_i) \\ &= \sum_{i=1}^n \sum_{x_1^n} q(x_1^n) \underbrace{\sum_{j \neq i} \log p_i(x_i)}_{= \log p^{(i)}(x_{\setminus i})} = \sum_{i=1}^n \sum_{x_{\setminus i}} q^{(i)}(x_{\setminus i}) \log p^{(i)}(x_{\setminus i}). \end{aligned}$$

Noting that

$$\sum_{x_{\setminus i}} q^{(i)}(x_{\setminus i}) \log q^{(i)}(x_{\setminus i}) - \sum_{x_{\setminus i}} q^{(i)}(x_{\setminus i}) \log p^{(i)}(x_{\setminus i}) = D_{\rm kl}\left(Q^{(i)} \| P^{(i)}\right)$$

and rearranging gives the desired result. □

Finally, we will prove the main result of this subsection: a tensorization identity for the entropy $\mathbb{H}(Y)$ for an arbitrary random variable Y that is a function of n independent random variables. For this result, we use a technique known as *tilting*, in combination with the two variants of Han's inequality we have shown, to obtain the result. The tilting technique is one used to transform problems of random variables into one of distributions, allowing us to bring the tools of information and entropy to bear more directly. This technique is a common one, and used frequently in large deviation theory, statistics, for heavy-tailed data, amont other areas. More concretely, let $Y = f(X_1, \ldots, X_n)$ for some non-negative function f. Then we may always define a tilted density

$$q(x_1, \ldots, x_n) := \frac{f(x_1, \ldots, x_n) p(x_1, \ldots, x_n)}{\mathbb{E}_P[f(X_1, \ldots, X_n)]} \tag{3.3.5}$$

which, by inspection, satisfies $\int q(x_1^n) = 1$ and $q \ge 0$. In our context, if $f \approx$ constant under the distribution P, then we should have $f(x_1^n) p(x_1^n) \approx c p(x_1^n)$ and so $D_{\rm kl}\left(Q \| P\right)$ should be small; we can make this rigorous via the following tensorization theorem.

Theorem 3.25. *Let $X_1, \ldots, X_n$ be independent random variables and $Y = f(X_1^n)$, where f is a non-negative function. Define $\mathbb{H}(Y \mid X_{\setminus i}) = \mathbb{E}[Y \log Y \mid X_{\setminus i}]$. Then*

$$\mathbb{H}(Y) \le \mathbb{E}\Bigg[\sum_{i=1}^n \mathbb{H}(Y \mid X_{\setminus i})\Bigg]. \tag{3.3.6}$$

Proof Inequality (3.3.6) holds for Y if and only if holds identically for cY for any $c > 0$, so we assume without loss of generality that $\mathbb{E}_P[Y] = 1$. We thus obtain that $\mathbb{H}(Y) = \mathbb{E}[Y \log Y] = \mathbb{E}[\phi(Y)]$, where assign $\phi(t) = t \log t$. Let P have density p with respect to a base measure μ. Then by defining the tilted distribution (density) $q(x_1^n) = f(x_1^n)p(x_1^n)$, we have $Q(\mathcal{X}^n) = 1$, and moreover, we have

$$D_{\rm kl}\left(Q \| P\right) = \int q(x_1^n) \log \frac{q(x_1^n)}{p(x_1^n)} d\mu(x_1^n) = \int f(x_1^n)p(x_1^n) \log f(x_1^n) d\mu(x_1^n) = \mathbb{E}_P[Y \log Y] = \mathbb{H}(Y).$$

Similarly, if $\phi(t) = t \log t$, then

$$\begin{aligned}
&D_{\rm kl}\left(Q^{(i)} \| P^{(i)}\right) \\
&= \int_{\mathcal{X}^{n-1}} \left(\int f(x_1^{i-1}, x, x_{i+1}^n) p_i(x) d\mu(x)\right) \log \frac{p^{(i)}(x_{\setminus i}) \int f(x_1^{i-1}, x, x_{i+1}^n) p_i(x) d\mu(x)}{p^{(i)}(x_{\setminus i})} p^{(i)}(x_{\setminus i}) d\mu(x_{\setminus i}) \\
&= \int_{\mathcal{X}^{n-1}} \mathbb{E}[Y \mid x_{\setminus i}] \log \mathbb{E}[Y \mid x_{\setminus i}] p^{(i)}(x_{\setminus i}) d\mu(x_{\setminus i}) \\
&= \mathbb{E}[\phi(\mathbb{E}[Y \mid X_{\setminus i}])].
\end{aligned}$$

The tower property of expectations then yields that

$$\mathbb{E}[\phi(Y)] - \mathbb{E}[\phi(\mathbb{E}[Y \mid X_{\setminus i}])] = \mathbb{E}[\mathbb{E}[\phi(Y) \mid X_{\setminus i}] - \phi(\mathbb{E}[Y \mid X_{\setminus i}])] = \mathbb{E}[\mathbb{H}(Y \mid X_{\setminus i})].$$

Using Han's inequality for relative entropies (Proposition 3.23) then immediately gives

$$\mathbb{H}(Y) = D_{\rm kl}\left(Q \| P\right) \le \sum_{i=1}^n \left[D_{\rm kl}\left(Q \| P\right) - D_{\rm kl}\left(Q^{(i)} \| P^{(i)}\right)\right] = \sum_{i=1}^n \mathbb{E}[\mathbb{H}(Y \mid X_{\setminus i})],$$

which is our desired result. □

Theorem 3.25 shows that if we can show that individually the conditional entropies $\mathbb{H}(Y \mid X_{\setminus i})$ are not too large, then the Herbst argument (Proposition 3.21 or its variant Proposition 3.22) allows us to provide strong concentration inequalities for general random variables Y.

Examples and consequences

We now show how to use some of the preceding results to derive strong concentration inequalities, showing as well how we may give convergence guarantees for a variety of procedures using these techniques.

We begin with our most straightforward example, which is the bounded differences inequality. In particular, we consider an arbitrary function f of n independent random variables, and we assume that for all $x_{1:n} = (x_1, \ldots, x_n)$, we have the bounded differences condition:

$$\sup_{x \in \mathcal{X}, x' \in \mathcal{X}} \left|f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x', x_{i+1}, \ldots, x_n)\right| \le c_i \ \text{ for all } x_{\setminus i}. \tag{3.3.7}$$

Then we have the following result.

Proposition 3.26 (Bounded differences)**.** *Assume that f satisfies the bounded differences condition* (3.3.7)*, where $\frac{1}{4}\sum_{i=1}^n c_i^2 \le \sigma^2$. Let X_i be independent. Then $Y = f(X_1, \ldots, X_n)$ is σ^2-sub-Gaussian.*

Proof We use a similar integration argument to the Herbst argument of Proposition 3.21, and we apply the tensorization inequality (3.3.6). First, let U be an arbitrary random variable taking values in $[a, b]$. We claim that if $\varphi_U(\lambda) = \mathbb{E}[e^{\lambda U}]$ and $\psi(\lambda) = \log \varphi_U(\lambda)$ is its cumulant generating function, then

$$\frac{\mathbb{H}(e^{\lambda U})}{\mathbb{E}[e^{\lambda U}]} \le \frac{\lambda^2(b-a)^2}{8}. \tag{3.3.8}$$

To see this, note that

$$\frac{\partial}{\partial\lambda}[\lambda\psi'(\lambda) - \psi(\lambda)] = \psi''(\lambda), \quad \text{so} \quad \lambda\psi'(\lambda) - \psi(\lambda) = \int_0^\lambda t\psi''(t)dt \le \frac{\lambda^2(b-a)^2}{8},$$

where we have used the homework exercise **XXXX** (recall Hoeffding's Lemma, Example 3.6), to argue that $\psi''(t) \le \frac{(b-a)^2}{4}$ for all t. Recalling that

$$\mathbb{H}(e^{\lambda U}) = \lambda\varphi_U'(\lambda) - \varphi_U(\lambda)\psi(\lambda) = \left[\lambda\psi'(\lambda) - \psi(\lambda)\right]\varphi_U(\lambda)$$

gives inequality (3.3.8).

Now we apply the tensorization identity. Let $Z = e^{\lambda Y}$. Then we have

$$\mathbb{H}(Z) \le \mathbb{E}\bigg[\sum_{i=1}^n \mathbb{H}(Z \mid X_{\setminus i})\bigg] \le \mathbb{E}\bigg[\sum_{i=1}^n \frac{c_i^2\lambda^2}{8}\mathbb{E}[e^{\lambda Z} \mid X_{\setminus i}]\bigg] = \sum_{i=1}^n \frac{c_i^2\lambda^2}{8}\mathbb{E}[e^{\lambda Z}].$$

Applying the Herbst argument gives the final result. □

As an immediate consequence of this inequality, we obtain the following dimension independent concentration inequality.

Example 3.27: Let $X_1, \ldots, X_n$ be independent vectors in $\mathbb{R}^d$, where d is arbitrary, and assume that $\|X_i\|_2 \le c_i$ with probability 1. (This could be taken to be a general Hilbert space with no loss of generality.) We claim that if we define

$$\sigma^2 := \sum_{i=1}^n c_i^2, \quad \text{then} \quad \mathbb{P}\bigg(\bigg\|\sum_{i=1}^n X_i\bigg\|_2 \ge t\bigg) \le \exp\left(-2\frac{[t - \sqrt{\sigma}]_+^2}{\sigma^2}\right).$$

Indeed, we have that $Y = \|\sum_{i=1}^n X_i\|_2$ satisfies the bounded differences inequality with parameters c_i, and so

$$\begin{aligned}\mathbb{P}\bigg(\bigg\|\sum_{i=1}^n X_i\bigg\|_2 \ge t\bigg) &= \mathbb{P}\bigg(\bigg\|\sum_{i=1}^n X_i\bigg\|_2 - \mathbb{E}\bigg\|\sum_{i=1}^n X_i\bigg\|_2 \ge t - \mathbb{E}\bigg\|\sum_{i=1}^n X_i\bigg\|_2\bigg) \\ &\le \exp\left(-2\frac{[t - \mathbb{E}\|\sum_{i=1}^n X_i\|_2]_+^2}{\sum_{i=1}^n c_i^2}\right).\end{aligned}$$

Noting that $\mathbb{E}[\|\sum_{i=1}^n X_i\|_2] \le \sqrt{\mathbb{E}[\|\sum_{i=1}^n X_i\|_2^2]} = \sqrt{\sum_{i=1}^n \mathbb{E}[\|X_i\|_2^2]}$ gives the result. ♣

3.3.3 Concentration of convex functions

We provide a second theorem on the concentration properties of a family of functions that are quite useful, for which other concentration techniques do not appear to give results. In particular, we say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *separately convex* if for each $i \in \{1, \ldots, n\}$ and all $x_{\setminus i} \in \mathbb{R}^{n-1}$ (or the domain of f), we have that

$$x \mapsto f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$$

is convex. We also recall that a function is L-Lipschitz if $|f(x) - f(y)| \le \|x - y\|_2$ for all $x, y \in \mathbb{R}^n$; any L-Lipschitz function is almost everywhere differentiable, and is L-Lipschitz if and only if $\|\nabla f(x)\|_2 \le L$ for (almost) all x. With these preliminaries in place, we have the following result.

Theorem 3.28. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [a, b]$ for all i. Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is separately convex and L-Lipschitz with respect to the $\|\cdot\|_2$ norm. Then*

$$\mathbb{E}[\exp(\lambda(f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]))] \le \exp\left(\lambda^2 (b-a)^2 L^2\right) \quad \textit{for all } \lambda \ge 0.$$

We defer the proof of the theorem temporarily, giving two example applications. The first is to the matrix concentration problem that motivates the beginning of this section.

Example 3.29: Let $X \in \mathbb{R}^{m \times n}$ be a matrix with independent entries, where $X_{ij} \in [-1, 1]$ for all i, j, and let $|||\cdot|||$ denote the operator norm on matrices, that is, $|||A||| = \sup_{u,v}\{u^\top A v : \|u\|_2 \le 1, \|v\|_2 \le 1\}$. Then Theorem 3.28 implies

$$\mathbb{P}(|||X||| \ge \mathbb{E}[|||X|||] + t) \le \exp\left(-\frac{t^2}{16}\right)$$

for all $t \ge 0$. Indeed, we first observe that

$$|\, |||X||| - |||Y|||\, | \le |||X - Y||| \le \|X - Y\|_{\rm Fr},$$

where $\|\cdot\|_{\rm Fr}$ denotes the Frobenius norm of a matrix. Thus the matrix operator norm is 1-Lipschitz. Therefore, we have by Theorem 3.28 and the Chernoff bound technique that

$$\mathbb{P}(|||X||| \ge \mathbb{E}[|||X|||] + t) \le \exp(4\lambda^2 - \lambda t)$$

for all $\lambda \ge 0$. Taking $\lambda = t/8$ gives the desired result. ♣

As a second example, we consider *Rademacher complexity.* These types of results are important for giving generalization bounds in a variety of statistical algorithms, and form the basis of a variety of concentration and convergence results. We defer further motivation of these ideas to subsequent chapters, just mentioning here that we can provide strong concentration guarantees for Rademacher complexity or Rademacher chaos.

Example 3.30: Let $\mathcal{A} \subset \mathbb{R}^n$ be any collection of vectors. The the *Rademacher complexity* of the class $\mathcal{A}$ is

$$R_n(\mathcal{A}) := \mathbb{E}\left[\sup_{a \in \mathcal{A}} \sum_{i=1}^n a_i \varepsilon_i\right], \tag{3.3.9}$$

where ε_i are i.i.d. Rademacher (sign) variables. Let $\widehat{R}_n(\mathcal{A}) = \sup_{a\in\mathcal{A}} \sum_{i=1}^n a_i\varepsilon_i$ denote the empirical version of this quantity. We claim that

$$\mathbb{P}(\widehat{R}_n(\mathcal{A}) \ge R_n(\mathcal{A}) + t) \le \exp\left(-\frac{t^2}{16\,\mathrm{diam}(\mathcal{A})^2}\right),$$

where $\mathrm{diam}(\mathcal{A}) := \sup_{a\in\mathcal{A}} \|a\|_2$. Indeed, we have that $\varepsilon \mapsto \sup_{a\in\mathcal{A}} a^\top\varepsilon$ is a convex function, as it is the maximum of a family of linear functions. Moreover, it is Lipschitz, with Lipschitz constant bounded by $\sup_{a\in\mathcal{A}} \|a\|_2$. Applying Theorem 3.28 as in Example **??** gives the result. ♣

Proof of Theorem 3.28 The proof relies on our earlier tensorization identity and a symmetrization lemma.

Lemma 3.31. *Let $X, Y \overset{\text{i.i.d.}}{\sim} P$ be independent. Then for any function $g : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{H}(e^{\lambda g(X)}) \le \lambda^2 \mathbb{E}[(g(X) - g(Y))^2 e^{\lambda g(X)} \mathbf{1}\{g(X) \ge g(Y)\}] \quad \text{for } \lambda \ge 0.$$

Moreover, if g is convex, then

$$\mathbb{H}(e^{\lambda g(X)}) \le \lambda^2 \mathbb{E}[(X - Y)^2 (g'(X))^2 e^{\lambda g(X)}] \quad \text{for } \lambda \ge 0.$$

Proof For the first result, we use the convexity of the exponential in an essential way. In particular, we have

$$\begin{aligned}\mathbb{H}(e^{\lambda g(X)}) &= \mathbb{E}[\lambda g(X) e^{\lambda g(X)}] - \mathbb{E}[e^{\lambda g(X)}] \log \mathbb{E}[e^{\lambda g(Y)}] \\ &\le \mathbb{E}[\lambda g(X) e^{\lambda g(X)}] - \mathbb{E}[e^{\lambda g(X)} \lambda g(Y)],\end{aligned}$$

because log is concave and $e^x \ge 0$. Using symmetry, that is, that $g(X) - g(Y)$ has the same distribution as $g(Y) - g(X)$, we then find

$$\mathbb{H}(e^{\lambda g(X)}) \le \frac{1}{2}\mathbb{E}[\lambda(g(X)-g(Y))(e^{\lambda g(X)}-e^{\lambda g(Y)})] = \mathbb{E}[\lambda(g(X)-g(Y))(e^{\lambda g(X)}-e^{\lambda g(Y)})\mathbf{1}\{g(X) \ge g(Y)\}].$$

Now we use the classical first order convexity inequality—that a convex function f satisfies $f(t) \ge f(s) + f'(s)(t-s)$ for all t and s, Theorem A.14 in the appendices—which gives that $e^t \ge e^s + e^s(t-s)$ for all s and t. Rewriting, we have $e^s - e^t \le e^s(s-t)$, and whenever $s \ge t$, we have $(s-t)(e^s - e^t) \le e^s(s-t)^2$. Replacing s and t with $\lambda g(X)$ and $\lambda g(Y)$, respectively, we obtain

$$\lambda(g(X) - g(Y))(e^{\lambda g(X)} - e^{\lambda g(Y)})\mathbf{1}\{g(X) \ge g(Y)\} \le \lambda^2 (g(X) - g(Y))^2 e^{\lambda g(X)} \mathbf{1}\{g(X) \ge g(Y)\}.$$

This gives the first inequality of the lemma.

To obtain the second inequality, note that if g is convex, then whenever $g(x) - g(y) \ge 0$, we have $g(y) \ge g(x) + g'(x)(y - x)$, or $g'(x)(x - y) \ge g(x) - g(y) \ge 0$. In particular,

$$(g(X) - g(Y))^2 \mathbf{1}\{g(X) \ge g(Y)\} \le (g'(X)(X - Y))^2,$$

which gives the second result. □

Returning to the main thread of the proof, we note that the separate convexity of f and the tensorization identity of Theorem 3.25 imply

$$\mathbb{H}(e^{\lambda f(X_{1:n})}) \le \mathbb{E}\bigg[\sum_{i=1}^n \mathbb{H}(e^{\lambda f(X_{1:n})} \mid X_{\setminus i})\bigg] \le \mathbb{E}\bigg[\sum_{i=1}^n \lambda^2 \mathbb{E}\left[(X_i - Y_i)^2 \left(\frac{\partial}{\partial x_i} f(X_{1:n})\right)^2 e^{\lambda f(X_{1:n})} \mid X_{\setminus i}\right]\bigg],$$

where Y_i are independent copies of the X_i. Now, we use that $(X_i - Y_i)^2 \le (b-a)^2$ and the definition of the partial derivative to obtain

$$\mathbb{H}(e^{\lambda f(X_{1:n})}) \le \lambda^2 (b-a)^2 \mathbb{E}[\|\nabla f(X_{1:n})\|_2^2 \, e^{\lambda f(X_{1:n}))}].$$

Noting that $\|\nabla f(X)\|_2^2 \le L^2$, and applying the Herbst argument, gives the result. □

3.A Technical proofs

3.A.1 Proof of Theorem 3.10

(1) implies (2) Let $K_1 = 1$. Using the change of variables identity that for a nonnegative random variable Z and any $k \ge 1$ we have $\mathbb{E}[Z^k] = k \int_0^\infty t^{k-1} \mathbb{P}(Z \ge t) dt$, we find

$$\mathbb{E}[|X|^k] = k \int_0^\infty t^{k-1} \mathbb{P}(|X| \ge t) dt \le 2k \int_0^\infty t^{k-1} \exp\left(-\frac{t^2}{\sigma^2}\right) dt = k\sigma^k \int_0^\infty u^{k/2-1} e^{-u} du,$$

where for the last inequality we made the substitution $u = t^2/\sigma^2$. Noting that this final integral is $\Gamma(k/2)$, we have $\mathbb{E}[|X|^k] \le k\sigma^k \Gamma(k/2)$. Because $\Gamma(s) \le s^s$ for $s \ge 1$, we obtain

$$\mathbb{E}[|X|^k]^{1/k} \le k^{1/k} \sigma \sqrt{k/2} \le e^{1/e} \sigma \sqrt{k}.$$

Thus (2) holds with $K_2 = e^{1/e}$.

(2) implies (3) Let $\sigma = \|X\|_{\psi_2} = \sup_{k \ge 1} k^{-\frac{1}{2}} \mathbb{E}[|X|^k]^{1/k}$, so that $K_2 = 1$ and $\mathbb{E}[|X|^k] \le k^{\frac{k}{2}} \sigma$ for all k. For $K_3 \in \mathbb{R}_+$, we thus have

$$\mathbb{E}[\exp(X^2/(K_3 \sigma^2))] = \sum_{k=0}^\infty \frac{\mathbb{E}[X^{2k}]}{k! K_3^{2k} \sigma^{2k}} \le \sum_{k=0}^\infty \frac{\sigma^{2k} (2k)^k}{k! K_3^{2k} \sigma^{2k}} \overset{(i)}{\le} \sum_{k=0}^\infty \left(\frac{2e}{K_3^2}\right)^k$$

where inequality (i) follows because $k! \ge (k/e)^k$, or $1/k! \le (e/k)^k$. Noting that $\sum_{k=0}^\infty \alpha^k = \frac{1}{1-\alpha}$, we obtain (3) by taking $K_3 = e\sqrt{2/(e-1)} \approx 2.933$.

(3) implies (4) Let us take $K_3 = 1$. We claim that (4) holds with $K_4 = \frac{3}{4}$. We prove this result for both small and large λ. First, note the (highly non-standard, but true!) inequality that $e^x \le x + e^{\frac{9x^2}{16}}$ for all x. Then we have

$$\mathbb{E}[\exp(\lambda X)] \le \underbrace{\mathbb{E}[\lambda X]}_{=0} + \mathbb{E}\left[\exp\left(\frac{9\lambda^2 X^2}{16}\right)\right]$$

Now note that for $|\lambda| \le \frac{4}{3\sigma}$, we have $9\lambda^2\sigma^2/16 \le 1$, and so by Jensen's inequality,

$$\mathbb{E}\left[\exp\left(\frac{9\lambda^2 X^2}{16}\right)\right] = \mathbb{E}\left[\exp(X^2/\sigma^2)^{\frac{9\lambda^2\sigma^2}{16}}\right] \le e^{\frac{9\lambda^2\sigma^2}{16}}.$$

For large λ, we use the simpler Fenchel-Young inequality, that is, that $\lambda x \le \frac{\lambda^2}{2c} + \frac{cx^2}{2}$, valid for all $c \ge 0$. Then we have for any $0 \le c \le 2$ that

$$\mathbb{E}[\exp(\lambda X)] \le e^{\frac{\lambda^2\sigma^2}{2c}} \mathbb{E}\left[\exp\left(\frac{cX^2}{2\sigma^2}\right)\right] \le e^{\frac{\lambda^2\sigma^2}{2c}} e^{\frac{c}{2}},$$

where the final inequality follows from Jensen's inequality. If $|\lambda| \ge \frac{4}{3\sigma}$, then $\frac{1}{2} \le \frac{9}{32}\lambda^2\sigma^2$, and we have

$$\mathbb{E}[\exp(\lambda X)] \le \inf_{c \in [0,2]} e^{[\frac{1}{2c} + \frac{9c}{32}]\lambda^2\sigma^2} = \exp\left(\frac{3\lambda^2\sigma^2}{4}\right).$$

(4) implies (1) This is the content of Proposition 3.7, with $K_4 = \frac{1}{2}$ and $K_1 = 2$.

3.A.2 Proof of Theorem 3.14

(1) implies (2) As in the proof of Theorem 3.10, we use that for a nonnegative random variable Z we have $\mathbb{E}[Z^k] = k\int_0^\infty t^{k-1}\mathbb{P}(Z \ge t)dt$. Let $K_1 = 1$. Then

$$\mathbb{E}[|X|^k] = k\int_0^\infty t^{k-1}\mathbb{P}(|X| \ge t)dt \le 2k\int_0^\infty t^{k-1}\exp(-t/\sigma)dt = 2k\sigma^k \int_0^\infty u^{k-1}\exp(-u)du,$$

where we used the substitution $u = t/\sigma$. Thus we have $\mathbb{E}[|X|^k] \le 2\Gamma(k+1)\sigma^k$, and using $\Gamma(k+1) \le k^k$ yields $\mathbb{E}[|X|^k]^{1/k} \le 2^{1/k}k\sigma$, so that (2) holds with $K_2 \le 2$.

(2) implies (3) Let $K_2 = 1$, and note that

$$\mathbb{E}[\exp(X/(K_3\sigma))] = \sum_{k=0}^\infty \frac{\mathbb{E}[X^k]}{K_3^k\sigma^k k!} \le \sum_{k=0}^\infty \frac{k^k}{k!}\cdot\frac{1}{K_3^k} \overset{(i)}{\le} \sum_{k=0}^\infty \left(\frac{e}{K_3}\right)^k,$$

where inequality (i) used that $k! \ge (k/e)^k$. Taking $K_3 = e^2/(e-1) < 5$ gives the result.

(3) implies (1) If $\mathbb{E}[\exp(X/\sigma)] \le e$, then for $t \ge 0$

$$\mathbb{P}(X \ge t) \le \mathbb{E}[\exp(X/\sigma)]e^{-t/\sigma} \le e^{1-t/\sigma}.$$

With the same result for the negative tail, we have

$$\mathbb{P}(|X| \ge t) \le 2e^{1-t/\sigma} \wedge 1 \le 2e^{-\frac{2t}{5\sigma}},$$

so that (1) holds with $K_1 = \frac{5}{2}$.

(2) if and only if (4) Thus, we see that up to constant numerical factors, the definition $\left\|X\right\|_{\psi_1} = \sup_{k \ge 1} k^{-1}\mathbb{E}[|X|^k]^{1/k}$ has the equivalent statements

$$\mathbb{P}(|X| \ge t) \le 2\exp(-t/(K_1 \left\|X\right\|_{\psi_1})) \quad \text{and} \quad \mathbb{E}[\exp(X/(K_3 \left\|X\right\|_{\psi_1}))] \le e.$$

Now, let us assume that (2) holds with $K_2 = 1$, so that $\sigma = \left\|X\right\|_{\psi_1}$ and that $\mathbb{E}[X] = 0$. Then we have $\mathbb{E}[X^k] \le k^k \left\|X\right\|_{\psi_1}^k$, and

$$\mathbb{E}[\exp(\lambda X)] = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}[X^k]}{k!} \le 1 + \sum_{k=2}^{\infty} \lambda^k \left\|X\right\|_{\psi_1}^k \cdot \frac{k^k}{k!} \le 1 + \sum_{k=2}^{\infty} \lambda^k \left\|X\right\|_{\psi_1}^k e^k,$$

the final inequality following because $k! \ge (k/e)^k$. Now, if $|\lambda| \le \frac{1}{2e\left\|X\right\|_{\psi_1}}$, then we have

$$\mathbb{E}[\exp(\lambda X)] \le 1 + \lambda^2 e^2 \left\|X\right\|_{\psi_1} \sum_{k=0}^{\infty} (\lambda \left\|X\right\|_{\psi_1} e)^k \le 1 + 2e^2 \left\|X\right\|_{\psi_1}^2 \lambda^2,$$

as the final sum is at most $\sum_{k=0}^{\infty} 2^{-k} = 2$. Using $1 + x \le e^x$ gives that (2) implies (4). For the opposite direction, we may simply use that if (4) holds with $K_4 = 1$ and $K_4' = 1$, then $\mathbb{E}[\exp(X/\sigma)] \le \exp(1)$, so that (3) holds.

Bibliography

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66:671–687, 2003.

[2] N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[3] A. Andoni, M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni. Locality-sensitive hashing using stable distributions. In T. Darrell, P. Indyk, and G. Shakhnarovich, editors, *Nearest Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press, 2006.

[4] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[5] V. Buldygin and Y. Kozachenko. *Metric Characterization of Random Variables and Random Processes*, volume 188 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[6] S. Dasgupta and A. Gupta. An elementray proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2002.

[7] P. Indyk. Nearest neighbors in high-dimensional spaces. In *Handbook of Discrete and Computational Geometry*. CRC Press, 2004.

[8] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing*, 1998.

[9] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[10] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

Chapter 4

Beginning generalization bounds and complexity classes

Now that we have explored a variety of concentration inequalities, we show how to put them to use in demonstrating that a variety of estimation and learning procedures have nice convergence properties, focusing on some standard tasks from machine learning.

Throughout this section, we will focus on so-called *empirical risk minimization* procedures and problems. The setting is as follows: we have a sample $Z_1, \ldots, Z_n \in \mathcal{Z}$ drawn i.i.d. according to some (unknown) distribution P, and we have a collection of functions $\mathcal{F}$ from which we wish to select an f that "fits" the data well, according to some loss measure $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$. That is, we wish to find a function $f \in \mathcal{F}$ minimizing the *risk*

$$R(f) := \mathbb{E}_P[\ell(f, Z)]. \tag{4.0.1}$$

In general, however, we only have access to the risk via the empirical distribution of the Z_i, and we often choose f by minimizing the empirical risk

$$\widehat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i). \tag{4.0.2}$$

As written, this formulation is quite abstract, so we provide a few examples to make it somewhat more concrete.

Example 4.1 (Binary classification problems)**:** One standard problem—still abstract—that motivates the formulation (4.0.1) is the *binary classification problem.* Here the data Z_i come in pairs (X, Y), where $X \in \mathcal{X}$ is some set of covariates (independent variables) and $Y \in \{-1, 1\}$ is the label of example X. The function class $\mathcal{F}$ consists of functions $f : \mathcal{X} \to \mathbb{R}$, and the goal is to find a function f such that

$$\mathbb{P}(\operatorname{sign}(f(X)) \neq Y)$$

is small, that is, minimizing the risk $\mathbb{E}[\ell(f, Z)]$ where the loss is the 0-1 loss, $\ell(f, (x, y)) = \mathbf{1}\{f(x)y \leq 0\}$. ♣

Example 4.2 (Multiclass classification)**:** The multiclass classifcation problem is identical to the binary problem, but instead of $Y \in \{-1, 1\}$ we assume that $Y \in [k] = \{1, \ldots, k\}$ for some $k \geq 2$, and the function class $\mathcal{F}$ consists of (a subset of) functions $f : \mathcal{X} \to \mathbb{R}^k$. The goal is to find a function f such that, if $Y = y$ is the correct label for a datapoint x, then $f_y(x) > f_l(x)$ for all $l \neq y$. That is, we wish to find $f \in \mathcal{F}$ minimizing

$$\mathbb{P}\left(\exists\, l \neq Y \text{ such that } f_l(X) \geq f_Y(X)\right).$$

In this case, the loss function is the zero-one loss $\ell(f, (x, y)) = \mathbf{1}\left\{\max_{l \neq y} f_l(x) \geq f_y(x)\right\}$. ♣

Example 4.3 (Binary classification with linear functions)**:** In the standard statistical learning setting, the data x belong to $\mathbb{R}^d$, and we assume that our function class $\mathcal{F}$ is indexed by a set $\Theta \subset \mathbb{R}^d$, so that $\mathcal{F} = \{f_\theta : f_\theta(x) = \theta^\top x, \theta \in \Theta\}$. In this case, we may use the zero-one loss, the convex hinge loss, or the (convex) logistic loss, which are variously $\ell_{\rm zo}(f_\theta, (x, y)) := \mathbf{1}\left\{y\theta^\top x \leq 0\right\}$, and the convex losses

$$\ell_{\rm hinge}(f_\theta, (x, y)) = \left[1 - yx^\top \theta\right]_+ \quad \text{and} \quad \ell_{\rm logit}(f_\theta, (x, y)) = \log(1 + \exp(-yx^\top \theta)).$$

The hinge and logistic losses, as they are convex, are substantially computationally easier to work with, and they are common choices in applications. ♣

The main motivating question that we ask in this question is the following: given a sample $Z_1, \ldots, Z_n$, if we choose some $\widehat{f}_n \in \mathcal{F}$ based on this sample, can we guarantee that it generalizes to unseen data? In particular, can we guarantee that (with high probability) we have the empirical risk bound

$$\widehat{R}_n(\widehat{f}_n) = \frac{1}{n} \sum_{i=1}^n \ell(\widehat{f}_n, Z_i) \leq R(\widehat{f}_n) + \epsilon \tag{4.0.3}$$

for some small ϵ? If we allow $\widehat{f}_n$ to be arbitrary, then this becomes clearly impossible: consider the classification example 4.1, and set $\widehat{f}_n$ to be the "hash" function that sets $\widehat{f}_n(x) = y$ if the pair (x, y) was in the sample, and otherwise $\widehat{f}_n(x) = -1$. Then clearly $\widehat{R}_n(\widehat{f}_n) = 0$, while there is no useful bound on $R(\widehat{f}_n)$.

4.1 Finite and countable classes of functions

In order to get bounds of the form (4.0.3), we require a few assumptions that are not too onerous. First, throughout this section, we will assume that for any fixed function f, the loss $\ell(f, Z)$ is σ^2-sub-Gaussian, that is,

$$\mathbb{E}_P\left[\exp\left(\lambda(\ell(f, Z) - R(f))\right)\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \tag{4.1.1}$$

for all $f \in \mathcal{F}$. (Recall that the risk functional $R(f) = \mathbb{E}_P[\ell(f, Z)]$.) For example, if the loss is the zero-one loss from classification problems, inequality (4.1.1) is satisfied with $\sigma^2 = \frac{1}{4}$ by Hoeffding's lemma. In order to guarantee a bound of the form (4.1.1) for a function $\widehat{f}$ chosen dependent on the data, in this section we give uniform bounds, that is, we would like to bound

$$\mathbb{P}\left(\text{there exists } f \in \mathcal{F} \text{ s.t. } R(f) > \widehat{R}_n(f) + t\right) \quad \text{or} \quad \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left|\widehat{R}_n(f) - R(f)\right| > t\right).$$

Such uniform bounds are certainly sufficient to guarantee that the empirical risk is a good proxy for the true risk R, even when $\widehat{f}_n$ is chosen based on the data.

Now, recalling that our set of functions or predictors $\mathcal{F}$ is finite or countable, let us suppose that for each $f \in \mathcal{F}$, we have a complexity measure $c(f)$—a penalty—such that

$$\sum_{f \in \mathcal{F}} e^{-c(f)} \le 1. \tag{4.1.2}$$

This inequality should look familiar to the Kraft inequality—which we will see in the coming chapters—from coding theory. As soon as we have such a penalty function, however, we have the following result.

Theorem 4.4. *Let the loss ℓ, distribution P on $\mathcal{Z}$, and function class $\mathcal{F}$ be such that $\ell(f, Z)$ is σ^2-sub-Gaussian for each $f \in \mathcal{F}$, and assume that the complexity inequality* (4.1.2) *holds. Then with probability at least $1 - \delta$ over the sample $Z_{1:n}$,*

$$R(f) \le \widehat{R}_n(f) + \sqrt{2\sigma^2 \frac{\log \frac{1}{\delta} + c(f)}{n}} \quad \textit{for all } f \in \mathcal{F}.$$

Proof First, we note that by the usual sub-Gaussian concentration inequality (Corollary 3.9) we have for any $t \ge 0$ and any $f \in \mathcal{F}$ that

$$\mathbb{P}\left(R(f) \ge \widehat{R}_n(f) + t\right) \le \exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

Now, if we replace t by $\sqrt{t^2 + 2\sigma^2 c(f)/n}$, we obtain

$$\mathbb{P}\left(R(f) \ge \widehat{R}_n(f) + \sqrt{t^2 + 2\sigma^2 c(f)/n}\right) \le \exp\left(-\frac{nt^2}{2\sigma^2} - c(f)\right).$$

Then using a union bound, we have

$$\begin{aligned} \mathbb{P}\left(\exists\, f \in \mathcal{F} \text{ s.t. } R(f) \ge \widehat{R}_n(f) + \sqrt{t^2 + 2\sigma^2 c(f)/n}\right) &\le \sum_{f \in \mathcal{F}} \exp\left(-\frac{nt^2}{2\sigma^2} - c(f)\right) \\ &= \exp\left(-\frac{nt^2}{2\sigma^2}\right) \underbrace{\sum_{f \in \mathcal{F}} \exp(-c(f))}_{\le 1}. \end{aligned}$$

Setting $t^2 = 2\sigma^2 \log \frac{1}{\delta}/n$ gives the result. □

As one classical example of this setting, suppose that we have a finite class of functions $\mathcal{F}$. Then we can set $c(f) = \log |\mathcal{F}|$, in which case we clearly have the summation guarantee (4.1.2), and we obtain

$$R(f) \le \widehat{R}_n(f) + \sqrt{2\sigma^2 \frac{\log \frac{1}{\delta} + \log |\mathcal{F}|}{n}} \quad \text{uniformly for } f \in \mathcal{F}$$

with probability at least $1 - \delta$. To make this even more concrete, consider the following example.

Example 4.5 (Floating point classifiers)**:** We implement a linear binary classifier using double-precision floating point values, that is, we have $f_\theta(x) = \theta^\top x$ for all $\theta \in \mathbb{R}^d$ that may be represented using d double-precision floating point numbers. Then for each coordinate of θ, there are at most 2^{64} representable numbers; in total, we must thus have $|\mathcal{F}| \le 2^{64d}$. Thus, for the zero-one loss $\ell_{\rm zo}(f_\theta, (x, y)) = \mathbf{1}\left\{\theta^\top xy \le 0\right\}$, we have

$$R(f_\theta) \le \widehat{R}_n(f_\theta) + \sqrt{\frac{\log\frac{1}{\delta} + 45d}{2n}}$$

for all representable classifiers simultaneously, with probability at least $1 - \delta$, as the zero-one loss is $1/4$-sub-Gaussian. (Here we have used that $64 \log 2 < 45$.) ♣

We also note in passing that by replacing δ with $\delta/2$ in the bounds of Theorem 4.4, a union bound yields the following two-sided corollary.

Corollary 4.6. *Under the conditions of Theorem 4.4, we have*

$$\left|\widehat{R}_n(f) - R(f)\right| \le \sqrt{2\sigma^2 \frac{\log\frac{2}{\delta} + c(f)}{n}} \quad \text{for all } f \in \mathcal{F}$$

with probability at least $1 - \delta$.

4.2 Structural risk minimization and adaptivity

In general, for a given function class $\mathcal{F}$, we can always decompose the excess risk into the *approximation/estimation* error decomposition. That is, let

$$R^* = \inf_f R(f),$$

where the preceding infimum is taken across *all* (measurable) functions. Then we have

$$R(\widehat{f}_n) - R^* = \underbrace{R(\widehat{f}_n) - \inf_{f\in\mathcal{F}} R(f)}_{\text{estimation}} + \underbrace{\inf_{f\in\mathcal{F}} R(f) - R^*}_{\text{approximation}}. \tag{4.2.1}$$

There is often a tradeoff between these two, analogous to the bias/variance tradeoff in classical statistics; if the approximation error is very small, then it is likely hard to guarantee that the estimation error converges quickly to zero, while certainly a constant function will have low estimation error, but may have substantial approximation error. With that in mind, we would like to develop procedures that, rather than simply attaining good performance for the class $\mathcal{F}$, are guaranteed to trade-off in an appropriate way between the two types of error. This leads us to the idea of *structural risk minimization*.

In this scenario, we assume we have a sequence of classes of functions, $\mathcal{F}_1, \mathcal{F}_2, \ldots$, of increasing complexity, meaning that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$. For example, in a linear classification setting with vectors $x \in \mathbb{R}^d$, we might take a sequence of classes allowing increasing numbers of non-zeros in the classification vector θ:

$$\mathcal{F}_1 := \left\{f_\theta(x) = \theta^\top x \text{ such that } \|\theta\|_0 \le 1\right\}, \ \mathcal{F}_2 := \left\{f_\theta(x) = \theta^\top x \text{ such that } \|\theta\|_0 \le 2\right\}, \ldots.$$

More broadly, let $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$ be a (possibly infinite) increasing sequence of function classes. We assume that for each $\mathcal{F}_k$ and each $n \in \mathbb{N}$, there exists a constant $C_{n,k}(\delta)$ such that we have the uniform generalization guarantee

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_k}\left|\widehat{R}_n(f) - R(f)\right| \ge C_{n,k}(\delta)\right) \le \delta \cdot 2^{-k}.$$

For example, by Corollary 4.6, if $\mathcal{F}$ is finite we may take

$$C_{n,k}(\delta) = \sqrt{2\sigma^2 \frac{\log|\mathcal{F}_k| + \log\frac{1}{\delta} + k\log 2}{n}}.$$

(We will see in subsequent sections of the course how to obtain other more general guarantees.)

We consider the following *structural risk minimization* procedure. First, given the empirical risk $\widehat{R}_n$, we find the model collection $\widehat{k}$ minimizing the penalized risk

$$\widehat{k} := \mathop{\mathrm{argmin}}_{k\in\mathbb{N}}\left\{\inf_{f\in\mathcal{F}_k}\widehat{R}_n(f) + C_{n,k}(\delta)\right\}. \tag{4.2.2a}$$

We then choose $\widehat{f}$ to minimize the risk over the estimated "best" class $\mathcal{F}_{\widehat{k}}$, that is, set

$$\widehat{f} := \mathop{\mathrm{argmin}}_{f\in\mathcal{F}_{\widehat{k}}}\widehat{R}_n(f). \tag{4.2.2b}$$

With this procedure, we have the following theorem.

Theorem 4.7. *Let $\widehat{f}$ be chosen according to the procedure* (4.2.2a)–(4.2.2b)*. Then with probability at least $1-\delta$, we have*

$$R(\widehat{f}) \le \inf_{k\in\mathbb{N}}\inf_{f\in\mathcal{F}_k}\left\{R(f) + 2C_{n,k}(\delta)\right\}.$$

Proof First, we have by the assumed guarantee on $C_{n,k}(\delta)$ that

$$\begin{aligned}&\mathbb{P}\left(\exists\ k\in\mathbb{N} \text{ and } f\in\mathcal{F}_k \text{ such that } \sup_{f\in\mathcal{F}_k}\left|\widehat{R}_n(f) - R(f)\right| \ge C_{n,k}(\delta)\right)\\ &\qquad\le \sum_{k=1}^{\infty}\mathbb{P}\left(\exists\ f\in\mathcal{F}_k \text{ such that } \sup_{f\in\mathcal{F}_k}\left|\widehat{R}_n(f) - R(f)\right| \ge C_{n,k}(\delta)\right) \le \sum_{k=1}^{\infty}\delta\cdot 2^{-k} = \delta.\end{aligned}$$

On the event that $\sup_{f\in\mathcal{F}_k}|\widehat{R}_n(f) - R(f)| < C_{n,k}(\delta)$ for all k, which occurs with probability at least $1-\delta$, we have

$$R(\widehat{f}) \le \widehat{R}_n(f) + C_{n,\widehat{k}}(\delta) = \inf_{k\in\mathbb{N}}\inf_{f\in\mathcal{F}_k}\left\{\widehat{R}_n(f) + C_{n,k}(\delta)\right\} \le \inf_{k\in\mathbb{N}}\inf_{f\in\mathcal{F}_k}\left\{R(f) + 2C_{n,k}(\delta)\right\}$$

by our choice of $\widehat{f}$. This is the desired result. □

We conclude with a final example, using our earlier floating point bound from Example 4.5, coupled with Corollary 4.6 and Theorem 4.7.

Example 4.8 (Structural risk minimization with floating point classifiers)**:** Consider again our floating point example, and let the function class $\mathcal{F}_k$ consist of functions defined by at most k double-precision floating point values, so that $\log |\mathcal{F}_k| \le 45d$. Then by taking

$$C_{n,k}(\delta) = \sqrt{\frac{\log \frac{1}{\delta} + 65k \log 2}{2n}}$$

we have that $|\widehat{R}_n(f) - R(f)| \le C_{n,k}(\delta)$ simultaneously for all $f \in \mathcal{F}_k$ and all $\mathcal{F}_k$, with probability at least $1 - \delta$. Then the empirical risk minimization procedure (4.2.2) guarantees that

$$R(\widehat{f}) \le \inf_{k \in \mathbb{N}} \left\{ \inf_{f \in \mathcal{F}_k} R(f) + \sqrt{\frac{2 \log \frac{1}{\delta} + 91k}{n}} \right\}.$$

Roughly, we trade between small risk $R(f)$—as the risk $\inf_{f \in \mathcal{F}_k} R(f)$ must be decreasing in k—and the estimation error penalty, which scales as $\sqrt{(k + \log \frac{1}{\delta})/n}$. ♣

Chapter 5

Basics of source coding

In this chapter, we explore the basic results in source coding—that is, given a sequence of random variables $X_1, X_2, \ldots$ distributed according to some known distribution P, how much storage is required for us to encode the random variables? The material in this chapter is covered in a variety of sources; standard references include Cover and Thomas [1] and Csiszár and Körner [2].

5.1 The source coding problem

The source coding problem—in its simplest form—is that of most efficiently losslessly encoding a sequence of symbols (generally random variables) drawn from a known distribution. In particular, we assume that the data consist of a sequence of symbols $X_1, X_2, \ldots$, drawn from a known distribution P on a finite or countable space $\mathcal{X}$. We wish to choose an encoding, represented by a *d-ary code function* C that maps $\mathcal{X}$ to finite strings consisting of the symbols $\{0, 1, \ldots, d-1\}$. We denote this by $\mathsf{C} : \mathcal{X} \to \{0, 1, \ldots, d-1\}^*$, and use $\ell_{\mathsf{C}}(x)$ to denote the length of the string $\mathsf{C}(x)$.

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

Definition 5.1. *A d-ary code* $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$ *is* non-singular *if for each* $x, x' \in \mathcal{X}$ *we have*

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

While Definition 5.1 is natural, generally speaking, we wish to transmit or encode a variety of codewords simultaneously, that is, we wish to encode a sequence $X_1, X_2, \ldots$ using the natural *extension* of the code C as the string $\mathsf{C}(X_1)\mathsf{C}(X_2)\mathsf{C}(X_3)\cdots$, where $\mathsf{C}(x_1)\mathsf{C}(x_2)$ denotes the concatenation of the strings $\mathsf{C}(x_1)$ and $\mathsf{C}(x_2)$. In this case, we require that the code be uniquely decodable:

Definition 5.2. *A d-ary code* $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$ *is* uniquely decodable *if for all sequences* $x_1, \ldots, x_n \in \mathcal{X}$ *and* $x'_1, \ldots, x'_n \in \mathcal{X}$ *we have*

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

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

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

Definition 5.3. *A d-ary code $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$ is* uniquely decodable *or* instantaneous *if no codeword is the prefix to another codeword.*

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

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

Example 5.1 (Encoding strategies)**:** Consider the encoding schemes below, which encode the letters a, b, c, and d.

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

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

5.2 The Kraft-McMillan inequalities

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

Theorem 5.2. *Let $\mathcal{X}$ be a finite or countable set, and let $\ell : \mathcal{X} \to \mathbb{N}$ be a function. If $\ell(x)$ is the length of the encoding of the symbol x in a uniquely decodable d-ary code, then*

$$\sum_{x \in \mathcal{X}} d^{-\ell(x)} \le 1. \tag{5.2.1}$$

Conversely, given any function $\ell : \mathcal{X} \to \mathbb{N}$ satisfying inequality (5.2.1)*, there is a prefix code whose codewords have length $\ell(x)$ for each $x \in \mathcal{X}$.*

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

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

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

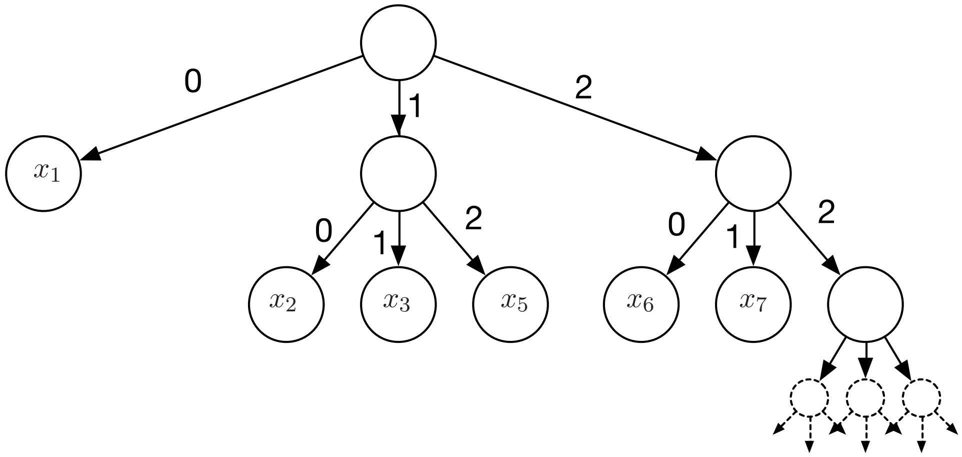

Figure 5.1. Prefix-tree encoding of a set of symbols. The encoding for x_1 is `0`, for x_2 is `10`, for x_3 is `11`, for x_4 is `12`, for x_5 is `20`, for x_6 is `21`, and nothing is encoded as `1`, `2`, or `22`.

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

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

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

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

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

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

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

We remark in passing if $\text{card}(\mathcal{X}) = \infty$, then by defining the sequence

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

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

The achievability of such a code is straightforward by a pictorial argument (recall Figure 5.1), so we sketch the result non-rigorously. Indeed, let $\mathcal{T}_d$ be an (infinite) d-ary tree. Then, at each level m of the tree, assign one of the nodes at that level to each symbol $x \in \mathcal{X}$ such that $\ell(x) = m$. Eliminate the subtree below that node, and repeat with the remaining symbols. The codeword corresponding to symbol x is then the path to the symbol in the tree. □

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

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

Theorem 5.3. *Let $X \in \mathcal{X}$ be a discrete random variable distributed according to P and let ℓ_C be the length function associated with a d-ary encoding $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$. In addition, let $\mathcal{C}$ be the set of all uniquely decodable d-ary codes for $\mathcal{X}$. Then*

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

Proof The lower bound is an argument by convex optimization, while for the upper bound we give an explicit length function and (implicit) prefix code attaining the bound. For the lower bound, we assume for simplicity that $\mathcal{X}$ is finite, and we identify $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$ (let $m = |\mathcal{X}|$ for shorthand). Then as $\mathcal{C}$ consists of *uniquely decodable* codebooks, all the associated length functions must satisfy the Kraft-McMillan inequality (5.2.1). Letting $\ell_i = \ell(i)$, the minimal encoding length is at least

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

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

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

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

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

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

as desired. □

5.3 Entropy rates and longer codes

Finally, we show that it is possible, at least for appropriate distributions on random variables X_i, to achieve a per-symbol encoding length that approaches a limiting version of the Shannon entropy of a random variable. To that end, we give two definitions capturing the limiting entropy properties of sequences of random variables.

Definition 5.4. *The* entropy rate *of a sequence $X_1, X_2, \ldots$ of random variables is*

$$H(\{X_i\}) := \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) \tag{5.3.1}$$

whenever the limit exists.

In some situations, the limit (5.3.1) may not exist. However, there are a variety of situations in which it does, and we focus generally on a specific but common instance in which the limit does exist. First, we recall the definition of a stationary sequence of random variables.

Definition 5.5. *We say a sequence $X_1, X_2, \ldots$ of random variable is stationary if for all n and all $k \in \mathbb{N}$ and all measurable sets $A_1, \ldots, A_k \subset \mathcal{X}$ we have*

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

With this definition, we have the following result.

Proposition 5.4. *Let the sequence of random variables $\{X_i\}$, taking values in the discrete space $\mathcal{X}$, be stationary. Then*

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

and the limits (5.3.1) *and above exist.*

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

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

and using that conditioning decreases entropy, we have

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

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

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

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

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

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

Proposition 5.5. *Let the sequence of random variables $X_1, X_2, \ldots$ be stationary. Then for any $\epsilon > 0$, there exists an $m \in \mathbb{N}$ and a d-ary (prefix) block encoder $\mathsf{C} : \mathcal{X}^m \to \{0, \ldots, d-1\}^*$ such that*

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

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

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

Then whenever n/m is an integer, we have

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

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

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

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

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

Bibliography

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[2] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 1981.

Chapter 6

Exponential families and maximum entropy

In this set of notes, we give a very brief introduction to exponential family models, which are a broad class of distributions that have been extensively studied in the statistics literature [4, 1, 2, 7]. There are deep connections between exponential families, convex analysis [7], and information geometry and the geometry of probability measures [1], and we will only touch briefly on a few of those here.

6.1 Review or introduction to exponential family models

We begin by defining exponential family distributions, giving several examples to illustrate a few of their properties. To define an exponential family distribution, we always assume there is some base measure μ on a space $\mathcal{X}$, and there exists a *sufficient statistic* $\phi : \mathcal{X} \to \mathbb{R}^d$, where $d \in \mathbb{N}$ is some fixed integer. For a given sufficient statistic function ϕ, let $\theta \in \mathbb{R}^d$ be an associated vector of *canonical* parameters. Then with this notation, we have the following.

Definition 6.1. *The* exponential family *associated with the function ϕ and base measure μ is defined as the set of distributions with densities p_θ with respect to μ, where*

$$p_\theta(x) = \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right), \tag{6.1.1}$$

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

$$A(\theta) := \log \int_{\mathcal{X}} \exp\left(\langle \theta, \phi(x) \rangle\right) d\mu(x), \tag{6.1.2}$$

whenever A is finite.

In some settings, it is convenient to define a base function $h : \mathcal{X} \to \mathbb{R}_+$ and define

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

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

We now give a few examples of exponential family models.

Example 6.1 (Bernoulli distribution)**:** In this case, we have $X \in \{0, 1\}$ and $P(X = 1) = p$ for some $p \in [0, 1]$ in the classical version of a Bernoulli. Thus we take μ to be the counting measure on $\{0, 1\}$, and by setting $\theta = \log \frac{p}{1-p}$ to obtain a canonical representation, we have

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

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

Example 6.2 (Poisson distribution)**:** The Poisson distribution (for count data) is usually parameterized by some $\lambda > 0$, and for $x \in \mathbb{N}$ has distribution $P_\lambda(X = x) = (1/x!)\lambda^x e^-\lambda$. Thus by taking μ to be counting (discrete) measure on $\{0, 1, \ldots\}$ and setting $\theta = \log \lambda$, we find the density (probability mass function in this case)

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

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

Example 6.3 (Normal distribution)**:** For the normal distribution, we take μ to be Lebesgue measure on $(-\infty, \infty)$. Then $\mathsf{N}(\mu, \Sigma)$ can be re-parameterized as as $\Theta = \Sigma^{-1}$ and $\theta = \Sigma^{-1}\mu$, and we have density

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

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. ♣

6.1.1 Why exponential families?

There are many reasons for us to study exponential families. As we see presently, they arise as the solutions to several natural optimization problems on the space of probability distributions. They also enjoy certain robustness properties related to optimal Bayes' procedures (more to come on this topic). Moreover, they are analytically very tractable, and have been the objects of substantial study for nearly the past hundred years. As one example, the following result is well-known (see, e.g., Wainwright and Jordan [7, Proposition 3.1] or Brown [4]):

Proposition 6.4. *The log-partition function $\theta \mapsto A(\theta)$ is infinitely differentiable on its open domain $\Theta := \{\theta \in \mathbb{R}^d : A(\theta) < \infty\}$. Moreover, A is convex.*

Proof We show convexity; the proof of the infinite differentiability follows from an argument using the dominated convergence theorem that allows passing the derivative through the integral defining A. For convexity, let let $\theta_\lambda = \lambda\theta_1 + (1 - \lambda)\theta_2$, where $\theta_1, \theta_2 \in \Theta$. Then $1/\lambda \geq 1$ and $1/(1 - \lambda) \geq 1$, and Hölder's inequality implies

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

as desired. □

As a final remark, we note that this convexity makes estimation in exponential families substantially easier. Indeed, given a sample $X_1, \ldots, X_n$, assume that we estimate θ by maximizing the likelihood (equivalently, minimizing the log-loss):

$$\underset{\theta}{\text{minimize}} \ \sum_{i=1}^n \log \frac{1}{p_\theta(X_i)} = \sum_{i=1}^n \left[- \langle \theta, \phi(X_i) \rangle + A(\theta) \right],$$

which is thus convex in θ. This means there are no local minima, and tractable algorithms exist for solving maximum likelihood. Later we will explore some properties of these types of minimization and log-loss problems.

6.2 Shannon entropy

We now explore a generalized version of entropy known as Shannon entropy, which allows us to define an entropy functional for essentially arbitrary distributions. This comes with a caveat, however: to define this entropy, we must fix a base measure μ ahead of time against which we integrate. In this case, we have

Definition 6.2. *Let μ be a base measure on $\mathcal{X}$ and assume P has density p with respect to μ. Then the* Shannon entropy *of P is*

$$H(P) = -\int p(x) \log p(x) d\mu(x).$$

Notably, if $\mathcal{X}$ is a discrete set and μ is counting measure, then $H(P) = -\sum_x p(x) \log p(x)$ is simply the standard entropy. However, for other base measures the calculation is different. For example, if we take μ to be Lebesgue measure, meaning that $d\mu(x) = dx$ and giving rise to the usual integral on $\mathbb{R}$ (or $\mathbb{R}^d$), then we obtain *differential entropy* [5, Chapter 8].

Example 6.5: Let P be the uniform distribution on $[0, a]$. Then the differential entropy $H(P) = -\log(1/a) = \log a$. ♣

Example 6.6: Let P be the normal distribution $\mathsf{N}(\mu, \Sigma)$ and μ be Lebesgue measure. Then

$$\begin{aligned} H(P) &= -\int p(x) \left[\log \frac{1}{\sqrt{2\pi \det(\Sigma)}} - \frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x - \mu) \right] dx \\ &= \frac{1}{2} \log(2\pi \det(\Sigma)) + \frac{1}{2} \mathbb{E}[(X-\mu)^\top \Sigma^{-1} (X-\mu)] \\ &= \frac{1}{2} \log(2\pi \det(\Sigma)) + \frac{d}{2}. \end{aligned}$$

♣

6.3 Maximizing Entropy

The maximum entropy principle, proposed by Jaynes in the 1950s (see Jaynes [6]), originated in statistical mechanics, where Jaynes showed that (in a sense) entropy in statistical mechanics and information theory were equivalent. The maximum entropy principle is this: given some constraints (prior information) about a distribution P, we consider all probability distributions satisfying said constraints. Then to encode our prior information while being as “objective” or “agnostic” as possible (essentially being as uncertain as possible), we should choose the distribution P satisfying the constraints to maximize the Shannon entropy.

While there are many arguments for and against the maximum entropy principle, we shall not dwell on them here, instead showing how maximizing entropy naturally gives rise to exponential family models. We will later see connections to Bayesian and minimax procedures. The one thing that we must consider, about which we will be quite explicit, is that the base measure μ is *essential* to all our derivations: it radically effects the distributions P we consider.

6.3.1 The maximum entropy problem

We begin by considering linear (mean-value) constraints on our distributions. In this case, we are given a function $\phi : \mathcal{X} \to \mathbb{R}^d$ and vector $\alpha \in \mathbb{R}^d$, we wish to solve

$$\operatorname{maximize} H(P) \quad \text{subject to} \quad \mathbb{E}_P[\phi(X)] = \alpha \tag{6.3.1}$$

over all distributions P having densities with respect to the base measure μ, that is, we have the (equivalent) absolute continuity condition $P \ll \mu$. Rewriting problem (6.3.1), we see that it is equivalent to

$$\begin{aligned} &\text{maximize} && -\int p(x) \log p(x) d\mu(x) \\ &\text{subject to} && \int p(x)\phi_i(x) d\mu(x) = \alpha_i, \quad p(x) \geq 0 \text{ for } x \in \mathcal{X}, \quad \int p(x) d\mu(x) = 1. \end{aligned}$$

Let

$$\mathcal{P}_\alpha^{\text{lin}} := \{P \ll \mu : \mathbb{E}_P[\phi(X)] = \alpha\}$$

be distributions with densities w.r.t. μ satisfying the expectation (linear) constraint $\mathbb{E}[\phi(X)] = \alpha$. We then obtain the following theorem.

Theorem 6.7. *For $\theta \in \mathbb{R}^d$, let P_θ have density*

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)), \quad A(\theta) = \log \int \exp(\langle \theta, \phi(x) \rangle) d\mu(x),$$

with respect to the measure μ. If $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$, then P_θ maximizes $H(P)$ over $\mathcal{P}_\alpha^{\text{lin}}$; moreover, the distribution P_θ is unique.

Proof We first give a heuristic derivation—which is not completely rigorous—and then check to verify that our result is exact. First, we write a Lagrangian for the problem (6.3.1). Introducing Lagrange multipliers $\lambda(x) \geq 0$ for the constraint $p(x) \geq 0$, $\theta_0 \in \mathbb{R}$ for the normalization constraint that $P(\mathcal{X}) = 1$, and θ_i for the constraints that $\mathbb{E}_P[\phi_i(X)] = \alpha_i$, we obtain the following Lagrangian:

$$\begin{aligned}\mathcal{L}(p, \theta, \theta_0, \lambda) &= \int p(x) \log p(x) d\mu(x) + \sum_{i=1}^d \theta_i \left(\alpha_i - \int p(x)\phi_i(x) d\mu(x)\right) \\ &\quad + \theta_0 \left(\int p(x) d\mu(x) - 1\right) - \int \lambda(x) p(x) d\mu(x).\end{aligned}$$

Now, heuristically treating the density $p = [p(x)]_{x \in \mathcal{X}}$ as a finite-dimensional vector (in the case that $\mathcal{X}$ is finite, this is completely rigorous), we take derivatives and obtain

$$\frac{\partial}{\partial p(x)} \mathcal{L}(p, \theta, \theta_0, \lambda) = 1 + \log p(x) - \sum_{i=1}^d \theta_i \phi_i(x) + \theta_0 - \lambda(x) = 1 + \log p(x) - \langle \theta, \phi(x) \rangle + \theta_0 - \lambda(x).$$

To find the minimizing p for the Lagrangian (the function is convex in p), we set this equal to zero to find that

$$p(x) = \exp\left(\langle \theta, \phi(x) \rangle - 1 - \theta_0 - \lambda(x)\right).$$

Now, we note that with this setting, we always have $p(x) > 0$, so that the constraint $p(x) \ge 0$ is unnecessary and (by complementary slackness) we have $\lambda(x) = 0$. In particular, by taking $\theta_0 = -1 + A(\theta) = -1 + \log \int \exp(\langle \theta, \phi(x) \rangle) d\mu(x)$, we have that (according to our heuristic derivation) the optimal density p should have the form

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

So we see the form of distribution we would like to have.

Let us now consider any other distribution $P \in \mathcal{P}_\alpha^{\rm lin}$, and assume that we have some θ satisfying $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$. In this case, we may expand the entropy $H(P)$ as

$$\begin{aligned}H(P) &= -\int p \log p d\mu = -\int p \log \frac{p}{p_\theta} d\mu - \int p \log p_\theta d\mu \\ &= -D_{\rm kl}(P \| P_\theta) - \int p(x)[\langle \theta, \phi(x) \rangle - A(\theta)] d\mu(x) \\ &\overset{(\star)}{=} -D_{\rm kl}(P \| P_\theta) - \int p_\theta(x)[\langle \theta, \phi(x) \rangle - A(\theta)] d\mu(x) \\ &= -D_{\rm kl}(P \| P_\theta) - H(P_\theta),\end{aligned}$$

where in the step $(\star)$ we have used the fact that $\int p(x)\phi(x) d\mu(x) = \int p_\theta(x) \phi(x) d\mu(x) = \alpha$. As $D_{\rm kl}(P \| P_\theta) > 0$ unless $P = P_\theta$, we have shown that P_θ is the unique distribution maximizing the entropy, as desired. □

6.3.2 Examples of maximum entropy

We now give three examples of maximum entropy, showing how the choice of the base measure μ strongly effects the resulting maximum entropy distribution. For all three, we assume that the space $\mathcal{X} = \mathbb{R}$ is the real line. We consider maximizing the entropy over all distributions P satisfying

$$\mathbb{E}_P[X^2] = 1.$$

Example 6.8: Assume that the base measure μ is counting measure on the support $\{-1, 1\}$, so that $\mu(\{-1\}) = \mu(\{1\}) = 1$. Then the maximum entropy distribution is given by $P(X = x) = \frac{1}{2}$ for $x \in \{-1, 1\}$. ♣

Example 6.9: Assume that the base measure μ is Lebesgue measure on $\mathcal{X} = \mathbb{R}$, so that $\mu([a, b]) = b - a$ for $b \ge a$. Then by Theorem 6.7, we have that the maximum entropy distribution has the form $p_\theta(x) \propto \exp(-\theta x^2)$; recognizing the normal, we see that the optimal distribution is simply $\mathsf{N}(0, 1)$. ♣

Example 6.10: Assume that the base measure μ is counting measure on the integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, \ldots\}$. Then Theorem 6.7 shows that the optimal distribution is a discrete version of the normal: we have $p_\theta(x) \propto \exp(-\theta x^2)$ for $x \in \mathbb{Z}$. That is, we choose $\theta > 0$ so that the distribution $p_\theta(x) = \exp(-\theta x^2) / \sum_{j=-\infty}^{\infty} \exp(-\theta j^2)$ has variance 1. ♣

6.3.3 Generalization to inequality constraints

It is possible to generalize Theorem 6.7 in a variety of ways. In this section, we show how to generalize the theorem to general (finite-dimensional) convex cone constraints (cf. Boyd and Vandenberghe [3, Chapter 5]). To remind the reader, we say a set $\mathcal{C}$ is a *convex cone* if for any two points $x, y \in \mathcal{C}$, we have $\lambda x + (1 - \lambda) y \in \mathcal{C}$ for all $\lambda \in [0, 1]$, and $\mathcal{C}$ is closed under positive scaling: $x \in \mathcal{C}$ implies that $tx \in \mathcal{C}$ for all $t \ge 0$. While this level of generality may seem a bit extreme, it does give some nice results. In most cases, we will always use one of the following two standard examples of cones (the positive orthant and the semi-definite cone):

i. *The orthant.* Take $\mathcal{C} = \mathbb{R}^d_+ = \{x \in \mathbb{R}^d : x_j \ge 0, j = 1, \ldots, d\}$. Then clearly $\mathcal{C}$ is convex and closed under positive scaling.

ii. *The semidefinite cone.* Take $\mathcal{C} = \{X \in \mathbb{R}^{d \times d} : X = X^\top, X \succeq 0\}$, where a matrix $X \succeq 0$ means that $a^\top X a \ge 0$ for all vectors a. Then we have that $\mathcal{C}$ is convex and closed under positive scaling as well.

Given a convex cone $\mathcal{C}$, we associate a cone ordering $\succeq$ with the cone and say that for two elements $x, y \in \mathcal{C}$, we have $x \succeq y$ if $x - y \succeq 0$, that is, $x - y \in \mathcal{C}$. In the orthant case, this simply means that x is component-wise larger than y. For a given inner product $\langle \cdot, \cdot \rangle$, we define the dual cone

$$\mathcal{C}^* := \{y : \langle y, x \rangle \ge 0 \text{ for all } x \in \mathcal{C}\}.$$

For the standard (Euclidean) inner product, the positive orthant is thus self-dual, and similarly the semidefinite cone is also self-dual. For a vector y, we write $y \succeq_* 0$ if $y \in \mathcal{C}^*$ is in the dual cone.

With this generality in mind, we may consider the following linearly constrained maximum entropy problem, which is predicated on a particular cone $\mathcal{C}$ with associated cone ordering $\preceq$ and a function ψ mapping into the ambient space in which $\mathcal{C}$ lies:

$$\text{maximize } H(P) \quad \text{subject to} \quad \mathbb{E}_P[\phi(X)] = \alpha, \;\; \mathbb{E}_P[\psi(X)] \preceq \beta, \tag{6.3.2}$$

where the base measure μ is implicit. We denote the family of distributions (with densities w.r.t.

μ) satisfying the two above constraints by $\mathcal{P}^{\rm lin}_{\alpha,\beta}$. Equivalently, we wish to solve

$$\begin{aligned}
&\text{maximize} \;\; -\int p(x)\log p(x)d\mu(x) \\
&\text{subject to} \;\; \int p(x)\phi(x)d\mu(x) = \alpha, \;\; \int p(x)\psi(x)d\mu(x) \preceq \beta, \\
&\qquad\qquad p(x) \ge 0 \text{ for } x \in \mathcal{X}, \;\; \int p(x)d\mu(x) = 1.
\end{aligned}$$

We then obtain the following theorem:

Theorem 6.11. *For $\theta \in \mathbb{R}^d$ and $K \in \mathcal{C}^*$, the dual cone to $\mathcal{C}$, let $P_{\theta,K}$ have density*

$$p_{\theta,K}(x) = \exp\left(\langle \theta, \phi(x)\rangle - \langle K, \psi(x)\rangle - A(\theta, K)\right), \quad A(\theta,K) = \log \int \exp(\langle \theta, \phi(x)\rangle - \langle K, \psi(x)\rangle) d\mu(x),$$

with respect to the measure μ. If

$$\mathbb{E}_{P_{\theta,K}}[\phi(X)] = \alpha \quad \textit{and} \quad \mathbb{E}_{P_{\theta,K}}[\psi(X)] = \beta,$$

then $P_{\theta,K}$ maximizes $H(P)$ over $\mathcal{P}^{\rm lin}_{\alpha,\beta}$. Moreover, the distribution $P_{\theta,K}$ is unique.

We make a few remarks in passing before proving the theorem. First, we note that we must assume both equalities are attained for the theorem to hold. We may also present an example.

> **Example 6.12** (Normal distributions maximize entropy subject to covariance constraints)**:** Suppose that the cone $\mathcal{C}$ is the positive semidefinite cone in $\mathbb{R}^{d \times d}$, that $\alpha = 0$, that we use the Lebesgue measure as our base measure, and that $\psi(x) = xx^\top \in \mathbb{R}^{d\times d}$. Let us fix $\beta = \Sigma$ for some positive definite matrix Σ. This gives us the problem
>
> $$\text{maximize} \;\; -\int p(x)\log p(x)dx \;\; \text{subject to } \mathbb{E}_P[XX^\top] \preceq \Sigma$$
>
> Then we have by Theorem 6.11 that if we can find a density $p_K(x) \propto \exp(-\left\langle K, xx^\top \right\rangle) = \exp(-x^\top K x)$ satisfying $\mathbb{E}[XX^\top] = \Sigma$, this distribution maximizes the entropy. But this is not hard: simply take the normal distribution $\mathsf{N}(0, \Sigma)$, which gives $K = \frac{1}{2}\Sigma^{-1}$. ♣

Now we provide the proof of Theorem 6.11.

Proof We can provide a heuristic derivation of the form of $p_{\theta,K}$ identically as in the proof of Theorem 6.7, where we also introduce the dual variable $K \in \mathcal{C}^*$ for the constraint $\int p(x)\psi(x)d\mu(x) \preceq \beta$. Rather than going through this, however, we simply show that the distribution $P_{\theta,K}$ maximizes $H(P)$. Indeed, we have for any $P \in \mathcal{P}^{\rm lin}_{\alpha,\beta}$ that

$$\begin{aligned}
H(P) = -\int p(x)\log p(x)d\mu(x) &= -\int p(x)\log\frac{p(x)}{p_{\theta,K}(x)}d\mu(x) - \int p(x)\log p_{\theta,K}(x)d\mu(x) \\
&= -D_{\rm kl}\left(P\|P_{\theta,K}\right) - \int p(x)\left[\langle\theta,\phi(x)\rangle - \langle K, \psi(x)\rangle - A(\theta,K)\right]d\mu(x) \\
&\le -D_{\rm kl}\left(P\|P_{\theta,K}\right) - \left[\langle\theta,\alpha\rangle - \langle K,\beta\rangle - A(\theta,K)\right],
\end{aligned}$$

where the inequality follows because $K \succeq_* 0$ so that if $\mathbb{E}[\psi(X)] \preceq \beta$, we have

$$\langle K, \mathbb{E}[\psi(X) - \beta]\rangle \le \langle K, 0\rangle = 0 \quad \text{or} \quad \langle K, \mathbb{E}[\psi(X)]\rangle \le \langle K, \beta\rangle.$$

Now, we note that $\int p_{\theta,K}(x)\phi(x)d\mu(x) = \alpha$ and $\int p_{\theta,K}(x)\psi(x)d\mu(x) = \beta$ by assumption. Then we have

$$\begin{aligned}
H(P) &\le -D_{\rm kl}\left(P\|P_{\theta,K}\right) - \left[\langle\theta,\alpha\rangle - \langle K,\beta\rangle - A(\theta,K)\right] \\
&= -D_{\rm kl}\left(P\|P_{\theta,K}\right) - \int p_{\theta,K}(x)\left[\langle\theta,\phi(x)\rangle - \langle K,\psi(x)\rangle - A(\theta,K)\right]d\mu(x) \\
&= -D_{\rm kl}\left(P\|P_{\theta,K}\right) - \int p_{\theta,K}(x)\log p_{\theta,K}(x)d\mu(x) = -D_{\rm kl}\left(P\|P_{\theta,K}\right) + H(P_{\theta,K}).
\end{aligned}$$

As $D_{\rm kl}\left(P\|P_{\theta,K}\right) > 0$ unless $P = P_{\theta,K}$, this gives the result. □

Bibliography

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[2] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, 1978.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] L. D. Brown. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, 1986.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[6] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9): 939–952, Sept. 1982.

[7] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Chapter 7

Robustness, duality, maximum entropy, and exponential families

In this lecture, we continue our study of exponential families, but now we investigate their properties in somewhat more depth, showing how exponential family models provide a natural robustness against model mis-specification, enjoy natural projection properties, and arise in other settings.

7.1 The existence of maximum entropy distributions

As in the previous chapter of these notes, we again consider exponential family models. For simplicity throughout this chapter, and with essentially no loss of generality, we assume that all of our exponential family distributions have (standard) densities. Moreover, we assume there is some fixed density (or, more generally, an arbitrary function) p satisfying $p(x) \geq 0$ and for which

$$p_\theta(x) = p(x)\exp(\langle \theta, \phi(x)\rangle - A(\theta)), \tag{7.1.1}$$

where the log-partition function or cumulant generating function $A(\theta) = \log \int p(x)\exp(\langle \theta, \phi(x)\rangle)dx$ as usual, and ϕ is the usual vector of sufficient statistics. In the previous chapter, we saw that if we restricted consideration to distributions satisfying the mean-value (linear) constraints of the form

$$\mathcal{P}_\alpha^{\rm lin} := \left\{Q : q(x) = p(x)f(x), \quad \text{where } f \geq 0 \text{ and } \int q(x)\phi(x)dx = \alpha, \int q(x)dx = 1\right\},$$

then the distribution with density $p_\theta(x) = p(x)\exp(\langle \theta, \phi(x)\rangle - A(\theta))$ uniquely maximized the (Shannon) entropy over the family $\mathcal{P}_\alpha^{\rm lin}$ if we could find any θ satisfying $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$. (Recall Theorem 6.7.) Now, of course, we must ask: does this actually happen? For if it does not, then all of this work is for naught.

Luckily for us, the answer is that we often find ourselves in the case that such results occur. Indeed, it is possible to show that, except for pathological cases, we are essentially *always* able to find such a solution. To that end, define the mean space

$$\mathcal{M}_\phi := \left\{\alpha \in \mathbb{R}^d : \exists Q \text{ s.t. } q(x) = f(x)p(x), f \geq 0, \quad \text{and} \quad \int q(x)\phi(x)dx = \alpha\right\}$$

Then we have the following result, which is well-known in the literature on exponential family modeling; we refer to Wainwright and Jordan [5, Proposition 3.2 and Theorem 3.3] for the proof. In the statement of the theorem, we recall that the domain dom A of the log partition function is defined as those points θ for which the integral $\int p(x)\exp(\langle \theta, \phi(x)\rangle)dx < \infty$.

Theorem 7.1. *Assume that there exists some point* $\theta_0 \in \operatorname{int}\operatorname{dom} A$*, where* $\operatorname{dom} A := \{\theta \in \mathbb{R}^d : A(\theta) < \infty\}$*. Then for any* α *in the interior of* $\mathcal{M}_\phi$*, there exists some* $\theta = \theta(\alpha)$ *such that* $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$*.*

Using tools from convex analysis, it is possible to extend this result to the case that $\operatorname{dom} A$ has no interior but only a relative interior, and similarly for $\mathcal{M}_\phi$ (see Hiriart-Urruty and Lemaréchal [3] or Rockafellar [4] for discussions of interior and relative interior). Moreover, it is also possible to show that for any $\alpha \in \mathcal{M}_\phi$ (not necessarily the interior), there exists a sequence $\theta_1, \theta_2, \ldots$ satisfying the limiting guarantee $\lim_n \mathbb{E}_{P_{\theta_n}}[\phi(X)] = \alpha$. Regardless, we have our desired result: if $\mathcal{P}^{\rm lin}$ is not empty, maximum entropy distributions exist and exponential family models attain these maximum entropy solutions.

7.2 I-projections and maximum likelihood

We first show one variant of the robustness of exponential family distributions by showing that they are (roughly) projections onto constrained families of distributions, and that they arise naturally in the context of maximum likelihood estimation. First, suppose that we have a family Π of distributions and some fixed distribution P (this last assumption of a fixed distribution P is not completely essential, but it simplifies our derivation). Then the *I-Projection* (for information projection) of the distribution P onto the family Π is

$$P^* := \operatorname*{argmin}_{Q \in \Pi} D_{\rm kl}\left(Q \| P\right), \tag{7.2.1}$$

when such a distribution exists. (In nice cases, it does.)

Perhaps unsurprisingly, given our derivations with maximum entropy distributions and exponential family models, we have the next proposition. The proposition shows that I-Projection is essentially the same as maximum entropy, and the projection of a distribution P onto a family of linearly constrained distributions yields exponential family distributions.

Proposition 7.2. *Suppose that* $\Pi = \mathcal{P}^{\rm lin}_\alpha$*. If* $p_\theta(x) = p(x)\exp(\langle \theta, \phi(x)\rangle - A(\theta))$ *satisfies* $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$*, then* p_θ *solves the I-projection problem* (7.2.1)*. Moreover we have (the Pythagorean identity)*

$$D_{\rm kl}\left(Q \| P\right) = D_{\rm kl}\left(P_\theta \| P\right) + D_{\rm kl}\left(Q \| P_\theta\right)$$

for $Q \in \mathcal{P}^{\rm lin}_\alpha$*.*

Proof Our proof is to perform an expansion of the KL-divergence that is completely parallel to that we performed in the proof of Theorem 6.7. Indeed, we have

$$\begin{aligned}
D_{\rm kl}\left(Q \| P\right) &= \int q(x) \log \frac{q(x)}{p(x)} dx \\
&= \int q(x) \log \frac{p_\theta(x)}{p(x)} dx + \int q(x) \log \frac{q(x)}{p_\theta(x)} dx \\
&= \int q(x)[\langle \theta, \phi(x)\rangle - A(\theta)] dx + D_{\rm kl}\left(Q \| P_\theta\right) \\
&\overset{(\star)}{=} \int p_\theta(x)[\langle \theta, \phi(x)\rangle - A(\theta)] dx + D_{\rm kl}\left(Q \| P_\theta\right) \\
&= \int p_\theta(x) \log \frac{p_\theta(x)}{p(x)} + D_{\rm kl}\left(Q \| P_\theta\right),
\end{aligned}$$

where equality ($\star$) follows by assumption that $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$. □

Now we consider maximum likelihood estimation, showing that—in a completely handwavy fashion—approximates I-projection. First, suppose that we have an exponential family $\{P_\theta\}_{\theta \in \Theta}$ of distributions, and suppose that the data comes from a true distribution P. Then to maximizing the likelihood of the data is equivalent to maximizing the log likelihood, which, in the population case, gives us the following sequence of equivalences:

$$\begin{aligned} \text{maximize } \mathbb{E}_P[\log p_\theta(X)] &\equiv \text{minimize } \mathbb{E}_P[\log \frac{1}{p_\theta(X)}] \\ &\equiv \text{minimize } \mathbb{E}_P\Big[\log \frac{p(X)}{p_\theta(X)}\Big] + H(P) \\ &\equiv \underset{\theta}{\text{minimize}}\ D_{\text{kl}}\left(P \| P_\theta\right), \end{aligned}$$

so that maximum likelihood is essentially a different type of projection.

Now, we also consider the empirical variant of maximum likelihood, where we maximize the likelihood of a given sample $X_1, \ldots, X_n$. In particular, we may study the structure of maximum likelihood exponential family estimators, and we see that they correspond to simple moment matching in exponential families. Indeed, consider the sample-based maximum likelihood problem of solving

$$\underset{\theta}{\text{maximize}}\ \prod_{i=1}^n p_\theta(X_i) \equiv \text{maximize} \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i), \tag{7.2.2}$$

where as usual we assume the exponential family model $p_\theta(x) = p(x)\exp(\langle \theta, \phi(x)\rangle - A(\theta))$. We have the following result.

Proposition 7.3. *Let $\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^n \phi(X_i)$. Then the maximum likelihood solution is given by any θ such that $\mathbb{E}_{P_\theta}[\phi(X)] = \widehat{\alpha}$.*

Proof The proof follows immediately upon taking derivatives. We define the empirical negative log likelihood (the empirical risk) as

$$\widehat{R}_n(\theta) := -\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) = -\frac{1}{n}\sum_{i=1}^n \langle \theta, \phi(X_i)\rangle + A(\theta) - \frac{1}{n}\sum_{i=1}^n \log p(X_i),$$

which is convex as $\theta \mapsto A(\theta)$ is convex (recall Proposition 6.4). Taking derivatives, we have

$$\begin{aligned} \nabla_\theta \widehat{R}_n(\theta) &= -\frac{1}{n}\sum_{i=1}^n \phi(X_i) + \nabla A(\theta) \\ &= -\frac{1}{n}\sum_{i=1}^n \phi(X_i) + \frac{1}{\int p(x)\exp(\langle \theta, \phi(x)\rangle)dx}\int \phi(x)p(x)\exp(\langle \theta, \phi(x)\rangle)dx \\ &= -\frac{1}{n}\sum_{i=1}^n \phi(X_i) + \mathbb{E}_{P_\theta}[\phi(X)]. \end{aligned}$$

In particular, finding any θ such that $\nabla A(\theta) = \mathbb{E}_{\widehat{P}_n}[\phi(X)]$ gives the result. □

As a consequence of the result, we have the following rough equivalences tying together the preceding material. In short, maximum entropy subject to (linear) empirical moment constraints (Theorem 6.7) is equivalent to maximum likelihood estimation in exponential families (Proposition 7.3), which is equivalent to I-projection of a fixed base distribution onto a linearly constrained family of distributions (Proposition 7.2).

7.3 Basics of minimax game playing with log loss

The final set of problems we consider in which exponential families make a natural appearance are in so-called minimax games under the log loss. In particular, we consider the following general formulation of a two-player minimax game. First, we choose a distribution Q on a set $\mathcal{X}$ (with density q). Then nature (or our adversary) chooses a distribution $P \in \mathcal{P}$ on the set $\mathcal{X}$, where $\mathcal{P}$ is a collection of distributions on $\mathcal{X}$, so we suffer loss

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[-\log q(X)] = \sup_{P \in \mathcal{P}} \int p(x) \log \frac{1}{q(x)} dx. \tag{7.3.1}$$

In particular, we would like to solve the minimax problem

$$\underset{Q}{\text{minimize}} \sup_{P \in \mathcal{P}} \mathbb{E}[-\log q(X)].$$

To motivate this abstract setting we give two examples, the first abstract and the second somewhat more concrete.

Example 7.4: Suppose that receive n random variables $X_i \overset{\text{i.i.d.}}{\sim} P$; in this case, we have the sequential prediction loss

$$\mathbb{E}_P[-\log q(X_1^n)] = \sum_{i=1}^n \mathbb{E}_P\left[\log \frac{1}{q(X_i \mid X_1^{i-1})}\right],$$

which corresponds to predicting X_i given X_1^{i-1} as well as possible, when the X_i follow an (unknown or adversarially chosen) distribution P. ♣

Example 7.5 (Coding)**:** Expanding on the preceding example, suppose that the set $\mathcal{X}$ is finite, and we wish to encode $\mathcal{X}$ into $\{0, 1\}$-valued sequences using as few bits as possible. In this case, the Kraft inequality (recall Theorem 5.2) tells us that if $C : \mathcal{X} \to \{0, 1\}^*$ is an uniquely decodable code, and $\ell_C(x)$ denotes the length of the encoding for the symbol $x \in \mathcal{X}$, then

$$\sum_x 2^{-\ell_C(x)} \le 1.$$

Conversely, given any length function $\ell : \mathcal{X} \to \mathbb{N}$ satisfying $\sum_x 2^{-\ell(x)} \le 1$, there exists an instantaneous (prefix) code C with the given length function. Thus, if we define the p.m.f. $q_C(x) = 2^{-\ell_C(x)} / \sum_x 2^{-\ell_C(x)}$, we have

$$-\log_2 q_C(x_1^n) = \sum_{i=1}^n \left[\ell_C(x_i) + \log \sum_x 2^{-\ell_C(x)}\right] \le \sum_{i=1}^n \ell_C(x_i).$$

In particular, we have a coding game where we attempt to choose a distribution Q (or sequential coding scheme C) that has as small an expected length as possible, uniformly over distributions P. (The field of universal coding studies such questions in depth; see Tsachy Weissman's course EE376b.) ♣

We now show how the minimax game (7.3.1) naturally gives rise to exponential family models, so that exponential family distributions are so-called robust Bayes procedures (cf. Grünwald and Dawid [2]). Specifically, we say that Q is a robust Bayes procedure for the class $\mathcal{P}$ of distributions if it minimizes the supremum risk (7.3.1) taken over the family $\mathcal{P}$; that is, it is uniformly good for all distributions $P \in \mathcal{P}$. If we restrict our class $\mathcal{P}$ to be a linearly constrained family of distributions, then we see that the exponential family distributions are natural robust Bayes procedures: they uniquely solve the minimax game. More concretely, assume that $\mathcal{P} = \mathcal{P}_\alpha^{\rm lin}$ and that P_θ denotes the exponential family distribution with density $p_\theta(x) = p(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta))$, where p denotes the base density. We have the following.

Proposition 7.6. *If $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$, then*

$$\inf_Q \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\log q(X)] = \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\log p_\theta(X)] = \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \inf_Q \mathbb{E}_P[-\log q(X)].$$

Proof This is a standard saddle-point argument (cf. [4, 3, 1]). First, note that

$$\begin{aligned}\sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\log p_\theta(X)] &= \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\langle \phi(X), \theta \rangle + A(\theta)] \\ &= -\langle \alpha, \theta \rangle + A(\theta) = \mathbb{E}_{P_\theta}[-\langle \theta, \phi(X) \rangle + A(\theta)] = H(P_\theta),\end{aligned}$$

where H denotes the Shannon entropy, for any distribution $P \in \mathcal{P}_\alpha^{\rm lin}$. Moreover, for any $Q \neq P_\theta$, we have

$$\sup_P \mathbb{E}_P[-\log q(X)] \ge \mathbb{E}_{P_\theta}[-\log q(X)] > \mathbb{E}_{P_\theta}[-\log p_\theta(X)] = H(P_\theta),$$

where the inequality follows because $D_{\rm kl}(P_\theta \| Q) = \int p_\theta(x) \log \frac{p_\theta(x)}{q(x)} dx > 0$. This shows the first equality in the proposition.

For the second equality, note that

$$\inf_Q \mathbb{E}_P[-\log q(X)] = \underbrace{\inf_Q \mathbb{E}_P\Big[\log \frac{p(X)}{q(X)}\Big]}_{=0} - \mathbb{E}_P[\log p(x)] = H(P).$$

But we know from our standard maximum entropy results (Theorem 6.7) that P_θ maximizes the entropy over $\mathcal{P}_\alpha^{\rm lin}$, that is, $\sup_{P \in \mathcal{P}_\alpha^{\rm lin}} H(P) = H(P_\theta)$. □

In short: maximum entropy is equivalent to robust prediction procedures for linear families of distributions $\mathcal{P}_\alpha^{\rm lin}$, which is equivalent to maximum likelihood in exponential families, which in turn is equivalent to I-projection.

Bibliography

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] P. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32(4):1367–1433, 2004.

[3] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

[4] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[5] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Chapter 8

Fisher Information

Having explored the definitions associated with exponential families and their robustness properties, we now turn to a study of somewhat more general parameterized distributions, developing connections between divergence measures and other geometric ideas such as the Fisher information. After this, we illustrate a few consequences of Fisher information for optimal estimators, which gives a small taste of the deep connections between information geometry, Fisher information, exponential family models. In the coming chapters, we show how Fisher information measures come to play a central role in sequential (universal) prediction problems.

8.1 Fisher information: definitions and examples

We begin by defining the Fisher information. Let $\{P_\theta\}_{\theta \in \Theta}$ denote a parametric family of distributions on a space $\mathcal{X}$, each where $\theta \in \Theta \subset \mathbb{R}^d$ indexes the distribution. Throughout this lecture and the next, we assume (with no real loss of generality) that each P_θ has a density given by p_θ. Then the *Fisher information* associated with the model is the matrix given by

$$I_\theta := \mathbb{E}_\theta \left[\nabla_\theta \log p_\theta(X) \nabla \log p_\theta(X)^\top \right] = \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^\top], \tag{8.1.1}$$

where the score function $\dot{\ell}_\theta = \nabla_\theta \log p_\theta(x)$ is the gradient of the log likelihood at θ (implicitly depending on X) and the expectation $\mathbb{E}_\theta$ denotes expectation taken with respect to P_θ. Intuitively, the Fisher information captures the variability of the gradient $\nabla \log p_\theta$; in a family of distributions for which the score function $\dot{\ell}_\theta$ has high variability, we intuitively expect estimation of the parameter θ to be easier—different θ change the behavior of $\dot{\ell}_\theta$—though the log-likelihood functional $\theta \mapsto \mathbb{E}_{\theta_0}[\log p_\theta(X)]$ varies more in θ.

Under suitable smoothness conditions on the densities p_θ (roughly, that derivatives pass through expectations; see Remark 8.1 at the end of this chapter), there are a variety of alternate definitions of Fisher information. These smoothness conditions hold for exponential families, so at least in the exponential family case, everything in this chapter is rigorous. (We note in passing that there are more general definitions of Fisher information for more general families under quadratic mean differentiability; see, for example, van der Vaart [4].) First, we note that the score function has mean zero under P_θ: we have

$$\begin{aligned}\mathbb{E}_\theta[\dot{\ell}_\theta] &= \int p_\theta(x) \nabla_\theta \log p_\theta(x) dx = \int \frac{\nabla p_\theta(x)}{p_\theta(x)} p_\theta(x) dx \\ &= \int \nabla p_\theta(x) dx \overset{(\star)}{=} \nabla \int p_\theta(x) dx = \nabla 1 = 0,\end{aligned}$$

where in equality $(\star)$ we have assumed that integration and derivation may be exchanged. Under similar conditions, we thus attain an alternate definition of Fisher information as the negative expected hessian of $\log p_\theta(X)$. Indeed,

$$\nabla^2 \log p_\theta(x) = \frac{\nabla^2 p_\theta(x)}{p_\theta(x)} - \frac{\nabla p_\theta(x) \nabla p_\theta(x)^\top}{p_\theta(x)^2} = \frac{\nabla^2 p_\theta(x)}{p_\theta(x)} - \dot{\ell}_\theta \dot{\ell}_\theta^\top,$$

so we have that the Fisher information is equal to

$$\begin{aligned}I_\theta = \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^\top] &= -\int p_\theta(x) \nabla^2 \log p_\theta(x) dx + \int \nabla^2 p_\theta(x) dx \\ &= -\mathbb{E}[\nabla^2 \log p_\theta(x)] + \nabla^2 \underbrace{\int p_\theta(x) dx}_{=1} = -\mathbb{E}[\nabla^2 \log p_\theta(x)]. \qquad (8.1.2)\end{aligned}$$

Summarizing, we have that

$$I_\theta = \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta] = -\mathbb{E}_\theta[\nabla^2 \log p_\theta(X)].$$

This representation also makes clear the additional fact that, if we have n i.i.d. observations from the model P_θ, then the information content similarly grows linearly, as $\log p_\theta(X_1^n) = \sum_{i=1}^n \log p_\theta(X_i)$.

We now give two examples of Fisher information, the first somewhat abstract and the second more concrete.

Example 8.1 (Canonical exponential family)**:** In a canonical exponential family model, we have $\log p_\theta(x) = \langle \theta, \phi(x) \rangle - A(\theta)$, where ϕ is the sufficient statistic and A is the log-partition function. Because $\dot{\ell}_\theta = \phi(x) - \nabla A(\theta)$ and $\nabla^2 \log p_\theta(x) = -\nabla^2 A(\theta)$ is a constant, we obtain

$$I_\theta = \nabla^2 A(\theta).$$

♣

Example 8.2 (Two parameterizations of a Bernoulli)**:** In the canonical parameterization of a Bernoulli as an exponential family model (Example 6.1), we had $p_\theta(x) = \exp(\theta x - \log(1 + e^\theta))$ for $x \in \{0, 1\}$, so by the preceding example the associated Fisher information is $\frac{e^\theta}{1+e^\theta} \frac{1}{1+e^\theta}$. If we make the change of variables $p = P_\theta(X = 1) = e^\theta / (1 + e^\theta)$, or $\theta = \log \frac{p}{1-p}$, we have $I_\theta = p(1 - p)$. On the other hand, if $P(X = x) = p^x (1 - p)^{1-x}$ for $p \in [0, 1]$, the standard formulation of the Bernoulli, then $\nabla \log P(X = x) = \frac{x}{p} - \frac{1-x}{1-p}$, so that

$$I_p = \mathbb{E}_p\left[\left(\frac{X}{p} - \frac{1-X}{1-p}\right)^2\right] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.$$

That is, the parameterization can change the Fisher information. ♣

8.2 Estimation and Fisher information: elementary considerations

The Fisher information has intimate connections to estimation, both in terms of classical estimation and the information games that we discuss subsequently. As a motivating calculation, we consider estimation of the mean of a $\mathsf{Bernoulli}(p)$ random variable, where $p \in [0,1]$, from a sample $X_1^n \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(p)$. The sample mean $\widehat{p}$ satisfies

$$\mathbb{E}[(\widehat{p} - p)^2] = \frac{1}{n} \operatorname{Var}(X) = \frac{p(1-p)}{n} = \frac{1}{I_p} \cdot \frac{1}{n},$$

where I_p is the Fisher information for the single observation $\mathsf{Bernoulli}(p)$ family as in Example 8.2. In fact, this inverse dependence on Fisher information is unavoidable, as made clear by the Cramér Rao Bound, which provides lower bounds on the mean squared error of all unbiased estimators.

Proposition 8.3 (Cramér Rao Bound)**.** *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary differentiable function and assume that the random function (estimator) T is unbiased for $\phi(\theta)$ under P_θ. Then*

$$\operatorname{Var}(T) \ge \nabla\phi(\theta)^\top I_\theta^{-1} \nabla\phi(\theta).$$

As an immediate corollary to Proposition 8.3, we may take $\phi(\theta) = \langle \lambda, \theta \rangle$ for $\lambda \in \mathbb{R}^d$. Then varying λ over all of $\mathbb{R}^d$, and we obtain that for any unbiased estimator T for the parameter $\theta \in \mathbb{R}^d$, we have $\operatorname{Var}(\langle \lambda, T \rangle) \ge \lambda^\top I_\theta^{-1} \lambda$. That is, we have

Corollary 8.4. *Let T be unbiased for the parameter θ under the distribution P_θ. Then the covariance of T has lower bound*

$$\operatorname{Cov}(T) \succeq I_\theta^{-1}.$$

In fact, the Cramér-Rao bound and Corollary 8.4 hold, in an asymptotic sense, for substantially more general settings (without the unbiasedness requirement). For example, see the books of van der Vaart [4] or Le Cam and Yang [3, Chapters 6 & 7], which show that under appropriate conditions (known variously as quadratic mean differentiability and local asymptotic normality) that no estimator can have smaller mean squared error than Fisher information in any uniform sense.

We now prove the proposition, where, as usual, we assume that it is possible to exchange differentiation and integration.

Proof Throughout this proof, all expectations and variances are computed with respect to P_θ. The idea of the proof is to choose $\lambda \in \mathbb{R}^d$ to minimize the variance

$$\operatorname{Var}(T - \langle \lambda, \dot{\ell}_\theta \rangle) \ge 0,$$

then use this λ to provide a lower bound on $\operatorname{Var}(T)$.

To that end, let $\dot{\ell}_{\theta,j} = \frac{\partial}{\partial \theta_j} \log p_\theta(X)$ denote the jth component of the score vector. Because $\mathbb{E}_\theta[\dot{\ell}_\theta] = 0$, we have the covariance equality

$$\begin{aligned}\operatorname{Cov}(T - \phi(\theta), \dot{\ell}_{\theta,j}) = \mathbb{E}[(T - \phi(\theta))\dot{\ell}_{\theta,j}] = \mathbb{E}[T\dot{\ell}_{\theta,j}] &= \int T(x) \frac{\frac{\partial}{\partial \theta_j} p_\theta(x)}{p_\theta(x)} p_\theta(x) dx \\ &= \frac{\partial}{\partial \theta_j} \int T(x) p_\theta(x) dx = \frac{\partial}{\partial \theta_j} \phi(\theta),\end{aligned}$$

where in the final step we used that T is unbiased for $\phi(\theta)$. Using the preceding equality,

$$\mathrm{Var}(T - \langle \lambda, \dot{\ell}_\theta \rangle) = \mathrm{Var}(T) + \lambda^\top I_\theta \lambda - 2\mathbb{E}[(T - \phi(\theta))\langle \lambda, \dot{\ell}_\theta \rangle] = \mathrm{Var}(T) + \lambda^\top I_\theta \lambda - 2\langle \lambda, \nabla \phi(\theta) \rangle.$$

Taking $\lambda = I_\theta^{-1} \nabla \phi(\theta)$ gives $0 \le \mathrm{Var}(T - \langle \lambda, \dot{\ell}_\theta \rangle) = \mathrm{Var}(T) - \nabla \phi(\theta)^\top I_\theta^{-1} \nabla \phi(\theta)$, and rearranging gives the result. □

8.3 Connections between Fisher information and divergence measures

By making connections between Fisher information and certain divergence measures, such as KL-divergence and mutual (Shannon) information, we gain additional insights into the structure of distributions, as well as optimal estimation and encoding procedures. As a consequence of the asymptotic expansions we make here, we see that estimation of 1-dimensional parameters is governed (essentially) by moduli of continuity of the loss function with respect to the metric induced by Fisher information; in short, Fisher information is an unavoidable quantity in estimation. We motivate our subsequent development with the following example.

Example 8.5 (Divergences in exponential families)**:** Consider the exponential family density $p_\theta(x) = h(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta))$. Then a straightforward calculation implies that for any θ_1 and θ_2, the KL-divergence between distributions P_{θ_1} and P_{θ_2} is

$$D_{\mathrm{kl}}\left(P_{\theta_1} \| P_{\theta_2}\right) = A(\theta_2) - A(\theta_1) - \left\langle \nabla A(\theta_1), \theta_2 - \theta_1 \right\rangle.$$

That is, the divergence is simply the difference between $A(\theta_2)$ and its first order expansion around θ_1. This suggests that we may approximate the KL-divergence via the quadratic remainder in the first order expansion. Indeed, as A is infinitely differentiable (it is an exponential family model), the Taylor expansion becomes

$$\begin{aligned} D_{\mathrm{kl}}\left(P_{\theta_1} \| P_{\theta_2}\right) &= \frac{1}{2} \left\langle \theta_1 - \theta_2, \nabla^2 A(\theta_1)(\theta_1 - \theta_2) \right\rangle + O(\|\theta_1 - \theta_2\|^3) \\ &= \frac{1}{2} \left\langle \theta_1 - \theta_2, I_{\theta_1}(\theta_1 - \theta_2) \right\rangle + O(\|\theta_1 - \theta_2\|^3). \end{aligned}$$

♣

In particular, KL-divergence is roughly quadratic for exponential family models, where the quadratic form is given by the Fisher information matrix. We also remark in passing that for a convex function f, the Bregman divergence (associated with f) between points x and y is given by $B_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$; such divergences are common in convex analysis, optimization, and differential geometry. Making such connections deeper and more rigorous is the goal of the field of information geometry (see the book of Amari and Nagaoka [1] for more).

We can generalize this example substantially under appropriate smoothness conditions. Indeed, we have

Proposition 8.6. *For appropriately smooth families of distributions* $\{P_\theta\}_{\theta \in \Theta}$,

$$D_{\mathrm{kl}}\left(P_{\theta_1} \| P_{\theta_2}\right) = \frac{1}{2} \left\langle \theta_1 - \theta_2, I_{\theta_1}(\theta_1 - \theta_2) \right\rangle + o(\|\theta_1 - \theta_2\|^2). \tag{8.3.1}$$

We only sketch the proof, as making it fully rigorous requires measure-theoretic arguments and Lebesgue's dominated convergence theorem.

Sketch of Proof By a Taylor expansion of the log density $\log p_{\theta_2}(x)$ about θ_1, we have

$$\begin{aligned}\log p_{\theta_2}(x) &= \log p_{\theta_1}(x) + \langle \nabla \log p_{\theta_1}(x), \theta_1 - \theta_2 \rangle \\ &\quad + \frac{1}{2}(\theta_1 - \theta_2)^\top \nabla^2 \log p_{\theta_1}(x)(\theta_1 - \theta_2) + R(\theta_1, \theta_2, x),\end{aligned}$$

where $R(\theta_1, \theta_2, x) = O_x(\|\theta_1 - \theta_2\|^3)$ is the remainder term, where O_x denotes a hidden dependence on x. Taking expectations and assuming that we can interchange differentiation and expectation appropriately, we have

$$\begin{aligned}\mathbb{E}_{\theta_1}[\log p_{\theta_2}(X)] &= \mathbb{E}_{\theta_1}[\log p_{\theta_1}(X)] + \left\langle \mathbb{E}_{\theta_1}[\dot{\ell}_{\theta_1}], \theta_1 - \theta_2 \right\rangle \\ &\quad + \frac{1}{2}(\theta_1 - \theta_2)^\top \mathbb{E}_{\theta_1}[\nabla^2 \log p_{\theta_1}(X)](\theta_1 - \theta_2) + \mathbb{E}_{\theta_1}[R(\theta_1, \theta_2, X)] \\ &= \mathbb{E}_{\theta_1}[\log p_{\theta_1}(X)] - \frac{1}{2}(\theta_1 - \theta_2)^\top I_{\theta_1}(\theta_1 - \theta_2) + o(\|\theta_1 - \theta_2\|^2),\end{aligned}$$

where we have assumed that the $O(\|\theta_1 - \theta_2\|^3)$ remainder is uniform enough in X that $\mathbb{E}[R] = o(\|\theta_1 - \theta_2\|^2)$ and used that the score function $\dot{\ell}_\theta$ is mean zero under P_θ. □

We may use Proposition 8.6 to give a somewhat more general version of the Cramér-Rao bound (Proposition 8.3) that applies to more general (sufficiently smooth) estimation problems. Indeed, we will show that Le Cam's method (recall Chapter 13.3) is (roughly) performing a type of discrete second-order approximation to the KL-divergence, then using this to provide lower bounds. More concretely, suppose we are attempting to estimate a parameter θ parameterizing the family $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$, and assume that $\Theta \subset \mathbb{R}^d$ and $\theta_0 \in \operatorname{int} \Theta$. Consider the minimax rate of estimation of θ_0 in a neighborhood around θ_0; that is, consider

$$\inf_{\widehat{\theta}} \sup_{\theta = \theta_0 + v \in \Theta} \mathbb{E}_\theta[\|\widehat{\theta}(X_1^n) - \theta\|^2],$$

where the observations X_i are drawn i.i.d. P_θ. Fixing $v \in \mathbb{R}^d$ and setting $\theta = \theta_0 + \delta v$ for some $\delta > 0$, Le Cam's method (13.3.3) then implies that

$$\inf_{\widehat{\theta}} \max_{\theta \in \{\theta_0, \theta + \delta v\}} \mathbb{E}_\theta[\|\widehat{\theta}(X_1^n) - \theta\|^2] \ge \frac{\delta^2 \|v\|^2}{8} \left[1 - \left\|P_{\theta_0}^n - P_{\theta_0 + \delta v}^n\right\|_{\rm TV}\right].$$

Using Pinsker's inequality that $2\|P - Q\|_{\rm TV}^2 \le D_{\rm kl}(P\|Q)$ and the asymptotic quadratic approximation (8.3.1), we have

$$\left\|P_{\theta_0}^n - P_{\theta_0 + \delta v}^n\right\|_{\rm TV} \le \sqrt{\frac{n}{2} D_{\rm kl}(P_{\theta_0} \| P_{\theta_0 + \delta v})} = \frac{\sqrt{n}}{2}\left(\delta^2 v^\top I_{\theta_0} v + o(\delta^2 \|v\|^2)\right)^{\frac{1}{2}}.$$

By taking $\delta^2 = (n v^\top I_{\theta_0} v)^{-1}$, for large enough v and n we know that $\theta_0 + \delta v \in \operatorname{int} \Theta$ (so that the distribution $P_{\theta_0 + \delta v}$ exists), and for large n, the remainder term $o(\delta^2 \|v\|^2)$ becomes negligible. Thus we obtain

$$\inf_{\widehat{\theta}} \max_{\theta \in \{\theta_0, \theta + \delta v\}} \mathbb{E}_\theta[\|\widehat{\theta}(X_1^n) - \theta\|^2] \gtrsim \frac{\delta^2 \|v\|^2}{16} = \frac{1}{16} \frac{\|v\|^2}{n v^\top I_{\theta_0} v}. \tag{8.3.2}$$

In particular, in one-dimension, inequality (8.3.2) implies a result generalizing the Cramér-Rao bound. We have the following asymptotic local minimax result:

Corollary 8.7. *Let $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$, where $\Theta \subset \mathbb{R}$, be a family of distributions satisfying the quadratic approximation condition of Proposition 8.6. Then there exists a constant $c > 0$ such that*

$$\lim_{v \to \infty} \lim_{n \to \infty} \inf_{\widehat{\theta}_n} \sup_{\theta : |\theta - \theta_0| \le v / \sqrt{n}} \mathbb{E}_\theta \left[(\widehat{\theta}_n(X_1^n) - \theta)^2 \right] \ge c \frac{1}{n} I_{\theta_0}^{-1}.$$

Written differently (and with minor extension), Corollary 8.7 gives a lower bound based on a local modulus of continuity of the loss function with respect to the metric induced by the Fisher information. Indeed, suppose we wish to estimate a parameter θ in the neighborhood of θ_0 (where the neighborhood size decreases as $1/\sqrt{n}$) according to some loss function $\ell : \Theta \times \Theta \to \mathbb{R}$. Then if we define the modulus of continuity of ℓ with respect to the Fisher information metric as

$$\omega_\ell(\delta, \theta_0) := \sup_{v : \|v\| \le 1} \frac{\ell(\theta_0, \theta_0 + \delta v)}{\delta^2 v^\top I_{\theta_0} v},$$

the combination of Corollary 8.7 and inequality (8.3.2) shows that the local minimax rate of estimating $\mathbb{E}_\theta[\ell(\widehat{\theta}_n, \theta)]$ for θ near θ_0 must be at least $\omega_\ell(n^{-1/2}, \theta_0)$. For more on connections between moduli of continuity and estimation, see, for example, Donoho and Liu [2].

Remark 8.1: In order to make all of our exchanges of differentiation and expectation rigorous, we must have some conditions on the densities we consider. One simple condition sufficient to make this work is via Lebesgue's dominated convergence theorem. Let $f : \mathcal{X} \times \Theta \to \mathbb{R}$ be a differentiable function. For a fixed base measure μ assume there exists a function g such that $g(x) \ge \|\nabla_\theta f(x, \theta)\|$ for all θ, where

$$\int_{\mathcal{X}} g(x) d\mu(x) < \infty.$$

Then in this case, we have $\nabla_\theta \int f(x, \theta) d\mu(x) = \int \nabla_\theta f(x, \theta) d\mu(x)$ by the mean-value theorem and definition of a derivative. (Note that for all θ_0 we have $\sup_{v : \|v\|_2 \le \delta} \|\nabla_\theta f(x, \theta)\|_2 \big|_{\theta = \theta_0 + v} \le g(x)$.) More generally, this type of argument can handle absolutely continuous functions, which are differentiable almost everywhere. ♢

Bibliography

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[2] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence I. Technical Report 137, University of California, Berkeley, Department of Statistics, 1987.

[3] L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[4] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

Chapter 9

Universal prediction and coding

In this chapter, we explore sequential game playing and online probabilistic prediction schemes. These have applications in coding when the true distribution of the data is unknown, biological algorithms (encoding genomic data, for example), control, and a variety of other areas. The field of universal prediction is broad; in addition to this chapter touching briefly on a few of the techniques therein and their relationships with statistical modeling and inference procedures, relevant reading includes the survey by Merhav and Feder [9], the more recent book of Grünwald [5], and Tsachy Weissman's EE376c course at Stanford.

9.1 Universal and sequential prediction

We begin by defining the universal prediction (and universal coding) problems. In this setting, we assume we are playing a game in which given a sequence X_1^n of data, we would like to predict the data (which, as we saw in Example 7.5, is the same as encoding the data) as as if we *knew* the true distribution of the data. Or, in more general settings, we would like to predict the data as well as all predictive distributions P from some family of distributions $\mathcal{P}$, even if *a priori* we know little about the coming sequence of data.

We consider two versions of this game: the probabilistic version and the adversarial version. We shall see that they have similarities, but there are also a few important distinctions between the two. For both of the following definitions of sequential prediction games, we assume that p and q are densities or probability mass functions in the case that $\mathcal{X}$ is continuous or discrete (this is no real loss of generality) for distributions P and Q.

We begin with the adversarial case. Given a sequence $x_1^n \in \mathcal{X}^n$, the *regret* of the distribution Q for the sequence x_1^n with respect to the distribution P is

$$\mathsf{Reg}(Q, P, x_1^n) := \log \frac{1}{q(x_1^n)} - \log \frac{1}{p(x_1^n)} = \sum_{i=1}^n \log \frac{1}{q(x_i \mid x_1^{i-1})} - \log \frac{1}{p(x_i \mid x_1^{i-1})}, \tag{9.1.1}$$

where we have written it as the sum over $q(x_i \mid x_1^{i-1})$ to emphasize the sequential nature of the game. Associated with the regret of the sequence x_1^n is the *adversarial regret* (usually simply called the regret) of Q with respect to the family $\mathcal{P}$ of distributions, which is

$$\mathfrak{R}_n^{\mathcal{X}}(Q, \mathcal{P}) := \sup_{P \in \mathcal{P}, x_1^n \in \mathcal{X}^n} \mathsf{Reg}(Q, P, x_1^n). \tag{9.1.2}$$

In more generality, we may which to use a loss function ℓ different than the log loss; that is, we might wish to measure a loss-based version the regret as

$$\sum_{i=1}^{n} \ell(x_i, Q(\cdot \mid x_1^{i-1})) - \ell(x_i, P(\cdot \mid x_1^{i-1})),$$

where $\ell(x_i, P)$ indicates the loss suffered on the point x_i when the distribution P over X_i is played, and $P(\cdot \mid x_1^{i-1})$ denotes the conditional distribution of X_i given x_1^{i-1} according to P. We defer discussion of such extensions later, focusing on the log loss for now because of its natural connections with maximum likelihood and coding.

A less adversarial problem is to minimize the *redundancy*, which is the expected regret under a distribution P. In this case, we define the redunancy of Q with respect to P as the expected regret of Q with respect to P under the distribution P, that is,

$$\mathsf{Red}_n(Q, P) := \mathbb{E}_P\left[\log \frac{1}{q(X_1^n)} - \log \frac{1}{p(X_1^n)}\right] = D_{\rm kl}\left(P \| Q\right), \tag{9.1.3}$$

where the dependence on n is implicit in the KL-divergence. The worst-case redundancy with respect to a class $\mathcal{P}$ is then

$$\mathfrak{R}_n(Q, \mathcal{P}) := \sup_{P \in \mathcal{P}} \mathsf{Red}_n(Q, P). \tag{9.1.4}$$

We now give two examples to illustrate the redundancy.

Example 9.1 (Example 7.5 on coding, continued)**:** We noted in Example 7.5 that for any p.m.f.s p and q on the set $\mathcal{X}$, it is possible to define coding schemes C_p and C_q with code lengths

$$\ell_{C_p}(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil \quad \text{and} \quad \ell_{C_q}(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil.$$

Conversely, given (uniquely decodable) encoding schemes C_p and $C_q : \mathcal{X} \to \{0,1\}^*$, the functions $p_{C_p}(x) = 2^{-\ell_{C_p}(x)}$ and $q_{C_q}(x) = 2^{-\ell_{C_q}(x)}$ satisfy $\sum_x p_{C_p}(x) \le 1$ and $\sum_x q_{C_q}(x) \le 1$. Thus, the redundancy of Q with respect to P is the additional number of bits required to encode variables distributed according to P when we assume they have distribution Q:

$$\begin{aligned} \mathsf{Red}_n(Q, P) &= \sum_{i=1}^{n} \mathbb{E}_P\left[\log \frac{1}{q(X_i \mid X_1^{i-1})} - \log \frac{1}{p(X_i \mid X_1^{i-1})}\right] \\ &= \sum_{i=1}^{n} \mathbb{E}_P[\ell_{C_q}(X_i)] - \mathbb{E}_P[\ell_{C_p}(X_i)], \end{aligned}$$

where $\ell_C(x)$ denotes the number of bits C uses to encode x. Note that, as in Chapter 5, the code $\lceil -\log p(x) \rceil$ is (essentially) optimal. ♣

As another example, we may consider a filtering or prediction problem for a linear system.

Example 9.2 (Prediction in a linear system)**:** Suppose we believe that a sequence of random variables $X_i \in \mathbb{R}^d$ are Markovian, where X_i given X_{i-1} is normally distributed with mean $AX_{i-1} + g$, where A is an unknown matrix and $g \in \mathbb{R}^d$ is a constant drift term. Concretely, we assume $X_i \sim \mathsf{N}(AX_{i-1} + g, \sigma^2 I_{d \times d})$, where we assume σ^2 is fixed and known. For our class of predicting distributions Q, we may look at those that at iteration i predict $X_i \sim \mathsf{N}(\mu_i, \sigma^2 I)$. In this case, the regret is given by

$$\mathsf{Reg}(Q, P, x_1^n) = \sum_{i=1}^n \frac{1}{2\sigma^2} \left\|\mu_i - x_i\right\|_2^2 - \frac{1}{2\sigma^2} \left\|Ax_{i-1} + g - x_i\right\|_2^2,$$

while the redundancy is

$$\mathsf{Red}_n(Q, P) = \frac{1}{2\sigma^2} \sum_{i=1}^n \mathbb{E}[\|AX_{i-1} + g - \mu_i(X_1^{i-1})\|_2^2],$$

assuming that P is the linear Gaussian Markov chain specified. ♣

9.2 Minimax strategies for regret

Our definitions in place, we now turn to strategies for attaining the optimal regret in the adversarial setting. We discuss this only briefly, as optimal strategies are somewhat difficult to implement, and the redundancy setting allows (for us) easier exploration.

We begin by describing a notion of complexity that captures the best possible regret in the adversarial setting. In particular, assume without loss of generality that we have a set of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ parameterized by $\theta \in \Theta$, where the distributions are supported on $\mathcal{X}^n$. We define the complexity of the set $\mathcal{P}$ (viz. the complexity of Θ) as

$$\mathsf{Comp}_n(\Theta) := \log \int_{\mathcal{X}^n} \sup_{\theta \in \Theta} p_\theta(x_1^n) dx_1^n \quad \text{or generally} \quad \mathsf{Comp}_n(\Theta) := \log \int_{\mathcal{X}^n} \sup_{\theta \in \Theta} p_\theta(x_1^n) d\mu(x_1^n), \tag{9.2.1}$$

where μ is some base measure on $\mathcal{X}^n$. Note that we may have $\mathsf{Comp}_n(\Theta) = +\infty$, especially when Θ is non-compact. This is not particularly uncommon, for example, consider the case of a normal location family model over $\mathcal{X} = \mathbb{R}$ with $\Theta = \mathbb{R}$.

It turns out that the complexity is precisely the minimax regret in the adversarial setting.

Proposition 9.3. *The minimax regret*

$$\inf_Q \mathfrak{R}^{\mathcal{X}}(Q, \mathcal{P}) = \mathsf{Comp}_n(\Theta).$$

Moreover, if $\mathsf{Comp}_n(\Theta) < +\infty$*, then the* normalized maximum likelihood *distribution (also known as the* Shtarkov *distribution)* $\overline{Q}$*, defined with density*

$$\overline{q}(x_1^n) = \frac{\sup_{\theta \in \Theta} p_\theta(x_1^n)}{\int \sup_\theta p_\theta(x_1^n) dx_1^n},$$

is uniquely minimax optimal.

The proposition completely characterizes the minimax regret in the adversarial setting, and it gives the unique distribution achieving the regret. Unfortunately, in most cases it is challenging to compute the minimax optimal distribution $\overline{Q}$, so we must make approximations of some type. One approach is to make Bayesian approximations to $\overline{Q}$, as we do in the sequel when we consider redundancy rather than adversarial regret. See also the book of Grünwald [5] for more discussion of this and other issues.

Proof We begin by proving the result in the case that $\mathsf{Comp}_n < +\infty$. First, note that the normalized maximum likelihood distribution $\overline{Q}$ has constant regret:

$$\begin{aligned}\mathfrak{R}_n^{\mathcal{X}}(\overline{Q}, \mathcal{P}) &= \sup_{x_1^n \in \mathcal{X}^n} \left[\log \frac{1}{\overline{q}(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)}\right] \\ &= \sup_{x_1^n} \left[\log \frac{\int \sup_\theta p_\theta(x_1^n) dx_1^n}{\sup_\theta p_\theta(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)}\right] = \mathsf{Comp}_n(\mathcal{P}).\end{aligned}$$

Moreover, for any distribution Q on $\mathcal{X}^n$ we have

$$\begin{aligned}\mathfrak{R}_n^{\mathcal{X}}(Q, \mathcal{P}) &\ge \int \left[\log \frac{1}{q(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)}\right] \overline{q}(x_1^n) dx_1^n \\ &= \int \left[\log \frac{\overline{q}(x_1^n)}{q(x_1^n)} + \mathsf{Comp}_n(\Theta)\right] \overline{q}(x_1^n) dx_1^n \\ &= D_{\rm kl}\left(\overline{Q} \| Q\right) + \mathsf{Comp}_n(\Theta), \qquad (9.2.2)\end{aligned}$$

so that $\overline{Q}$ is uniquely minimax optimal, as $D_{\rm kl}\left(\overline{Q} \| Q\right) > 0$ unless $\overline{Q} = Q$.

Now we show how to extend the lower bound (9.2.2) to the case when $\mathsf{Comp}_n(\Theta) = +\infty$. Let us assume without loss of generality that $\mathcal{X}$ is countable and consists of points $x_1, x_2, \ldots$ (we can discretize $\mathcal{X}$ otherwise) and assume we have $n = 1$. Fix any $\epsilon \in (0, 1)$ and construct the sequence $\theta_1, \theta_2, \ldots$ so that $p_{\theta_j}(x_j) \ge (1 - \epsilon) \sup_{\theta \in \Theta} p_\theta(x)$, and define the sets $\Theta_j = \{\theta_1, \ldots, \theta_j\}$. Clearly we have $\mathsf{Comp}(\Theta_j) \le \log j$, and if we define $\overline{q}_j(x) = \max_{\theta \in \Theta_j} p_\theta(x) / \sum_{x \in \mathcal{X}} \max_{\theta \in \Theta_j} p_\theta(x)$, we may extend the reasoning yielding inequality (9.2.2) to obtain

$$\begin{aligned}\mathfrak{R}^{\mathcal{X}}(Q, \mathcal{P}) &= \sup_{x \in \mathcal{X}} \left[\log \frac{1}{q(x)} - \log \frac{1}{\sup_{\theta \in \Theta} p_\theta(x)}\right] \\ &\ge \sum_x \overline{q}_j(x) \left[\log \frac{1}{q(x)} - \log \frac{1}{\max_{\theta \in \Theta_j} p_\theta(x)}\right] \\ &= \sum_x \overline{q}_j(x) \left[\log \frac{\overline{q}_j(x)}{q(x)} + \log \sum_{x'} \max_{\theta \in \Theta_j} p_\theta(x')\right] = D_{\rm kl}\left(\overline{Q}_j \| Q\right) + \mathsf{Comp}(\Theta_j).\end{aligned}$$

But of course, by noting that

$$\mathsf{Comp}(\Theta_j) \ge (1 - \epsilon) \sum_{i=1}^j \sup_\theta p_\theta(x_i) + \sum_{i > j} \max_{\theta \in \Theta_j} p_\theta(x_i) \to +\infty$$

as $j \to \infty$, we obtain the result when $\mathsf{Comp}_n(\Theta) = \infty$. □

We now give an example where (up to constant factor terms) we can explicitly calculate the minimax regret in the adversarial setting. In this case, we compete with the family of i.i.d. Bernoulli distributions.

Example 9.4 (Complexity of the Bernoulli distribution)**:** In this example, we consider competing against the family of Bernoulli distributions $\{P_\theta\}_{\theta \in [0,1]}$, where for a point $x \in \{0, 1\}$, we have $P_\theta(x) = \theta^x(1-\theta)^{1-x}$. For a sequence $x_1^n \in \{0,1\}^n$ with m non-zeros, we thus have for $\widehat{\theta} = m/n$ that

$$\sup_{\theta \in [0,1]} P_\theta(x_1^n) = P_{\widehat{\theta}}(x_1^n) = \widehat{\theta}^m(1-\widehat{\theta})^{n-m} = \exp(-nh_2(\widehat{\theta})),$$

where $h_2(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy. Using this representation, we find that the complexity of the Bernoulli family is

$$\mathsf{Comp}_n([0,1]) = \log \sum_{m=0}^n \binom{n}{m} e^{-nh_2(\frac{m}{n})}.$$

Rather than explicitly compute with this, we now use Stirling's approximation (cf. Cover and Thomas [3, Chapter 17]): for any $p \in (0,1)$ with $np \in \mathbb{N}$, we have

$$\binom{n}{np} \in \frac{1}{\sqrt{n}}\left[\frac{1}{\sqrt{8p(1-p)}}, \frac{1}{\sqrt{\pi p(1-p)}}\right]\exp(nh_2(p)).$$

Thus, by dealing with the boundary cases $m = n$ and $m = 0$ explicitly, we obtain

$$\begin{aligned}\sum_{m=0}^n \binom{n}{m}\exp(-nh_2(\frac{m}{n})) &= 2 + \sum_{m=1}^{n-1}\binom{n}{m}\exp(-nh_2(\frac{m}{n})) \\ &\in 2 + \left[\frac{1}{\sqrt{8}}, \frac{1}{\sqrt{\pi}}\right]\frac{1}{\sqrt{n}}\underbrace{\sum_{m=1}^{n-1}\frac{1}{\sqrt{\frac{m}{n}(1-\frac{m}{n})}}}_{\to n\int_0^1(\theta(1-\theta))^{-\frac{1}{2}}},\end{aligned}$$

the noted asymptote occuring as $n \to \infty$ by the fact that this sum is a Riemann sum for the integral $\int_0^1 \theta^{-1/2}(1-\theta)^{-1/2}d\theta$. In particular, we have that as $n \to \infty$,

$$\begin{aligned}\inf_Q \mathfrak{R}_n^{\mathcal{X}}(Q, \mathcal{P}) = \mathsf{Comp}_n([0,1]) &= \log\left(2 + [8^{-1/2}, \pi^{-1/2}]n^{1/2}\int_0^1 \frac{1}{\sqrt{\theta(1-\theta)}}d\theta\right) + o(1) \\ &= \frac{1}{2}\log n + \log\int_0^1\frac{1}{\sqrt{\theta(1-\theta)}}d\theta + O(1).\end{aligned}$$

We remark in passing that this is equal to $\frac{1}{2}\log n + \log\int_0^1 \sqrt{I_\theta}d\theta$, where I_θ denotes the Fisher information of the Bernoulli family (recall Example 8.2). We will see that this holds in more generality, at least for redundancy, in the sequel. ♣

9.3 Mixture (Bayesian) strategies and redundancy

We now turn to a slightly less adversarial setting, where we assume that we compete against a random sequence X_1^n of data, drawn from some fixed distribution P, rather than an adversarially chosen sequence x_1^n. Thinking of this problem as a game, we choose a distribution Q according to which we make predictions (based on previous data), and nature chooses a distribution $P_\theta \in$

$\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$. In the simplest case—upon which we focus—the data X_1^n are then generated i.i.d. according to P_θ, and we suffer expected regret (or redundancy)

$$\mathsf{Red}_n(Q, P_\theta) = \mathbb{E}_\theta\left[\log\frac{1}{q(X_1^n)}\right] - \mathbb{E}_\theta\left[\log\frac{1}{p_\theta(X_1^n)}\right] = D_{\rm kl}\left(P_\theta^n \| Q_n\right), \tag{9.3.1}$$

where we use Q_n to denote that Q is applied on all n data points (in a sequential fashion, as $Q(\cdot \mid X_1^{i-1})$). In this expression, q and p denote the densities of Q and P, respectively. In a slightly more general setting, we may consider the expected regret of Q with respect to a distribution P_θ even under model mis-specification, meaning that the data is generated according to an alternate distribution P. In this case, the (more general) redundancy becomes

$$\mathbb{E}_P\left[\log\frac{1}{q(X_1^n)} - \log\frac{1}{p_\theta(X_1^n)}\right]. \tag{9.3.2}$$

In both cases (9.3.1) and (9.3.2), we would like to be able to guarantee that the redundancy grows more slowly than n as $n \to \infty$. That is, we would like to find distributions Q such that, for any $\theta_0 \in \Theta$, we have $\frac{1}{n} D_{\rm kl}\left(P_{\theta_0}^n \| Q_n\right) \to 0$ as $n \to \infty$. Assuming we could actually obtain such a distribution in general, this is interesting because (even in the i.i.d. case) for *any* fixed distribution $P_\theta \neq P_{\theta_0}$, we must have $D_{\rm kl}\left(P_{\theta_0}^n \| P_\theta^n\right) = n D_{\rm kl}\left(P_{\theta_0} \| P_\theta^n\right) = \Omega(n)$. A standard approach to attaining such guarantees is the *mixture approach*, which is based on choosing Q as a convex combination (mixture) of all the possible source distributions P_θ for $\theta \in \Theta$.

In particular, given a prior distribution π (weighting function integrating to 1) over Θ, we define the mixture distribution

$$Q_n^\pi(A) = \int_\Theta \pi(\theta) P_\theta(A) d\theta \quad \text{for} \quad A \subset \mathcal{X}^n. \tag{9.3.3}$$

Rewriting this in terms of densities p_θ, we have

$$q_n^\pi(x_1^n) = \int_\Theta \pi(\theta) p_\theta(x_1^n) d\theta.$$

Conceptually, this gives a simple prediction scheme, where at iteration i we play the density

$$q^\pi(x_i \mid x_1^{i-1}) = \frac{q^\pi(x_1^i)}{q^\pi(x_1^{i-1})},$$

which is equivalent to playing

$$q^\pi(x_i \mid x_1^{i-1}) = \int_\Theta q(x_i, \theta \mid x_1^{i-1}) d\theta = \int_\Theta p_\theta(x_i) \pi(\theta \mid x_1^{i-1}) d\theta,$$

by construction of the distributions Q^π as mixtures of i.i.d. P_θ. Here the posterior distribution $\pi(\theta \mid x_1^{i-1})$ is given by

$$\pi(\theta \mid x_1^{i-1}) = \frac{\pi(\theta) p_\theta(x_1^{i-1})}{\int_\Theta \pi(\theta') p_{\theta'}(x_1^{i-1}) d\theta'} = \frac{\pi(\theta) \exp\left(-\log\frac{1}{p_\theta(x_1^{i-1})}\right)}{\int_\Theta \pi(\theta') p_{\theta'}(x_1^{i-1}) d\theta'}, \tag{9.3.4}$$

where we have emphasized that this strategy exhibits an *exponential weighting* approach, where distribution weights are scaled exponentially by their previous loss performance of $\log 1/p_\theta(x_1^{i-1})$.

This mixture construction (9.3.3), with the weighting scheme (9.3.4), enjoys very good performance. In fact, we say that so long as the prior π puts non-zero mass over all of Θ, under some appropriate smoothness conditions, the scheme Q^π is universal, meaning that $D_{\rm kl}\left(P_\theta^n \| Q_n^\pi\right) = o(n)$. We have the following theorem illustrating this effect. In the theorem, we let π be a density on Θ, and we assume the Fisher information I_θ for the family $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ exists in a neighborhood of $\theta_0 \in \operatorname{int} \Theta$, and that the distributions P_θ are sufficiently regular that differentiation and integration can be interchanged. (See Clarke and Barron [2] for precise conditions.) We have

Theorem 9.5 (Clarke and Barron [2])**.** *Under the above conditions, if $Q_n^\pi = \int P_\theta^n \pi(\theta) d\theta$ is the mixture* (9.3.3)*, then*

$$D_{\rm kl}\left(P_{\theta_0}^n \| Q_n^\pi\right) - \frac{d}{2} \log \frac{n}{2\pi e} \to \log \frac{1}{\pi(\theta_0)} + \frac{1}{2} \log \det(I_{\theta_0}) \quad \textit{as } n \to \infty. \tag{9.3.5}$$

While we do not rigorously prove the theorem, we give a sketch showing the main components of the result based on asymptotic normality arguments for the maximum likelihood estimator in Section 9.4. See Clarke and Barron [2] for a full proof.

Example 9.6 (Bernoulli distributions with a Beta prior)**:** Consider the class of binary (i.i.d. or memoryless) Bernoulli sources, that is, the X_i are i.i.d $\mathsf{Bernoulli}(\theta)$, where $\theta = P_\theta(X = 1) \in [0, 1]$. The Beta$(\alpha, \beta)$-distribution prior on θ is the mixture π with density

$$\pi(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1}$$

on $[0, 1]$, where $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$ denotes the gamma function. We remark that that under the Beta(α, β) distribution, we have $\mathbb{E}_\pi[\theta] = \frac{\alpha}{\alpha + \beta}$. (See any undergraduate probability text for such results.)

If we play via a mixture of Bernoulli distributions under such a Beta-prior for θ, by Theorem 9.5 we have a universal prediction scheme. We may also explicitly calculate the predictive distribution Q. To do so, we first compute the posterior $\pi(\theta \mid X_1^i)$ as in expression (9.3.4). Let $S_i = \sum_{j=1}^i X_j$ be partial sum of the Xs up to iteration i. Then

$$\pi(\theta \mid x_1^i) = \frac{p_\theta(x_1^i)\pi(\theta)}{q(x_1^i)} \propto \theta^{S_i}(1 - \theta)^{i - S_i} \theta^{\alpha - 1} \theta^{\beta - 1} = \theta^{\alpha + S_i - 1}(1 - \theta)^{\beta + i - S_i - 1},$$

where we have ignored the denominator as we must simply normalize the above quantity in θ. But by inspection, the posterior density of $\theta \mid X_1^i$ is a Beta$(\alpha + S_i, \beta + i - S_i)$ distribution. Thus to compute the predictive distribution, we note that $\mathbb{E}_\theta[X_i] = \theta$, so we have

$$Q(X_i = 1 \mid X_1^i) = \mathbb{E}_\pi[\theta \mid X_1^i] = \frac{S_i + \alpha}{i + \alpha + \beta}.$$

Moreover, Theorem 9.5 shows that when we play the prediction game with a Beta(α, β)-prior, we have redundancy scaling as

$$D_{\rm kl}\left(P_{\theta_0}^n \| Q_n^\pi\right) = \frac{1}{2} \log \frac{n}{2\pi e} + \log \left[\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \frac{1}{\theta_0^{\alpha - 1}(1 - \theta_0)^{\beta - 1}}\right] + \frac{1}{2} \log \frac{1}{\theta_0(1 - \theta_0)} + o(1)$$

for $\theta_0 \in (0, 1)$. ♣

As one additional interesting result, we show that mixture models are actually quite robust, even under model mis-specification, that is, when the true distribution generating the data does not belong to the class $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$. That is, mixtures can give good performance for the generalized redundancy quantity (9.3.2). For this next result, we as usual define the mixture distribution Q^π over the set $\mathcal{X}$ via $Q^\pi(A) = \int_\Theta P_\theta(A)d\pi(\theta)$. We may also restrict this mixture distribution to a subset $\Theta_0 \subset \Theta$ by defining

$$Q^\pi_{\Theta_0}(A) = \frac{1}{\pi(\Theta_0)} \int_{\Theta_0} P_\theta(A)d\pi(\theta).$$

Then we obtain the following robustness result.

Proposition 9.7. *Assume that P_θ have densities p_θ over $\mathcal{X}$, let P be any distribution having density p over $\mathcal{X}$, and let q^π be the density associated with Q^π. Then for any $\Theta_0 \subset \Theta$,*

$$\mathbb{E}_P\left[\log\frac{1}{q^\pi(X)} - \log\frac{1}{p_\theta(X)}\right] \le \log\frac{1}{\pi(\Theta_0)} + D_{\rm kl}\left(P\|Q^\pi_{\Theta_0}\right) - D_{\rm kl}\left(P\|P_\theta\right).$$

In particular, Proposition 9.7 shows that so long as the mixture distributions $Q^\pi_{\Theta_0}$ can closely approximate P_θ, then we attain a convergence guarantee nearly as good as any in the family $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$. (This result is similar in flavor to the mutual information bound (16.1.3), Corollary 16.2, and the *index of resolvability* quantity.)

Proof Fix any $\Theta_0 \subset \Theta$. Then we have $q^\pi(x) = \int_\Theta p_\theta(x)d\pi(\theta) \ge \int_{\Theta_0} p_\theta(x)d\pi(\theta)$. Thus we have

$$\begin{aligned}
\mathbb{E}_P\left[\log\frac{p(X)}{q^\pi(X)}\right] &\le \mathbb{E}_P\left[\inf_{\Theta_0\subset\Theta}\log\frac{p(X)}{\int_{\Theta_0} p_\theta(x)d\pi(\theta)}\right] \\
&= \mathbb{E}_P\left[\inf_{\Theta_0}\log\frac{p(X)\pi(\Theta_0)}{\pi(\Theta_0)\int_{\Theta_0} p_\theta(x)d\pi(\theta)}\right] = \mathbb{E}_P\left[\inf_{\Theta_0}\log\frac{p(X)}{\pi(\Theta_0)q^\pi_{\Theta_0}(X)}\right].
\end{aligned}$$

This is certainly smaller than the same quantity with the infimum outside the expectation, and noting that

$$\mathbb{E}_P\left[\log\frac{1}{q^\pi(X)} - \log\frac{1}{p_\theta(X)}\right] = \mathbb{E}_P\left[\log\frac{p(X)}{q^\pi(X)}\right] - \mathbb{E}_P\left[\log\frac{p(X)}{p_\theta(X)}\right]$$

gives the result. □

9.3.1 Bayesian redundancy and objective, reference, and Jeffreys priors

We can also imagine a slight variant of the redundancy game we have described to this point. Instead of choosing a distribution Q and allowing nature to choose a distribution P_θ, we could switch the order of the game. In particular, we could assume that nature first chooses prior distribution π on θ, and without seeing θ (but with knowledge of the distribution π) we choose the predictive distribution Q. This leads to the *Bayesian redundancy*, which is simply the expected redundancy we suffer:

$$\int_\Theta \pi(\theta)D_{\rm kl}\left(P_\theta^n\|Q_n\right)d\theta.$$

However, recalling our calculations with mutual information (equations (13.4.4), (16.1.1), and (16.1.4)), we know that the Bayes-optimal prediction distribution is Q^π_n. In particular, if we let T denote a random variable distributed according to π, and conditional on $T = \theta$ assume that the X_i are drawn according to P_θ, we have that the mutual information between T and X_1^n is

$$I_\pi(T; X_1^n) = \int \pi(\theta) D_{\rm kl}\left(P_\theta^n \| Q_n^\pi\right) d\theta = \inf_Q \int \pi(\theta) D_{\rm kl}\left(P_\theta^n \| Q\right) d\theta. \tag{9.3.6}$$

With Theorem 9.5 in hand, we can give a somewhat more nuanced picture of this mutual information quantity. As a first consequence of Theorem 9.5, we have that

$$I_\pi(T; X_1^n) = \frac{d}{2} \log \frac{n}{2\pi e} + \int \log \frac{\sqrt{\det I_\theta}}{\pi(\theta)} \pi(\theta) d\theta + o(1), \tag{9.3.7}$$

where I_θ denotes the Fisher information matrix for the family $\{P_\theta\}_{\theta \in \Theta}$. One strand of Bayesian statistics—we will not delve too deeply into this now, instead referring to the survey by Bernardo [1]—known as reference analysis, advocates that in performing a Bayesian analysis, we should choose the prior π that maximizes the mutual information between the parameters θ about which we wish to make inferences and any observations X_1^n available. Moreover, in this set of strategies, one allows n to tend to ∞, as we wish to take advantage of any data we might actually see. The asymptotic formula (9.3.7) allows us to choose such a prior.

In a different vein, Jeffreys [7] proposed that if the square root of the determinant of the Fisher information was integrable, then one should take π as

$$\pi_{\rm jeffreys}(\theta) = \frac{\sqrt{\det I_\theta}}{\int_\Theta \sqrt{\det I_\theta} d\theta}$$

known as the *Jeffreys prior*. Jeffreys originally proposed this for invariance reasons, as the inferences made on the parameter θ under the prior $\pi_{\rm jeffreys}$ are identical to those made on a transformed parameter $\phi(\theta)$ under the appropriately transformed Jeffreys prior. The asymptotic expression (9.3.7), however, shows that the Jeffreys prior is the asymptotic reference prior. Indeed, computing the integral in (9.3.7), we have

$$\begin{aligned}
\int_\Theta \pi(\theta) \log \frac{\sqrt{\det I_\theta}}{\pi(\theta)} d\theta &= \int_\Theta \pi(\theta) \log \frac{\pi_{\rm jeffreys}(\theta)}{\pi(\theta)} d\theta + \log \int \sqrt{\det I_\theta} d\theta \\
&= -D_{\rm kl}\left(\pi \| \pi_{\rm jeffreys}\right) + \log \int \sqrt{\det I_\theta} d\theta,
\end{aligned}$$

whenever the Jeffreys prior exists. Moreover, we see that in an asymptotic sense, the worst-case prior distribution π for nature to play is given by the Jeffreys prior, as otherwise the $-D_{\rm kl}\left(\pi \| \pi_{\rm jeffreys}\right)$ term in the expected (Bayesian) redundancy is negative.

Example 9.8 (Jeffreys priors and the exponential distribution)**:** Let us now assume that our source distributions P_θ are exponential distributions, meaning that $\theta \in (0, \infty)$ and we have density $p_\theta(x) = \exp(-\theta x - \log \frac{1}{\theta})$ for $x \in [0, \infty)$. This is clearly an exponential family model, and the Fisher information is easy to compute as $I_\theta = \frac{\partial^2}{\partial \theta^2} \log \frac{1}{\theta} = 1/\theta^2$ (cf. Example 8.1).
In this case, the Jeffreys prior is $\pi_{\rm jeffreys}(\theta) \propto \sqrt{I} = 1/\theta$, but this "density" does not integrate over $[0, \infty)$. One approach to this difficulty, advocated by Bernardo [1, Definition 3] (among others) is to just proceed formally and notice that after observing a single datapoint, the

"posterior" distribution $\pi(\theta \mid X)$ is well-defined. Following this idea, note that after seeing some data $X_1, \ldots, X_i$, with $S_i = \sum_{j=1}^i X_j$ as the partial sum, we have

$$\pi(\theta \mid x_1^i) \propto p_\theta(x_1^i) \pi_{\text{jeffreys}}(\theta) = \theta^i \exp\bigg(-\theta \sum_{j=1}^i x_j \bigg) \frac{1}{\theta} = \theta^{i-1} \exp(-\theta S_i).$$

Integrating, we have for $s_i = \sum_{j=1}^i x_j$

$$q(x \mid x_1^i) = \int_0^\infty p_\theta(x) \pi(\theta \mid x_1^i) d\theta \propto \int_0^\infty \theta e^{-\theta x} \theta^{i-1} e^{-\theta s_i} d\theta = \frac{1}{(s_i + x)^{i+1}} \int_0^\infty u^i e^{-u} du,$$

where we made the change of variables $u = \theta(s_i + x)$. This is at least a distribution that normalizes, so often one simply assumes the existence of a piece of fake data. For example, by saying we "observe" $x_0 = 1$, we have prior proportional to $\pi(\theta) = e^{-\theta}$, which yields redundancy

$$D_{\text{kl}}\left(P_{\theta_0}^n \| Q_n^\pi\right) = \frac{1}{2} \log \frac{n}{2\pi e} + \theta_0 + \log \frac{1}{\theta_0} + o(1).$$

The difference is that, in this case, the redundancy bound is no longer uniform in θ_0, as it would be for the true reference (or Jeffreys, if it exists) prior. ♣

9.3.2 Redundancy capacity duality

Let us discuss Bayesian redundancy versus worst-case redundancy in somewhat more depth. If we play a game where nature chooses T according to the known prior π, and draws data $X_1^n \sim P_\theta$ conditional on $T = \theta$, then we know that as in expression (9.3.7), we have

$$\inf_Q \mathbb{E}_\pi \left[D_{\text{kl}}\left(P_T^n \| Q\right) \right] = \int D_{\text{kl}}\left(P_\theta^n \| Q_n^\pi\right) \pi(\theta) d\theta = I_\pi(T; X_1^n).$$

A natural question that arises from this expression is the following: if nature chooses a worst-case prior, can we swap the order of maximization and minimization? That is, do we ever have the equality

$$\sup_\pi I_\pi(T; X_1^n) = \inf_Q \sup_\theta D_{\text{kl}}\left(P_\theta^n \| Q\right),$$

so that the worst-case Bayesian redundancy is actually the minimax redundancy? It is clear that if nature can choose the worst case P_θ after we choose Q, the redundancy must be at least as bad as the Bayesian redundancy, so

$$\sup_\pi I_\pi(T; X_1^n) \le \inf_Q \sup_\theta D_{\text{kl}}\left(P_\theta^n \| Q\right) = \inf_Q \mathfrak{R}_n(Q, \mathcal{P}).$$

Indeed, if this inequality were an equality, then for the worst-case prior π^*, the mixture $Q_n^{\pi^*}$ would be minimax optimal.

In fact, the redundancy-capacity theorem, first proved by Gallager [4], and extended by Haussler [6] (among others) allows us to do just that. That is, if we must choose a distribution Q and then nature chooses P_θ adversarially, we can guarantee to worse redundancy than in the (worst-case) Bayesian setting. We state a simpler version of the result that holds when the random variables X take values in finite spaces; Haussler's more general version shows that the next theorem holds whenever $X \in \mathcal{X}$ and $\mathcal{X}$ is a complete separable metric space.

Theorem 9.9 (Gallager [4]). *Let X be a random variable taking on a finite number of values and Θ be a measurable space. Then*

$$\sup_\pi \inf_Q \int D_{\rm kl}\left(P_\theta \| Q\right) d\pi(\theta) = \sup_\pi I_\pi(T; X) = \inf_Q \sup_{\theta \in \Theta} D_{\rm kl}\left(P_\theta \| Q\right).$$

Moreover, the infimum on the right is uniquely attained by some distribution Q^*, and if π^* attains the supremum on the left, then $Q^* = \int P_\theta d\pi^*(\theta)$.

See Section 9.5 for a proof of Theorem 9.9.

This theorem is known as the *redundancy-capacity* theorem in the literature, because in classical information theory, the capacity of a noisy channel $T \to X_1^n$ is the maximal mutual informationx $\sup_\pi I_\pi(T; X_1^n)$. In the exercises, you explore some robustness properties of the optimal distribution Q^π in relation to this theorem. In short, though, we see that if there is a capacity achieving prior, then the associated mixture distribution Q^π is minimax optimal and attains the minimax redundancy for the game.

9.4 Asymptotic normality and Theorem 9.5

In this section, we very briefly (and very hand-wavily) justify the asymptotic expression (9.3.5). To do this, we argue that (roughly) the posterior distribution $\pi(\theta \mid X_1^n)$ should be roughly normally distributed with appropriate variance measure, which gives the result. We now give the intuition for this statement, first by heuristically deriving the asymptotics of a maximum likelihood estimator, then by looking at the Bayesian case. (Clarke and Barron [2] provide a fully rigorous proof.)

9.4.1 Heuristic justification of asymptotic normality

First, we sketch the asymptotic normality of the maximum likelihood estimator $\widehat{\theta}$, that is, $\widehat{\theta}$ is chosen to maximize $\log p_\theta(X_1^n)$. (See, for example, Lehmann and Casella [8] for more rigorous arguments.) Assume that the data are generated i.i.d. according to P_{θ_0}. Then by assumption that $\widehat{\theta}$ maximizes the log-likelihood, we have the stationary condition $0 = \nabla \log p_{\widehat{\theta}}(X_1^n)$. Performing a Taylor expansion of this quantity about θ_0, we have

$$0 = \nabla \log p_{\widehat{\theta}}(X_1^n) = \nabla \log p_{\theta_0}(X_1^n) + \nabla^2 \log p_{\theta_0}(X_1^n)(\widehat{\theta} - \theta_0) + R$$

where R is a remainder term. Assuming that $\widehat{\theta} \to \theta_0$ at any reasonable rate (this can be made rigorous), this remainder is negligible asymptotically.

Rearranging this equality, we obtain

$$\begin{aligned}
\widehat{\theta} - \theta_0 &\approx (-\nabla^2 \log p_{\theta_0}(X_1^n))^{-1} \nabla \log p_{\theta_0}(X_1^n) \\
&= \frac{1}{n}\underbrace{\left(-\frac{1}{n}\sum_{i=1}^n \nabla^2 \log p_{\theta_0}(X_i)\right)}_{\approx I_{\theta_0}}^{-1} \sum_{i=1}^n \nabla \log p_{\theta_0}(X_i) \\
&\approx \frac{1}{n} I_{\theta_0}^{-1} \sum_{i=1}^n \nabla \log p_{\theta_0}(X_i),
\end{aligned}$$

where we have used that the Fisher information $I_\theta = -\mathbb{E}_\theta[\nabla^2 \log p_\theta(X)]$ and the law of large numbers. By the (multivariate) central limit theorem, we then obtain the asymptotic normality result

$$\sqrt{n}(\widehat{\theta} - \theta_0) \approx \frac{1}{\sqrt{n}} I_{\theta_0}^{-1} \sum_{i=1}^n \nabla \log p_{\theta_0}(X_i) \overset{d}{\to} \mathsf{N}(0, I_{\theta_0}^{-1}),$$

where $\overset{d}{\to}$ denotes convergence in distribution, with asymptotic variance

$$I_{\theta_0}^{-1} \mathbb{E}_{\theta_0}[\nabla \log p_{\theta_0}(X) \nabla \log p_{\theta_0}(X)^\top] I_{\theta_0}^{-1} = I_{\theta_0}^{-1} I_{\theta_0} I_{\theta_0}^{-1} = I_{\theta_0}^{-1}.$$

Completely heuristically, we also write

$$\widehat{\theta} \text{ `` } \sim \text{ '' } \mathsf{N}(\theta_0, (nI_{\theta_0})^{-1}). \tag{9.4.1}$$

9.4.2 Heuristic calculations of posterior distributions and redundancy

With the asymptotic distributional heuristic (9.4.1), we now look at the redundancy and posterior distribution of θ conditioned on the data X_1^n when the data are drawn i.i.d. P_{θ_0}. When Q_n^π is the mixture distribution associated with π, the posterior density of $\theta \mid X_1^n$ is

$$\pi(\theta \mid X_1^n) = \frac{p_\theta(X_1^n)\pi(\theta)}{q_n(X_1^n)}.$$

By our heuristic calculation of the MLE, this density (assuming the data overwhelms the prior) is approximately a normal density with mean θ_0 and variance $(nI_{\theta_0})^{-1}$, where we have used expression (9.4.1). Expanding the redundancy, we obtain

$$\mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_1^n)}{q_n(X_1^n)}\right] = \mathbb{E}_{\theta_0}\left[\log \frac{p_{\widehat{\theta}}(X_1^n)\pi(\widehat{\theta})}{q_n(X_1^n)}\right] + \mathbb{E}_{\theta_0}\left[\log \frac{1}{\pi(\widehat{\theta})}\right] + \mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_1^n)}{p_{\widehat{\theta}}(X_1^n)}\right]. \tag{9.4.2}$$

Now we use our heuristic. We have that

$$\mathbb{E}_{\theta_0}\left[\log \frac{p_{\widehat{\theta}}(X_1^n)\pi(\widehat{\theta})}{q_n(X_1^n)}\right] \approx \log \frac{1}{(2\pi)^{d/2} \det(nI_{\theta_0})^{-1/2}} + \mathbb{E}_{\theta_0}\left[-\frac{1}{2}(\widehat{\theta} - \theta_0)^\top (nI_{\theta_0})^{-1} (\widehat{\theta} - \theta_0)\right],$$

by the asymptotic normality result, $\pi(\widehat{\theta}) = \pi(\theta_0) + O(1/\sqrt{n})$ again by the asymptotic normality result, and

$$\begin{aligned}
\log p_{\widehat{\theta}}(X_1^n) &\approx \log p_{\theta_0}(X_1^n) + \left(\sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)\right)^\top (\widehat{\theta} - \theta_0) \\
&\approx \log p_{\theta_0}(X_1^n) + \left(\sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)\right)^\top I_{\theta_0}^{-1} \bigg(\frac{1}{n} \sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)\bigg).
\end{aligned}$$

Substituting these three into the redundancy expression (9.4.2), we obtain

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_1^n)}{q_n(X_1^n)}\right] &\approx \log \frac{1}{(2\pi)^{d/2} \det(nI_{\theta_0})^{-1/2}} + \mathbb{E}_{\theta_0}\left[-\frac{1}{2}(\widehat{\theta} - \theta_0)^\top (nI_{\theta_0})^{-1} (\widehat{\theta} - \theta_0)\right] \\
&\quad + \log \frac{1}{\pi(\theta_0)} - \mathbb{E}_{\theta_0}\left[\left(\sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)\right)^\top I_{\theta_0}^{-1} \bigg(\frac{1}{n} \sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)\bigg)\right] \\
&= \frac{d}{2} \log \frac{n}{2\pi} + \frac{1}{2} \log\det(I_{\theta_0}) + \log \frac{1}{\pi(\theta_0)} - d + R,
\end{aligned}$$

where R is a remainder term. This gives the major terms in the asymptotic result in Theorem 9.5.

9.5 Proof of Theorem 9.9

In this section, we prove one version of the strong saddle point results associated with the universal prediction game as given by Theorem 9.9 (in the case that X belongs to a finite set). For shorthand, we recall the definition of the redundancy

$$\mathsf{Red}(Q, \theta) := \mathbb{E}_{P_\theta}\left[-\log Q(X) + \log P_\theta(X)\right] = D_{\rm kl}\left(P_\theta \| Q\right),$$

where we have assumed that X belongs to a finite set, so that $Q(X)$ is simply the probability of X. For a given prior distribution π on θ, we define the expected redundancy as

$$\mathsf{Red}(Q, \pi) = \int D_{\rm kl}\left(P_\theta \| Q\right) d\pi(\theta).$$

Our goal is to show that the max-min value of the prediction game is the same as the min-max value of the game, that is,

$$\sup_\pi I_\pi(T; X) = \sup_\pi \inf_Q \mathsf{Red}(Q, \pi) = \inf_Q \sup_{\theta \in \Theta} \mathsf{Red}(Q, \theta).$$

Proof We know that the max-min risk (worst-case Bayes risk) of the game is $\sup_\pi I_\pi(T; X)$; it remains to show that this is the min-max risk. To that end, define the *capacity* of the family $\{P_\theta\}_{\theta \in \Theta}$ as

$$C := \sup_\pi I_\pi(T; X). \tag{9.5.1}$$

Notably, this constant is finite (because $I_\pi(T; X) \le \log |\mathcal{X}|$), and there exists a sequence π_n of prior probabilities such that $I_{\pi_n}(T; X) \to C$. Now, let $\bar{Q}$ be any cluster point of the sequence of mixtures $Q^{\pi_n} = \int P_\theta d\pi_n(\theta)$; such a point exists because the space of probability distributions on the finite set $\mathcal{X}$ is compact. We will show that

$$\sum_x P_\theta(x) \log \frac{P_\theta(x)}{\bar{Q}(x)} \le C \quad \text{for all } \theta \in \Theta, \tag{9.5.2}$$

and we claim this is sufficient for the theorem. Indeed, suppose that inequality (9.5.2) holds. Then in this case, we have

$$\inf_Q \sup_{\theta \in \Theta} \mathsf{Red}(Q, \theta) \le \sup_{\theta \in \Theta} \mathsf{Red}(\bar{Q}, \theta) = \sup_{\theta \in \Theta} D_{\rm kl}\left(P_\theta \| \bar{Q}\right) \le C,$$

which implies the theorem, because it is always the case that

$$\sup_\pi \inf_Q \mathsf{Red}(Q, \theta) \le \inf_Q \sup_\pi \mathsf{Red}(Q, \pi) = \inf_Q \sup_{\theta \in \Theta} \mathsf{Red}(Q, \theta).$$

For the sake of contradiction, let us assume that there exists some $\theta \in \Theta$ such that inequality (9.5.2) fails, call it θ^*. We will then show that suitable mixtures $(1 - \lambda)\pi + \lambda\delta_{\theta^*}$, where δ_{θ^*} is the point mass on θ^*, could increase the capacity (9.5.1). To that end, for shorthand define the mixtures

$$\pi_{n,\lambda} = (1 - \lambda)\pi_n + \lambda\delta_{\theta^*} \quad \text{and} \quad Q^{\pi_{n,\lambda}} = (1 - \lambda)Q^{\pi_n} + \lambda P_{\theta^*}$$

for $\lambda \in [0, 1]$. Let us also use the notation $H_w(X \mid T)$ to denote the conditionaly entropy of the random variable X on T (when T is distributed as w), and we abuse notation by writing $H(X) = H(P)$ when X is distributed as P. In this case, it is clear that we have

$$H_{\pi_{n,\lambda}}(X \mid T) = (1 - \lambda)H_{\pi_n}(X \mid T) + \lambda H(X \mid T = \theta^*),$$

and by definition of the mutual information we have

$$\begin{aligned}
I_{\pi_{n,\lambda}}(T; X) &= H_{\pi_{n,\lambda}}(X) - H_{\pi_{n,\lambda}}(X \mid T) \\
&= H((1-\lambda)Q^{\pi_n} + \lambda P_{\theta^*}) - (1-\lambda)H_{\pi_n}(X \mid T) - \lambda H(X \mid T = \theta^*).
\end{aligned}$$

To demonstrate our contradiction, we will show two things: first, that at $\lambda = 0$ the limits of both sides of the preceding display are equal to the capacity C, and second, that the derivative of the right hand side is positive. This will contradict the definition (9.5.1) of the capacity.

To that end, note that

$$\lim_n H_{\pi_n}(X \mid T) = \lim_n H_{\pi_n}(X) - I_{\pi_n}(T; X) = H(\bar{Q}) - C,$$

by the continuity of the entropy function. Thus, we have

$$\lim_n I_{\pi_n,\lambda}(T; X) = H((1-\lambda)\bar{Q} + \lambda P_{\theta^*}) - (1-\lambda)(H(\bar{Q}) - C) - \lambda H(P_\theta). \tag{9.5.3}$$

It is clear that at $\lambda = 0$, both sides are equal to the capacity C, while taking derivatives with respect to λ we have

$$\frac{\partial}{\partial \lambda} H((1-\lambda)\bar{Q} + \lambda P_{\theta^*}) = -\sum_x (P_{\theta^*}(x) - \bar{Q}(x)) \log\left((1-\lambda)\bar{Q}(x) + \lambda P_{\theta^*}(x)\right).$$

Evaluating this derivative at $\lambda = 0$, we find

$$\begin{aligned}
&\frac{\partial}{\partial \lambda} \lim_n I_{\pi_n,\lambda}(T; X)\bigg|_{\lambda=0} \\
&\quad= -\sum_x P_{\theta^*}(x) \log \bar{Q}(x) + \sum_x \bar{Q}(x) \log \bar{Q}(x) + H(\bar{Q}) - C + \sum_x P_{\theta^*}(x) \log P_{\theta^*}(x) \\
&\quad= \sum_x P_{\theta^*}(x) \log \frac{P_{\theta^*}(x)}{\bar{Q}(x)} - C.
\end{aligned}$$

In particular, if inequality (9.5.2) fails to hold, then $\frac{\partial}{\partial \lambda} \lim_n I_{\pi_n,\lambda}(T; X)|_{\lambda=0} > 0$, contradicting the definition (9.5.1) of the channel capacity.

The uniqueness of the result follows from the strict convexity of the mutual information I in the mixture channel $\bar{Q}$. □

Bibliography

[1] J. M. Bernardo. Reference analysis. In D. Day and C. R. Rao, editors, *Bayesian Thinking, Modeling and Computation*, volume 25 of *Handbook of Statistics*, chapter 2, pages 17–90. Elsevier, 2005.

[2] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[4] R. Gallager. Source coding with side information and universal coding. Technical Report LIDS-P-937, MIT Laboratory for Information and Decision Systems, 1979.

[5] P. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[6] D. Haussler. A general minimax result for relative entropy. *IEEE Transactions on Information Theory*, 43(4):1276–1280, 1997.

[7] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A: Mathematical and Physical Sciences*, 186:453–461, 1946.

[8] E. L. Lehmann and G. Casella. *Theory of Point Estimation, Second Edition*. Springer, 1998.

[9] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44 (6):2124–2147, 1998.

Chapter 10

Universal prediction with other losses

Thus far, in our discussion of universal prediction and related ideas, we have focused (essentially) exclusively on making predictions with the logarithmic loss, so that we play a full distribution over the set $\mathcal{X}$ as our prediction at each time step in the procedure. This is natural in settings, such as coding (recall examples 7.5 and 9.1), in which the log loss corresponds to a quantity we directly care about, or when we do not necessarily know much about the task at hand but rather wish to simply model a process. (We will see this more shortly.) In many cases, however, we have a natural task-specific loss. The natural question that follows, then, is to what extent it is possible to extend the results of Chapter 9 to different settings in which we do not necessarily care about prediction of an entire distribution. (Relevant references include the paper of Cesa-Bianchi and Lugosi [3], which shows how complexity measures known as Rademacher complexity govern the regret in online prediction games; the book by the same authors [4], which gives results covering a wide variety of online learning, prediction, and other games; the survey by Merhav and Feder [12]; and the study of consequences of the choice of loss for universal prediction problems by Haussler et al. [7].)

10.1 Redudancy and expected regret

We begin by considering a generalization of the redundancy (9.1.3) to the case in which we do not use the log loss. In particular, we have as usual a space $\mathcal{X}$ and a loss function $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $\ell(\widehat{x}, x)$ is the penalty we suffer for playing $\widehat{x}$ when the instantaneous data is x. (In somewhat more generality, we may allow the loss to act on $\widehat{\mathcal{X}} \times \mathcal{X}$, where the prediction space $\widehat{\mathcal{X}}$ may be different from $\mathcal{X}$.) As a simple example, consider a weather prediction problem, where $X_i \in \{0, 1\}$ indicates whether it rained on day i and $\widehat{X}_i$ denotes our prediction of whether it will rain. Then a natural loss includes $\ell(\widehat{x}, x) = \mathbf{1}\{\widehat{x} \cdot x \le 0\}$, which simply counts the number of mistaken predictions.

Given the loss ℓ, our goal is to minimize the expected cumulative loss

$$\sum_{i=1}^{n} \mathbb{E}_P[\ell(\widehat{X}_i, X_i)],$$

where $\widehat{X}_i$ are the predictions of the procedure we use and P is the distribution generating the data X_1^n. In this case, if the distribution P is known, it is clear that the optimal strategy is to play the Bayes-optimal prediction

$$X_i^* \in \operatorname*{argmin}_{x \in \widehat{\mathcal{X}}} \mathbb{E}_P[\ell(x, X_i) \mid X_1^{i-1}] = \operatorname*{argmin}_{x \in \widehat{\mathcal{X}}} \int_{\mathcal{X}} \ell(x, x_i) dP(x_i \mid X_1^{i-1}). \tag{10.1.1}$$

In many cases, however, we do not know the distribution P, and so our goal (as in the previous chapter) is to simultaneously minimize the cumulative loss simultaneously for all source distributions in a family $\mathcal{P}$.

10.1.1 Universal prediction via the log loss

As our first idea, we adapt the same strategies as those in the previous section, using a distribution Q that has redundancy growing only sub-linearly against the class $\mathcal{P}$, and making Bayes optimal predictions with Q. That is, at iteration i, we assume that $X_i \sim Q(\cdot \mid X_1^{i-1})$ and play

$$\widehat{X}_i \in \operatorname*{argmin}_{x \in \widehat{\mathcal{X}}} \mathbb{E}_Q[\ell(x, X_i) \mid X_1^{i-1}] = \int_{\mathcal{X}} \ell(x, x_i) dQ(x_i \mid X_1^{i-1}). \tag{10.1.2}$$

Given such a distribution Q, we measure its loss-based redundancy against P via

$$\mathsf{Red}_n(Q, P, \ell) := \mathbb{E}_P\left[\sum_{i=1}^n \ell(\widehat{X}_i, X_i) - \sum_{i=1}^n \ell(X_i^*, X_i)\right], \tag{10.1.3}$$

where $\widehat{X}_i$ chosen according to $Q(\cdot \mid X_1^{i-1})$ as in expression (10.1.2). The natural question now, of course, is whether the strategy (10.1.2) has redundancy growing more slowly than n.

It turns out that in some situations, this is the case: we have the following theorem [12, Section III.A.2], which only requires that the usual redundancy (9.1.3) (with log loss) is sub-linear and the loss is suitably bounded. In the theorem, we assume that the class of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ is indexed by $\theta \in \Theta$.

Theorem 10.1. *Assume that the redundancy $\mathsf{Red}_n(Q, P_\theta) \le R_n(\theta)$ and that $|\ell(\widehat{x}, x) - \ell(x^*, x)| \le L$ for all x and predictions $\widehat{x}, x^*$. Then we have*

$$\frac{1}{n}\mathsf{Red}_n(Q, P_\theta, \ell) \le L\sqrt{\frac{2}{n}R_n(\theta)}.$$

To attain vanishing expected regret under the loss ℓ, then, Theorem 10.1 requires only that we play a Bayes' strategy (10.1.2) with a distribution Q for which the average (over n) of the usual redundancy (9.1.3) tends to zero, so long as the loss is (roughly) bounded. We give two examples of bounded losses. First, we might consider the 0-1 loss, which clearly satisfies $|\ell(\widehat{x}, x) - \ell(x^*, x)| \le 1$. Second, the absolute value loss (which is used for robust estimation of location parameters [14, 9]), given by $\ell(\widehat{x}, x) = |x - \widehat{x}|$, satisfies $|\ell(\widehat{x}, x) - \ell(x^*, x)| \le |\widehat{x} - x^*|$. If the distribution P_θ has median θ and Θ is compact, then $\mathbb{E}[|\widehat{x} - X|]$ is minimized by its median, and $|\widehat{x} - x^*|$ is bounded by the diameter of Θ.

Proof The theorem is essentially a consequence of Pinsker's inequality (Proposition 2.10). By expanding the loss-based redundancy, we have the following chain of equalities:

$$\begin{aligned}
\mathsf{Red}_n(Q, P_\theta, \ell) &= \sum_{i=1}^n \mathbb{E}_\theta[\ell(\widehat{X}_i, X_i)] - \mathbb{E}_\theta[\ell(X_i^*, X_i)] \\
&= \sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \int_{\mathcal{X}} p_\theta(x_i \mid x_1^{i-1}) \left[\ell(\widehat{X}_i, x_i) - \ell(X_i^*, x_i)\right] dx_i dx_1^{i-1} \\
&= \sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \int_{\mathcal{X}} (p_\theta(x_i \mid x_1^{i-1}) - q(x_i \mid x_1^{i-1})) \left[\ell(\widehat{X}_i, x_i) - \ell(X_i^*, x_i)\right] dx_i dx_1^{i-1} \\
&\quad + \sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \underbrace{\mathbb{E}_Q[\ell(\widehat{X}_i, X_i) - \ell(X_i^*, X_i) \mid x_1^{i-1}]}_{\le 0} dx_1^{i-1}, \qquad (10.1.4)
\end{aligned}$$

where for the inequality we used that the play $\widehat{X}_i$ minimizes

$$\mathbb{E}_Q[\ell(\widehat{X}_i, X_i) - \ell(X_i^*, X_i) \mid X_1^{i-1}]$$

by the construction (10.1.2).

Now, using Hölder's inequality on the innermost integral in the first sum of expression (10.1.4), we have

$$\begin{aligned}
&\int_{\mathcal{X}} (p_\theta(x_i \mid x_1^{i-1}) - q(x_i \mid x_1^{i-1})) \left[\ell(\widehat{X}_i, x_i) - \ell(X_i^*, x_i)\right] dx_i \\
&\qquad \le 2 \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\rm TV} \sup_{x \in \mathcal{X}} |\ell(\widehat{X}_i, x) - \ell(X_i^*, x)| \\
&\qquad \le 2L \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\rm TV},
\end{aligned}$$

where we have used the definition of total variation distance. Combining this inequality with (10.1.4), we obtain

$$\begin{aligned}
\mathsf{Red}_n(Q, P_\theta, \ell) &\le 2L \sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\rm TV} dx_1^{i-1} \\
&\overset{(\star)}{\le} 2L \sum_{i=1}^n \left(\int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) dx_1^{i-1}\right)^{\frac{1}{2}} \left(\int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\rm TV}^2\right)^{\frac{1}{2}} \\
&= 2L \sum_{i=1}^n \left(\int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\rm TV}^2\right)^{\frac{1}{2}},
\end{aligned}$$

where the inequality $(\star)$ follows by the Cauchy-Schwarz inequality applied to the integrands $\sqrt{p_\theta}$ and $\sqrt{p_\theta} \left\|P - Q\right\|_{\rm TV}$. Applying the Cauchy-Schwarz inequality to the final sum, we have

$$\begin{aligned}
\mathsf{Red}_n(Q, P_\theta, \ell) &\le 2L\sqrt{n} \left(\sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) \left\|P_\theta(\cdot \mid x_1^{i-1}) - Q(\cdot \mid x_1^{i-1})\right\|_{\rm TV}^2\right)^{\frac{1}{2}} \\
&\overset{(\star\star)}{\le} 2L\sqrt{n} \left(\frac{1}{2} \sum_{i=1}^n \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) D_{\rm kl}\left(P_\theta(\cdot \mid x_1^{i-1}) \| Q(\cdot \mid x_1^{i-1})\right) dx_1^{i-1}\right)^{\frac{1}{2}} \\
&= L\sqrt{2n} \sqrt{D_{\rm kl}\left(P_\theta^n \| Q\right)},
\end{aligned}$$

where inequality $(\star\star)$ is an application of Pinsker's inequality. But of course, we know by that $\mathsf{Red}_n(Q, P_\theta) = D_{\rm kl}\left(P_\theta^n \| Q\right)$ by definition (9.1.3) of the redundancy. □

Before proceding to examples, we note that in a variety of cases the bounds of Theorem 10.1 are loose. For example, under mean-squared error, universal linear predictors [6, 15] have redundancy $\mathcal{O}(\log n)$, while Theorem 10.1 gives at best a bound of $\mathcal{O}(\sqrt{n})$.

TODO: Add material on redundancy/capacity (Theorem 9.9) analogue in general loss case, which allows playing mixture distributions based on mixture of $\{P_\theta\}_{\theta\in\Theta}$.

10.1.2 Examples

We now give an example application of Theorem 10.1 with an application to a classification problem with side information. In particular, let us consider the 0-1 loss $\ell_{0-1}(\hat{y}, y) = \mathbf{1}\{\hat{y} \cdot y \le 0\}$, and assume that we wish to predict y based on a vector $x \in \mathbb{R}^d$ of regressors that are fixed ahead of time. In addition, we assume that the "true" distribution (or competitor) P_θ is that given x and θ, Y has normal distribution with mean $\langle \theta, x\rangle$ and variance σ^2, that is,

$$Y_i = \langle \theta, x_i\rangle + \varepsilon_i, \quad \varepsilon_i \stackrel{\rm i.i.d.}{\sim} \mathsf{N}(0, \sigma^2).$$

Now, we consider playing according to a mixture distribution (9.3.3), and for our prior π we choose $\theta \sim \mathsf{N}(0, \tau^2 I_{d\times d})$, where $\tau > 0$ is some parameter we choose.

Let us first consider the case in which we observe $Y_1, \ldots, Y_n$ directly (rather than simply whether we classify correctly) and consider the prediction scheme this generates. First, we recall as in the posterior calculation (9.3.4) that we must calculate the posterior on θ given $Y_1, \ldots, Y_i$ at step $i+1$. Assuming we have computed this posterior, we play

$$\begin{aligned}\widehat{Y}_i := \mathop{\rm argmin}_{y\in\mathbb{R}} \mathbb{E}_{Q^\pi}[\ell_{0-1}(y, Y_i) \mid Y_1^{i-1}] &= \mathop{\rm argmin}_{y\in\mathbb{R}} Q^\pi(\mathrm{sign}(Y_i) \neq \mathrm{sign}(y) \mid Y_1^{i-1}) \\ &= \mathop{\rm argmin}_{y\in\mathbb{R}} \int_{-\infty}^{\infty} P_\theta(\mathrm{sign}(Y_i) \neq \mathrm{sign}(y)) \pi(\theta \mid Y_1^{i-1}) d\theta. \quad (10.1.5)\end{aligned}$$

With this in mind, we begin by computing the posterior distribution on θ:

Lemma 10.2. *Assume that θ has prior $\mathsf{N}(0, \tau^2 I_{d\times d})$. Then conditional on $Y_1^i = y_1^i$ and the first i vectors $x_1^i = (x_1, \ldots, x_i) \subset \mathbb{R}^d$, we have*

$$\theta \mid y_1^i, x_1^i \sim \mathsf{N}\left(K_i^{-1}\sum_{j=1}^i x_j y_j, K_i^{-1}\right), \quad \textit{where} \quad K_i = \frac{1}{\tau^2} I_{d\times d} + \frac{1}{\sigma^2}\sum_{j=1}^i x_j x_j^\top.$$

Deferring the proof of Lemma 10.2 temporarily, we note that under the distribution Q^π, as by assumption we have $Y_i = \langle \theta, x_i\rangle + \varepsilon_i$, the posterior distribution (under the prior π for θ) on Y_{i+1} conditional on $Y_1^i = y_i^i$ and $x_1, \ldots, x_{i+1}$ is

$$Y_{i+1} = \langle \theta, x_{i+1}\rangle + \varepsilon_{i+1} \mid y_1^i, x_i^1 \sim \mathsf{N}\Big(\Big\langle x_{i+1}, K_i^{-1}\sum_{j=1}^i x_j y_j\Big\rangle, x_{i+1}^\top K_i^{-1} x_{i+1} + \sigma^2\Big).$$

Consequently, if we let $\widehat{\theta}_{i+1}$ be the posterior mean of $\theta \mid y_1^i, x_i^i$ (as given by Lemma 10.2), the optimal prediction (10.1.5) is to choose any $\widehat{Y}_{i+1}$ satisfying $\mathrm{sign}(\widehat{Y}_{i+1}) = \mathrm{sign}(\langle x_{i+1}, \widehat{\theta}_{i+1}\rangle)$. Another option is to simply play

$$\widehat{Y}_{i+1} = x_{i+1}^\top K_i^{-1} \bigg(\sum_{j=1}^i y_j x_j \bigg), \tag{10.1.6}$$

which is $\mathbb{E}[\widehat{Y}_{i+1} \mid Y_1^i, X_1^{i+1}] = \mathbb{E}[\langle \theta, X_{i+1}\rangle \mid Y_1^i, X_1^i]$, because this $\widehat{Y}_{i+1}$ has sign that is most probable for Y_{i+1} (under the mixture Q^π).

Let us now evaluate the 0-1 redundancy of the prediction scheme (10.1.6). We first compute the Fisher information for the distribution $Y_i \sim \mathsf{N}(\langle \theta, x_i\rangle, \sigma^2)$. By a straightforward calculation, we have $I_\theta = \frac{1}{\sigma^2} X^\top X$, where the matrix $X \in \mathbb{R}^{n\times d}$ is the data matrix $X = [x_1 \ \cdots \ x_n]^\top$. Then for any $\theta_0 \in \mathbb{R}^d$, Theorem 9.5 implies that for the prior $\pi(\theta) = \frac{1}{(2\pi\tau^2)^{d/2}} \exp(-\frac{1}{2\tau^2} \left\|\theta\right\|_2^2)$, we have (up to constant factors) the redundancy bound

$$\mathsf{Red}_n(Q^\pi, P_{\theta_0}) \lesssim d \log n + d \log \tau + \frac{1}{\tau^2} \left\|\theta_0\right\|_2^2 + \log\det(\sigma^{-2} X^\top X).$$

Thus the expected regret under the 0-1 loss ℓ_{0-1} is

$$\mathsf{Red}_n(Q^\pi, P_{\theta_0}, \ell_{0-1}) \lesssim \sqrt{n} \sqrt{d \log n + d \log(\sigma\tau) + \frac{1}{\tau^2} \left\|\theta_0\right\|_2^2 + \log\det(X^\top X)} \tag{10.1.7}$$

by Theorem 10.1. We can provide some intuition for this expected regret bound. First, for any θ_0, we can asymptotically attain vanishing expected regret, though larger θ_0 require more information to identify. In addition, the less informative the prior is (by taking $\tau \uparrow +\infty$), the less we suffer by being universal to all θ_0, but there is logarithmic penalty in τ. We also note that the bound (10.1.7) is not strongly universal, because by taking $\left\|\theta_0\right\| \to \infty$ we can make the bound vacuous.

We remark in passing that we can play a similar game when all we observe are truncated (signed) normal random variables, that is, we see only $\mathrm{sign}(Y_i)$ rather than Y_i. Unfortunately, in this case, there is no closed form for the posterior updates as in Lemma 10.2. That said, it is possible to play the game using sampling (Monte Carlo) or other strategies.

Finally, we prove Lemma 10.2:

Proof We use Bayes rule, ignoring normalizing constants that do not depend on θ. In this case, we have the posterior distribution proportional to the prior times the likelihood, so

$$\pi(\theta \mid y_1^i, x_1^i) \propto \pi(\theta) \prod_{i=1}^n p_\theta(y_i \mid x_i) \propto \exp\left(-\frac{1}{2\tau^2} \left\|\theta\right\|_2^2 - \frac{1}{2\sigma^2} \sum_{j=1}^i (y_j - \langle x_j, \theta\rangle)^2 \right).$$

Now, we complete the square in the exponent above, which yields

$$\begin{aligned}\frac{1}{2\tau^2} \left\|\theta\right\|_2^2 + \frac{1}{2\sigma^2} \sum_{j=1}^i (y_j - \langle x_j, \theta\rangle)^2 &= \frac{1}{2} \theta^\top \left(\frac{1}{\tau^2} I_{d\times d} + \frac{1}{\sigma^2} \sum_{j=1}^i x_j x_j^\top \right) \theta - \theta^\top \sum_{j=1}^i y_j x_j + C \\ &= \frac{1}{2} \left(\theta - K_i^{-1} \sum_{j=1}^i y_j x_j \right)^\top K_i \left(\theta - K_i^{-1} \sum_{j=1}^i y_j x_j \right) + C',\end{aligned}$$

where C, C' are constants depending only on the y_1^i and not x_1^i or θ, and we have recalled the definition of $K_i = \tau^{-2} I_{d\times d} + \sigma^{-2} \sum_{j=1}^i x_j x_j^\top$. By inspection, this implies our desired result. □

10.2 Individual sequence prediction and regret

Having discussed (in some minor detail) prediction games under more general losses in an expected sense, we now consider the more adversarial sense of Section 9.2, where we wish to compete against a family of prediction strategies and the data sequence observed is chosen adversarially. In this section, we look into the case in which the comparison class—set of strategies against which we wish to compete—is finite.

As a first observation, in the redundancy setting, we see that when the class $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ has $|\Theta| < \infty$, then the redundancy capacity theorem (Theorem 9.9) implies that

$$\inf_Q \sup_{\theta \in \Theta} \mathsf{Red}_n(Q, P_\theta) = \inf_Q \sup_{\theta \in \Theta} D_{\rm kl}\left(P_\theta^n \| Q\right) = \sup_\pi I_\pi(T; X_1^n) \le \log |\Theta|,$$

where $T \sim \pi$ and conditioned on $T = \theta$ we draw $X_1^n \sim P_\theta$. (Here we have used that $I(T; X_1^n) = H(T) - H(T \mid X_1^n) \le H(T) \le \log|\Theta|$, by definition (2.1.3) of the mutual information.) In particular, the redundancy is *constant* for any n.

Now we come to our question: is this possible in a purely sequential case? More precisely, suppose we wish to predict a sequence of variables $y_i \in \{-1, 1\}$, we have access to a finite collection of strategies, and we would like to guarantee that we perform as well in prediction as any single member of this class. Then, while it is not possible to achieve constant regret, it is possible to have regret that grows only logarithmically in the number of comparison strategies. To establish the setting, let us denote our collection of strategies, henceforth called "experts", by $\{x_{i,j}\}_{j=1}^d$, where i ranges in $1, \ldots, n$. Then at iteration i of the prediction game, we measure the loss of expert j by $\ell(x_{i,j}, y)$.

We begin by considering a mixture strategy that would be natural under the logarithmic loss, we assume the experts play points $x_{i,j} \in [0,1]$, where $x_{i,j} = P(Y_i = 1)$ according to expert j. (We remark in passing that while the notation is perhaps not completely explicit about this, the experts may adapt to the sequence Y_1^n.) In this case, the loss we suffer is the usual log loss, $\ell(x_{i,j}, y) = y \log \frac{1}{x_{i,j}} + (1-y) \log \frac{1}{1 - x_{i,j}}$. Now, if we assume we begin with the uniform prior distribution $\pi(j) = 1/d$ for all j, then the posterior distribution, denoted by $\pi_j^i = \pi(j \mid Y_1^{i-1})$, is

$$\begin{aligned}\pi_j^i \propto \pi(j) \prod_{l=1}^i x_{l,j}^{y_l} (1 - x_{l,j})^{1 - y_l} &= \pi(j) \exp\left(-\sum_{l=1}^i \left[y_l \log \frac{1}{x_{l,j}} + (1 - y_l) \log \frac{1}{1 - x_{l,j}}\right]\right) \\ &= \pi(j) \exp\left(-\sum_{l=1}^i \ell(x_{l,j}, y_l)\right).\end{aligned}$$

This strategy suggests what is known variously as the *multiplicative weights* strategy [1], exponentiated gradient descent method [10], or (after some massaging) a method known since the late 1970s as the mirror descent or non-Euclidean gradient descent method (entropic gradient descent) [13, 2].

In particular, we consider an algorithm for general losses where fix a stepsize $\eta > 0$ (as we cannot be as aggressive as in the probabilistic setting), and we then weight each of the experts j by exponentially decaying the weight assigned to the expert for the losses it has suffered. For the algorithm to work, unfortunately, we need a technical condition on the loss function and experts $x_{i,j}$. This loss function is analogous to a weakened version of exp-concavity, which is a common assumption in online game playing scenarios (see the logarithmic regret algorithms developed by Hazan et al. [8], as well as earlier work, for example, that by Kivinen and Warmuth [11] studying regression problems for which the loss is strongly convex in one variable but not simultaneously in all). In particular, exp-concavity is the assumption that

$$x \mapsto \exp(-\ell(x, y))$$

is a concave function. Because the exponent of the logarithm is linear, the log loss is obviously exp-concave, but for alternate losses, we make a slightly weaker assumption. In particular, we assume there are constants c, η such that for any vector π in the d-simplex (i.e. $\pi \in \mathbb{R}_+^d$ satisfies $\sum_{j=1}^d \pi_j = 1$) there is some way to choose $\widehat{y}$ so that for any y (that can be played in the game)

$$\exp\left(-\frac{1}{c}\ell(\widehat{y}, y)\right) \ge \sum_{j=1}^d \pi_j \exp(-\eta \ell(x_{i,j}, y)) \quad \text{or} \quad \ell(\widehat{y}, y) \le -c \log\left(\sum_{j=1}^d \pi_j \exp(-\eta \ell(x_{i,j}, y))\right). \tag{10.2.1}$$

By inspection, inequality (10.2.1) holds for the log loss with $c = \eta = 1$ and the choice $\widehat{y} = \sum_{j=1}^d \pi_j x_{i,j}$, because of the exp-concavity condition; any exp-concave loss also satisfies inequality (10.2.1) with $c = \eta = 1$ and the choice of the posterior mean $\widehat{y} = \sum_{j=1}^d \pi_j x_{i,j}$. The idea in this case is that losses satisfying inequality (10.2.1) behave enough like the logarithmic loss that a Bayesian updating of the experts works. (Condition (10.2.1) originates with the work of Haussler et al. [7], where they name such losses (c, η)-realizable.)

Example 10.3 (Squared error and exp-concavity)**:** Consider the squared error loss $\ell(\widehat{y}, y) = \frac{1}{2}(\widehat{y} - y)^2$, where $\widehat{y}, y \in \mathbb{R}$. We claim that if $x_j \in [0, 1]$ for each j, π is in the simplex, meaning $\sum_j \pi_j = 1$ and $\pi_j \ge 0$, and $y \in [0, 1]$, then the squared error $\pi \mapsto \ell(\langle \pi, x \rangle, y)$ is exp-concave, that is, inequality (10.2.1) holds with $c = \eta = 1$ and $\widehat{y} = \langle \pi, x \rangle$. Indeed, computing the Hessian of the exponent, we have

$$\begin{aligned}\nabla_\pi^2 \exp\left(-\frac{1}{2}(\langle \pi, x \rangle - y)^2\right) &= \nabla_\pi \left[-\exp\left(-\frac{1}{2}(\langle \pi, x \rangle - y)^2\right)(\langle \pi, x \rangle - y)x\right] \\ &= \exp\left(-\frac{1}{2}(\langle \pi, x \rangle - y)^2\right)\left((\langle \pi, x \rangle - y)^2 - 1\right) xx^\top.\end{aligned}$$

Noting that $|\langle \pi, x \rangle - y| \le 1$ yields that $(\langle \pi, x \rangle - y)^2 - 1 \le 0$, so we have

$$\nabla_\pi^2 \exp\left(-\frac{1}{2}(\langle \pi, x \rangle - y)^2\right) \preceq 0_{d \times d}$$

under the setting of the example. We thus have exp-concavity as desired. ♣

We can also show that the 0-1 loss satisfies the weakened version of exp-concavity in inequality (10.2.1), but we have to take the constant c to be larger (or η to be smaller).

Example 10.4 (Zero-one loss and weak exp-concavity)**:** Now suppose that we use the 0-1 loss, that is, $\ell_{0-1}(\widehat{y}, y) = \mathbf{1}\{y \cdot \widehat{y} \le 0\}$. We claim that if we take a weighted majority vote under the distribution π, meaning that we set $\widehat{y} = \sum_{j=1}^d \pi_j \operatorname{sign}(x_j)$ for a vector $x \in \mathbb{R}^d$, then inequality (10.2.1) holds with any c large enough that

$$c^{-1} \le \log \frac{2}{1 + e^{-\eta}}. \tag{10.2.2}$$

Demonstrating inequality (10.2.2) is, by inspection, equivalent to showing that

$$\ell_{0-1}(\widehat{y}, y) \le -c \log\left(\sum_{j=1}^d \pi_j e^{-\eta \ell_{0-1}(x_j, y)}\right).$$

If $\widehat{y}$ has the correct sign, meaning that $\mathrm{sign}(\widehat{y}) = \mathrm{sign}(y)$, the result is trivial. If $\mathrm{sign}(\widehat{y})$ is not equal to $\mathrm{sign}(y) \in \{-1, 1\}$, then we know at least (by the weights π_j) half of the values x_j have incorrect sign. Thus

$$\sum_{j=1}^d \pi_j e^{-\eta \ell_{0-1}(x_j, y)} = \sum_{j : x_j y \le 0} \pi_j e^{-\eta} + \sum_{j : x_j y > 0} \pi_j \le \frac{1}{2} e^{-\eta} + \frac{1}{2}.$$

Thus, to attain

$$\ell_{0-1}(\widehat{y}, y) = 1 \le -c \log\left(\sum_{j=1}^d \pi_j e^{-\eta \ell_{0-1}(x_j, y)}\right)$$

it is sufficient that

$$1 \le -c \log\left(\frac{1 + e^{-\eta}}{2}\right) \le -c \log\left(\sum_{j=1}^d \pi_j e^{-\eta \ell_{0-1}(x_j, y)}\right), \quad \text{or} \quad c^{-1} \le \log\left(\frac{2}{1 + e^{-\eta}}\right).$$

This is our desired claim (10.2.2). ♣

Having given general conditions and our motivation of exponential weighting scheme in the case of the logarithmic loss, we arrive at our algorithm. We simply weight the experts by exponentially decaying the losses they suffer. We begin the procedure by initializing a weight vector $w \in \mathbb{R}^d$ with $w_j = 1$ for $j = 1, \ldots, d$. After this, we repeat the following four steps at each time i, beginning with $i = 1$:

1. Set $w_j^i = \exp\left(-\eta \sum_{l=1}^{i-1} \ell(x_{l,j}, y_l)\right)$
2. Set $W^i = \sum_{j=1}^d w_j^i$ and $\pi_j^i = w_j^i / W^i$ for each $j \in \{1, \ldots, d\}$
3. Choose $\widehat{y}_i$ satisfying (10.2.1) for the weighting $\pi = \pi^i$ and expert values $\{x_{i,j}\}_{j=1}^d$
4. Observe y_i and suffer loss $\ell(\widehat{y}_i, y_i)$

With the scheme above, we have the following regret bound.

Theorem 10.5 (Haussler et al. [7]). *Assume condition* (10.2.1) *holds and that $\widehat{y}_i$ is chosen by the above scheme. Then for any $j \in \{1, \ldots, d\}$ and any sequence $y_1^n \in \mathbb{R}^n$,*

$$\sum_{i=1}^n \ell(\widehat{y}_i, y_i) \le c \log d + c\eta \sum_{i=1}^n \ell(x_{i,j}, y_i).$$

Proof This is an argument based on potentials. At each iteration, any loss we suffer implies that the potential W^i must decrease, but it cannot decrease too quickly (as otherwise the individual predictors $x_{i,j}$ would suffer too much loss). Beginning with condition (10.2.1), we observe that

$$\ell(\widehat{y}_i, y_i) \le -c \log\left(\sum_{j=1}^d \pi_j^i \exp(-\eta \ell(x_{i,j}, y_i))\right) = -c \log\left(\frac{W^{i+1}}{W^i}\right)$$

Summing this inequality from $i = 1$ to n and using that $W^1 = d$, we have

$$\begin{aligned}\sum_{i=1}^n \ell(\widehat{y}_i, y_i) \le -c \log\left(\frac{W^{n+1}}{W^1}\right) &= c \log d - c \log\left(\sum_{j=1}^d \exp\left(-\eta \sum_{i=1}^n \ell(x_{i,j}, y_i)\right)\right) \\ &\le c \log d - c \log \exp\left(-\eta \sum_{i=1}^n \ell(x_{i,j}, y_i)\right),\end{aligned}$$

where the inequality uses that $\exp(\cdot)$ is increasing. As $\log \exp(a) = a$, this is the desired result. □

We illustrate the theorem by continuing Example 10.4, showing how Theorem 10.5 gives a regret guarantee of at most $\sqrt{n \log d}$ for any set of at most d experts and any sequence $y_1^n \in \mathbb{R}^n$ under the zero-one loss.

Example (Example 10.4 continued)**:** By substituting the choice $c^{-1} = \log \frac{2}{1+e^{-\eta}}$ into the regret guarantee of Theorem 10.5 (which satisfies inequality (10.2.1) by our guarantee (10.2.2) from Example 10.4), we obtain

$$\sum_{i=1}^n \ell_{0-1}(\widehat{y}_i, y_i) - \ell_{0-1}(x_{i,j}, y_i) \le \frac{\log d}{\log \frac{2}{1+e^{-\eta}}} + \frac{\left(\eta - \log \frac{2}{1+e^{-\eta}}\right) \sum_{i=1}^n \ell_{0-1}(x_{i,j}, y_i)}{\log \frac{2}{1+e^{-\eta}}}.$$

Now, we make an asymptotic expansion to give the basic flavor of the result (this can be made rigorous, but it is sufficient). First, we note that

$$\log \frac{2}{1+e^{-\eta}} \approx \frac{\eta}{2} - \frac{\eta^2}{8},$$

and substituting this into the previous display, we have regret guarantee

$$\sum_{i=1}^n \ell_{0-1}(\widehat{y}_i, y_i) - \ell_{0-1}(x_{i,j}, y_i) \lesssim \frac{\log d}{\eta} + \eta \sum_{i=1}^n \ell_{0-1}(x_{i,j}, y_i). \tag{10.2.3}$$

By making the choice $\eta \approx \sqrt{\log d / n}$ and noting that $\ell_{0-1} \le 1$, we obtain

$$\sum_{i=1}^n \ell_{0-1}(\widehat{y}_i, y_i) - \ell_{0-1}(x_{i,j}, y_i) \lesssim \sqrt{n \log d}$$

for any collection of experts and any sequence y_1^n. ♣

We make a few remarks on the preceding example to close the chapter. First, ideally we would like to attain adaptive regret guarantees, meaning that the regret scales with the performance of the best predictor in inequality (10.2.3). In particular, we might expect that a good expert would satisfy $\sum_{i=1}^n \ell_{0-1}(x_{i,j}, y_i) \ll n$, which—if we could choose

$$\eta \approx \left(\frac{\log d}{\sum_{i=1}^n \ell_{0-1}(x_{i,j^*}, y_i)}\right)^{\frac{1}{2}},$$

where $j^* = \operatorname{argmin}_j \sum_{i=1}^n \ell_{0-1}(x_{i,j}, y_i)$—then we would attain regret bound

$$\sqrt{\log d \cdot \sum_{i=1}^n \ell_{0-1}(x_{i,j^*}, y_i)} \ll \sqrt{n \log d}.$$

For results of this form, see, for example, Cesa-Bianchi et al. [5] or the more recent work on mirror descent of Steinhardt and Liang [16].

Secondly, we note that it is actually possible to give a regret bound of the form (10.2.3) without relying on the near exp-concavity condition (10.2.1). In particular, performing mirror descent on the convex losses defined by

$$\pi \mapsto \left| \sum_{j=1}^d \operatorname{sign}(x_{i,j}) \pi_j - \operatorname{sign}(y_i) \right|,$$

which is convex, will give a regret bound of $\sqrt{n \log d}$ for the zero-one loss as well. We leave this exploration to the interested reader.

Bibliography

[1] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[2] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[3] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Annals of Statistics*, 27(6):1865–1895, 1999.

[4] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006.

[5] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2–3):321–352, 2007.

[6] L. D. Davisson. The prediction error of stationary gaussian time series of unknown covariance. *IEEE Transactions on Information Theory*, 11:527–532, 1965.

[7] D. Haussler, J. Kivinen, and M. K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.

[8] E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2006.

[9] P. J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

[10] J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–64, Jan. 1997.

[11] J. Kivinen and M. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45(3):301–329, July 2001.

[12] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[13] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

[14] B. T. Polyak and J. Tsypkin. Robust identification. *Automatica*, 16:53–63, 1980. doi: 10.1016/0005-1098(80)90086-2. URL `http://dx.doi.org/10.1016/0005-1098(80)90086-2`.

[15] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30:629–636, 1984.

[16] J. Steinhardt and P. Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Chapter 11

Online convex optimization

A related notion to the universal prediction problem with alternate losses is that of *online learning* and *online convex optimization*, where we modify the requirements of Chapter 10 further. In the current setting, we essentially do away with distributional assumptions at all, including prediction with a distribution, and we consider the following two player sequential game: we have a space $\mathcal{W}$ in which we—the learner or first player—can play points $w_1, w_2, \ldots$, while nature plays a sequence of loss functions $\ell_t : \mathcal{W} \to \mathbb{R}$. The goal is to guarantee that the regret

$$\sum_{t=1}^{n} \left[\ell_t(w_t) - \ell_t(w^\star)\right] \tag{11.0.1}$$

grows at most sub-linearly with n, for any $w^\star \in \mathcal{W}$ (often, we desire this guarantee to be uniform). As stated, this goal is too broad, so in this chapter we focus on a few natural restrictions, namely, that the sequence of losses ℓ_t are convex, and $\mathcal{W}$ is a convex subset of $\mathbb{R}^d$. In this setting, the problem (11.0.1) is known as *online convex programming.*

11.1 The problem of online convex optimization

Before proceeding, we provide a few relevant definitions to make our discussion easier; we refer to Appendix A for an overview of convexity and proofs of a variety of useful properties of convex sets and functions. First, we recall that a set $\mathcal{W}$ is *convex* if for all $\lambda \in [0, 1]$ and $w, w' \in \mathcal{W}$, we have

$$\lambda w + (1 - \lambda) w' \in \mathcal{W}.$$

Similarly, a function f is *convex* if

$$f(\lambda w + (1 - \lambda) w') \le \lambda f(w) + (1 - \lambda) f(w')$$

for all $\lambda \in [0, 1]$ and w, w'. The *subgradient set*, or *subdifferential*, of a convex function f at the point w is defined to be

$$\partial f(w) := \{g \in \mathbb{R}^d : f(v) \ge f(w) + \langle g, v - w\rangle \text{ for all } v\},$$

and we say that any vector $g \in \mathbb{R}^d$ satisfying $f(v) \ge f(w) + \langle g, v - w\rangle$ for all v is a *subgradient.* For convex functions, the subdifferential set $\partial f(w)$ is essentially always non-empty for any $w \in \operatorname{dom} f$.[1]

[1] Rigorously, we are guaranteed that $\partial f(w) \neq \emptyset$ at all points w in the relative interior of the domain of f.

We now give several examples of convex functions, losses, and corresponding subgradients. The first two examples are for *classification problems*, in which we receive data points $x \in \mathbb{R}^d$ and wish to predict associated labels $y \in \{-1, 1\}$.

Example 11.1 (Support vector machines): In the support vector machine problem, we receive data in pairs $(x_t, y_t) \in \mathbb{R}^d \times \{-1, 1\}$, and the loss function

$$\ell_t(w) = \left[1 - y_t \langle w, x_t \rangle\right]_+ = \max\{1 - y_t \langle w, x_t \rangle, 0\},$$

which is convex because it is the maximum of two linear functions. Moreover, the subgradient set is

$$\partial \ell_t(w) = \begin{cases} -y_t x_t & \text{if } y_t \langle w, x_t \rangle < 1 \\ -\lambda \cdot y_t x_t & \text{for } \lambda \in [0,1] \text{ if } y_t \langle w, x_t \rangle = 1 \\ 0 & \text{otherwise.} \end{cases}$$

♣

Example 11.2 (Logistic regression): As in the support vector machine, we receive data in pairs $(x_t, y_t) \in \mathbb{R}^d \times \{-1, 1\}$, and the loss function is

$$\ell_t(w) = \log(1 + \exp(-y_t \langle x_t, w \rangle)).$$

To see that this loss is convex, note that if $h(t) = \log(1 + e^t)$, then $h'(t) = \frac{1}{1+e^{-t}}$ and $h''(t) = \frac{e^{-t}}{(1+e^{-t})^2} \ge 0$, and ℓ_t is the composition of a linear transformation with h. In this case,

$$\partial \ell_t(w) = \nabla \ell_t(w) = -\frac{1}{1 + e^{y_t \langle x_t, w \rangle}} y_t x_t.$$

♣

Example 11.3 (Expert prediction and zero-one error): By randomization, it is possible to cast certain non-convex optimization problems as convex. Indeed, let us assume that there are d experts, each of which makes a prediction $x_{t,j}$ (for $j = 1, \ldots, d$) at time t, represented by the vector $x_t \in \mathbb{R}^d$, of a label $y_t \in \{-1, 1\}$. Each also suffers the (non-convex) loss $\ell_{0-1}(x_{t,j}, y_t) = \mathbf{1}\{x_{t,j} y_t \le 0\}$. By assigning a weight w_j to each expert $x_{t,j}$ subject to the constraint that $w \succeq 0$ and $\langle w, \mathbb{1} \rangle = 1$, then if we were to randomly choose to predict using expert j with probability w_j, we would suffer expected loss at time t of

$$\ell_t(w) = \sum_{j=1}^d w_j \ell_{0-1}(x_{t,j}, y_t) = \langle g_t, w \rangle,$$

where we have defined the vector $g_t = [\ell_{0-1}(x_{t,j}, y_t)]_{j=1}^d \in \{0, 1\}^d$. Notably, the expected zero-one loss is convex (even linear), so that its online minimization falls into the online convex programming framework. ♣

As we see in the sequel, online convex programming approaches are often quite simple, and, in fact, are often provably optimal in a variety of scenarios *outside* of online convex optimization. This motivates our study, and we will see that online convex programming approaches have a number of similarities to our regret minimization approaches in previous chapters on universal coding, regret, and redundancy.

11.2 Online gradient and non-Euclidean gradient (mirror) descent

We now turn to an investigation of the single approach we will use to solve online convex optimization problems, which is known as *mirror descent*.[2] Before describing the algorithm in its full generality, however, we first demonstrate a special case (though our analysis will be for the general algorithm).

Roughly, the intuition for our procedures is as follows: after observing a loss ℓ_t, we make a small update to move our estimate w_t in a direction to improve the value of the losses we have seen. However, so that we do not make progress too quickly—or too aggressively follow spurious information—we attempt to keep new iterates close to previous iterates. With that in mind, we present *(projected) online gradient descent*, which requires only that we specify a sequence η_t of non-increasing stepsizes.

Input: Parameter space $\mathcal{W}$, stepsize sequence η_t.
Repeat: for each iteration t, predict $w_t \in \mathcal{W}$, receive function ℓ_t and suffer loss $\ell_t(w_t)$. Compute any $g_t \in \partial \ell_t(w_t)$, and perform subgradient update

$$w_{t+\frac{1}{2}} = w_t - \eta_t g_t, \quad w_{t+1} = \Pi_{\mathcal{W}}(w_{t+\frac{1}{2}}), \tag{11.2.1}$$

where $\Pi_{\mathcal{W}}$ denotes (Euclidean) projection onto $\mathcal{W}$.

Figure 11.1: Online projected gradient descent.

An equivalent formulation of the update (11.2.1) is to write it as the single step

$$w_{t+1} = \operatorname*{argmin}_{w \in \mathcal{W}} \left\{ \langle g_t, w \rangle + \frac{1}{2\eta_t} \|w - w_t\|_2^2 \right\}, \tag{11.2.2}$$

which makes clear that we trade between improving performance on ℓ_t via the linear approximation of $\ell_t(w) \approx \ell_t(w_t) + g_t^\top (w - w_t)$ and remaining close to w_t according to the Euclidean distance $\|\cdot\|_2$. In a variety of scenarios, however, it is quite advantageous to measure distances in a way more amenable to the problem structure, for example, if $\mathcal{W}$ is a probability simplex or we have prior information about the loss functions ℓ_t that nature may choose. With this in mind, we present a slightly more general algorithm, which requires us to give a few more definitions.

Given a convex differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$, we define the *Bregman divergence* associated with ψ by

$$B_\psi(w, v) = \psi(w) - \psi(v) - \langle \nabla \psi(v), w - v \rangle. \tag{11.2.3}$$

The Bregman divergence is always non-negative, as $B_\psi(w, v)$ is the gap between the true function value $\psi(w)$ and its linear approximation at the point v (see Figure 11.2). A few examples illustrate its properties.

Example 11.4 (Euclidean distance as Bregman divergence)**:** Take $\psi(w) = \frac{1}{2} \|w\|_2^2$ to obtain $B(w, v) = \frac{1}{2} \|w - v\|_2^2$. More generally, if for a matrix A we define $\|w\|_A^2 = w^\top A w$, then takin $\psi(w) = \frac{1}{2} w^\top A w$, we have

$$B_\psi(w, v) = \frac{1}{2}(w - v)^\top A (w - v) = \frac{1}{2} \|w - v\|_A^2.$$

So Bregman divergences generalize (squared) Euclidean distance. ♣

[2] The reasons for this name are somewhat convoluted, and we do not dwell on them.

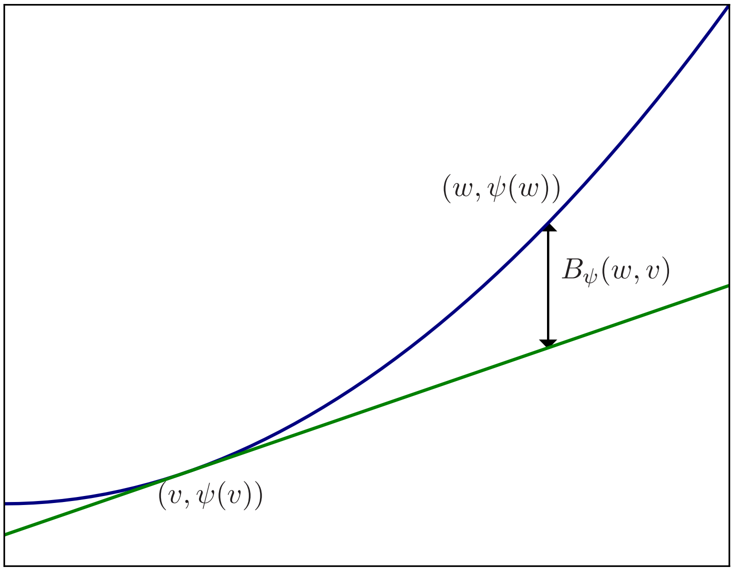

Figure 11.2: Illustration of Bregman divergence.

Example 11.5 (KL divergence as a Bregman divergence)**:** Take $\psi(w) = \sum_{j=1}^d w_j \log w_j$. Then ψ is convex over the positive orthant $\mathbb{R}_+^d$ (the second derivative of $w \log w$ is $1/w$), and for $w, v \in \Delta_d = \{u \in \mathbb{R}_+^d : \langle \mathbb{1}, u \rangle = 1\}$, we have

$$B_\psi(w, v) = \sum_j w_j \log w_j - \sum_j v_j \log v_j - \sum_j (1 + \log v_j)(w_j - v_j) = \sum_j w_j \log \frac{w_j}{v_j} = D_{\rm kl}\left(w \| v\right),$$

where in the final equality we treat w and v as probability distributions on $\{1, \ldots, d\}$. ♣

With these examples in mind, we now present the mirror descent algorithm, which is the natural generalization of online gradient descent.

Input: proximal function ψ, parameter space $\mathcal{W}$, and non-increasing stepsize sequence $\eta_1, \eta_2, \ldots$.
Repeat: for each iteration t, predict $w_t \in \mathcal{W}$, receive function ℓ_t and suffer loss $\ell_t(w_t)$. Compute any $g_t \in \partial \ell_t(w_t)$, and perform non-Euclidean subgradient update

$$w_{t+1} = \mathop{\rm argmin}_{w \in \mathcal{W}} \left\{ \langle g_t, w \rangle + \frac{1}{\eta_t} B_\psi(w, w_t) \right\}. \tag{11.2.4}$$

Figure 11.3: The online mirror descent algorithm

Before providing the analysis of Algorithm 11.3, we give a few examples of its implementation. First, by taking $\mathcal{W} = \mathbb{R}^d$ and $\psi(w) = \frac{1}{2}\left\|w\right\|_2^2$, we note that the mirror descent procedure simply corresponds to the gradient update $w_{t+1} = w_t - \eta_t g_t$. We can also recover the *exponentiated gradient* algorithm, also known as entropic mirror descent.

Example 11.6 (Exponentiated gradient algorithm)**:** Suppose that we have $\mathcal{W} = \Delta_d = \{w \in \mathbb{R}^d_+ : \langle \mathbb{1}, w \rangle = 1\}$, the probability simplex in $\mathbb{R}^d$. Then a natural choice for ψ is the negative entropy, $\psi(w) = \sum_j w_j \log w_j$, which (as noted previously) gives $B_\psi(w, v) = \sum_j w_j \log \frac{w_j}{v_j}$. We now consider the update step (11.2.4). In this case, fixing $v = w_t$ for notational simplicity, we must solve

$$\text{minimize} \;\; \langle g, w \rangle + \frac{1}{\eta} \sum_j w_j \log \frac{w_j}{v_j} \;\; \text{subject to } w \in \Delta_d$$

in w. Writing the Lagrangian for this problem after introducing multipliers $\tau \in \mathbb{R}$ for the contraint that $\langle \mathbb{1}, w \rangle = 1$ and $\lambda \in \mathbb{R}^d_+$ for $w \succeq 0$, we have

$$\mathcal{L}(w, \lambda, \tau) = \langle g, w \rangle + \frac{1}{\eta} \sum_{j=1}^d w_j \log \frac{w_j}{v_j} - \langle \lambda, w \rangle + \tau(\langle \mathbb{1}, w \rangle - 1),$$

which is minimized by taking

$$w_j = v_j \exp(-\eta g_j + \lambda_j \eta - \tau \eta - 1),$$

and as $w_j > 0$ certainly, the constraint $w \succeq 0$ is inactive and $\lambda_j = 0$. Thus, choosing τ to normalize the w_j, we obtain the *exponentiated gradient update*

$$w_{t+1,i} = \frac{w_{t,i} e^{-\eta_t g_{t,i}}}{\sum_j w_{t,j} e^{-\eta_t g_{t,j}}} \;\; \text{for } i = 1, \ldots, d,$$

as the explicit calculation of the mirror descent update (11.2.4). ♣

We now turn to an analysis of the mirror descent algorithm. Before presenting the analysis, we require two more definitions that allow us to relate Bregman divergences to various norms.

Definition 11.1. *Let $\|\cdot\|$ be a norm. The* dual norm $\|\cdot\|_*$ *associated with $\|\cdot\|$ is*

$$\|y\|_* := \sup_{x : \|x\| \le 1} x^\top y.$$

For example, a straightforward calculation shows that the dual to the ℓ_∞-norm is the ℓ_1-norm, and the Euclidean norm $\|\cdot\|_2$ is self-dual (by the Cauchy-Schwarz inequality). Lastly, we require a definition of functions of suitable curvature for use in mirror descent methods.

Definition 11.2. *A convex function $f : \mathbb{R}^d \to \mathbb{R}$ is* strongly convex with respect to the norm $\|\cdot\|$ over the set $\mathcal{W}$ *if for all $w, v \in \mathcal{W}$ and $g \in \partial f(w)$ we have*

$$f(v) \ge f(w) + \langle g, v - w \rangle + \frac{1}{2} \|w - v\|^2 .$$

That is, the function f is strongly convex if it grows at least quadratically fast at every point in its domain. It is immediate from the definition of the Bregman divergence that ψ is strongly convex if and only if

$$B_\psi(w, v) \ge \frac{1}{2} \|w - v\|^2 .$$

As two examples, we consider Euclidean distance and entropy. For the Euclidean distance, which uses $\psi(w) = \frac{1}{2} \|w\|_2^2$, we have $\nabla \psi(w) = w$, and

$$\frac{1}{2} \|v\|_2^2 = \frac{1}{2} \|w + v - w\|_2^2 = \frac{1}{2} \|w\|_2^2 + \langle w, v - w \rangle + \frac{1}{2} \|w - v\|_2^2$$

by a calculation, so that ψ is strongly convex with respect to the Euclidean norm. We also have the following observation.

Observation 11.7. *Let $\psi(w) = \sum_j w_j \log w_j$ be the negative entropy. Then ψ is strongly convex with respect to the ℓ_1-norm, that is,*

$$B_\psi(w, v) = D_{\rm kl}\left(w \| v\right) \ge \frac{1}{2} \left\|w - v\right\|_1^2.$$

Proof The result is an immediate consequence of Pinsker's inequality, Proposition 2.10. □

With these examples in place, we present the main theorem of this section.

Theorem 11.8 (Regret of mirror descent)**.** *Let ℓ_t be an arbitrary sequence of convex functions, and let w_t be generated according to the mirror descent algorithm 11.3. Assume that the proximal function ψ is strongly convex with respect to the norm $\|\cdot\|$, which has dual norm $\|\cdot\|_*$. Then*

(a) If $\eta_t = \eta$ for all t, then for any $w^\star \in \mathcal{W}$,

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{1}{\eta} B_\psi(w^\star, w_1) + \frac{\eta}{2} \sum_{t=1}^n \left\|g_t\right\|_*^2.$$

(b) If $\mathcal{W}$ is compact and $B_\psi(w^\star, w) \le R^2$ for any $w \in \mathcal{W}$, then

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{1}{2\eta_n} R^2 + \sum_{t=1}^n \frac{\eta_t}{2} \left\|g_t\right\|_*^2.$$

Before proving the theorem, we provide a few comments to exhibit its power. First, we consider the Euclidean case, where $\psi(w) = \frac{1}{2}\left\|w\right\|_2^2$, and we assume that the loss functions ℓ_t are all L-Lipschitz, meaning that $|\ell_t(w) - \ell_t(v)| \le L\left\|w - v\right\|_2$, which is equivalent to $\left\|g_t\right\|_2 \le L$ for all $g_t \in \partial \ell_t(w)$. In this case, the two regret bounds above become

$$\frac{1}{2\eta}\left\|w^\star - w_1\right\|_2^2 + \frac{\eta}{2} n L^2 \quad \text{and} \quad \frac{1}{2\eta_n} R^2 + \sum_{t=1}^n \frac{\eta_t}{2} L^2,$$

respectively, where in the second case we assumed that $\left\|w^\star - w_t\right\|_2 \le R$ for all t. In the former case, we take $\eta = \frac{R}{L\sqrt{n}}$, while in the second, we take $\eta_t = \frac{R}{L\sqrt{t}}$, which does not require knowledge of n ahead of time. Focusing on the latter case, we have the following corollary.

Corollary 11.9. *Assume that $\mathcal{W} \subset \{w \in \mathbb{R}^d : \left\|w\right\|_2 \le R\}$ and that the loss functions ℓ_t are L-Lipschitz with respect to the Euclidean norm. Take $\eta_t = \frac{R}{L\sqrt{t}}$. Then for all $w^\star \in \mathcal{W}$,*

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le 3RL\sqrt{n}.$$

Proof For any $w, w^\star \in \mathcal{W}$, we have $\left\|w - w^\star\right\|_2 \le 2R$, so that $B_\psi(w^\star, w) \le 4R^2$. Using that

$$\sum_{t=1}^n t^{-\frac{1}{2}} \le \int_0^n t^{-\frac{1}{2}} dt = 2\sqrt{n}$$

gives the result. □

Now that we have presented the Euclidean variant of online convex optimization, we turn to an example that achieves better performance in high dimensional settings, as long as the domain is the probability simplex. (Recall Example 11.3 for motivation.) In this case, we have the following corollary to Theorem 11.8.

Corollary 11.10. *Assume that $\mathcal{W} = \Delta_d = \{w \in \mathbb{R}^d_+ : \langle \mathbb{1}, w\rangle = 1\}$ and take the proximal function $\psi(w) = \sum_j w_j \log w_j$ to be the negative entropy in the mirror descent procedure 11.3. Then with the fixed stepsize η and initial point as the uniform distribution $w_1 = \mathbb{1}/d$, we have for any sequence of convex losses ℓ_t*

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2}\sum_{t=1}^n \|g_t\|_\infty^2 .$$

Proof Using Pinsker's inequality in the form of Observation 11.7, we have that ψ is strongly convex with respect to $\|\cdot\|_1$. Consequently, taking the dual norm to be the ℓ_∞-norm, part (a) of Theorem 11.8 shows that

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{1}{\eta}\sum_{j=1}^d w_j^\star \log\frac{w_j^\star}{w_{1,j}} + \frac{\eta}{2}\sum_{t=1}^n \|g_t\|_\infty^2 .$$

Noting that with $w_1 = \mathbb{1}/d$, we have $B_\psi(w^\star, w_1) \le \log d$ for any $w^\star \in \mathcal{W}$ gives the result. □

Corollary 11.10 yields somewhat sharper results than Corollary 11.9, though in the restricted setting that $\mathcal{W}$ is the probability simplex in $\mathbb{R}^d$. Indeed, let us assume that the subgradients $g_t \in [-1,1]^d$, the hypercube in $\mathbb{R}^d$. In this case, the tightest possible bound on their ℓ_2-norm is $\|g_t\|_2 \le \sqrt{d}$, while $\|g_t\|_\infty \le 1$ always. Similarly, if $\mathcal{W} = \Delta_d$, then while we are only guaranteed that $\|w^\star - w_1\|_2 \le 1$. Thus, the best regret guaranteed by the Euclidean case (Corollary 11.9) is

$$\frac{1}{2\eta}\|w^\star - w_1\|_2^2 + \frac{\eta}{2}nd \le \sqrt{nd} \quad \text{with the choice } \eta = \frac{1}{\sqrt{nd}},$$

while the entropic mirror descent procedure (Alg. 11.3 with $\psi(w) = \sum_j w_j \log w_j$) guarantees

$$\frac{\log d}{\eta} + \frac{\eta}{2}n \le \sqrt{2n\log d} \quad \text{with the choice } \eta = \frac{\sqrt{2\log d}}{2\sqrt{n}}. \tag{11.2.5}$$

The latter guarantee is *exponentially* better in the dimension. Moreover, the key insight is that we essentially maintain a "prior," and then perform "Bayesian"-like updating of the posterior distribution w_t at each time step, exactly as in the setting of redundancy minimization.

11.2.1 Proof of Theorem 11.8

The proof of the theorem proceeds in three lemmas, which are essentially inductive applications of optimality conditions for convex optimization problems. The first is the explicit characterization of optimality for a convex optimization problem. (For a proof of this lemma, see, for example, the books of Hiriart-Urruty and Lemaréchal [2, 3], or Section 2.5 of Boyd et al. [1].)

Lemma 11.11. *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a convex function and $\mathcal{W}$ be a convex set. Then $w^\star$ minimizes $h(w)$ over $\mathcal{W}$ if and only if there exists $g \in \partial h(w^\star)$ such that*

$$\langle g, w - w^\star\rangle \ge 0 \quad \textit{for all } w \in \mathcal{W}.$$

Lemma 11.12. *Let $\ell_t : \mathcal{W} \to \mathbb{R}$ be any sequence of convex loss functions and η_t be a non-increasing sequence, where $\eta_0 = \infty$. Then with the mirror descent strategy (11.2.4), for any $w^\star \in \mathcal{W}$ we have*

$$\sum_{t=1}^n \ell_t(w_t) - \ell_t(w^\star) \le \sum_{t=1}^n \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) B_\psi(w^\star, w_t) + \sum_{t=1}^n \left[-\frac{1}{\eta_t} B_\psi(w_{t+1}, w_t) + \langle g_t, w_t - w_{t+1}\rangle\right].$$

Proof Our proof follows by the application of a few key identities. First, we note that by convexity, we have for any $g_t \in \partial \ell_t(w_t)$ that

$$\ell_t(w_t) - \ell_t(w^\star) \le \langle g_t, w_t - w^\star\rangle. \tag{11.2.6}$$

Secondly, we have that because w_{t+1} minimizes

$$\langle g_t, w\rangle + \frac{1}{\eta_t} B_\psi(w, w_t)$$

over $w \in \mathcal{W}$, then Lemma 11.11 implies

$$\langle \eta_t g_t + \nabla\psi(w_{t+1}) - \nabla\psi(w_t), w - w_{t+1}\rangle \ge 0 \text{ for all } w \in \mathcal{W}. \tag{11.2.7}$$

Taking $w = w^\star$ in inequality (11.2.7) and making a substitution in inequality (11.2.6), we have

$$\begin{aligned} \ell_t(w_t) - \ell_t(w^\star) &\le \langle g_t, w_t - w^\star\rangle = \langle g_t, w_{t+1} - w^\star\rangle + \langle g_t, w_t - w_{t+1}\rangle \\ &\le \frac{1}{\eta_t} \langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t), w^\star - w_{t+1}\rangle + \langle g_t, w_t - w_{t+1}\rangle \\ &= \frac{1}{\eta_t} \left[B_\psi(w^\star, w_t) - B_\psi(w^\star, w_{t+1}) - B_\psi(w_{t+1}, w_t)\right] + \langle g_t, w_t - w_{t+1}\rangle \end{aligned} \tag{11.2.8}$$

where the final equality (11.2.8) follows from algebraic manipulations of $B_\psi(w, w')$. Summing inequality (11.2.8) gives

$$\begin{aligned} \sum_{t=1}^n \ell_t(w_t) - \ell_t(w^\star) &\le \sum_{t=1}^n \frac{1}{\eta_t} \left[B_\psi(w^\star, w_t) - B_\psi(w^\star, w_{t+1}) - B_\psi(w_{t+1}, w_t)\right] + \sum_{t=1}^n \langle g_t, w_t - w_{t+1}\rangle \\ &= \sum_{t=2}^n \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) B_\psi(w^\star, w_t) + \frac{1}{\eta_1} B_\psi(w^\star, w_1) - \frac{1}{\eta_n} B_\psi(w^\star, w_{n+1}) \\ &\quad + \sum_{t=1}^n \left[-\frac{1}{\eta_t} B_\psi(w_{t+1}, w_t) + \langle g_t, w_t - w_{t+1}\rangle\right] \end{aligned}$$

as desired. □

It remains to use the negative terms $-B_\psi(w_t, w_{t+1})$ to cancel the gradient terms $\langle g_t, w_t - w_{t+1}\rangle$. To that end, we recall Definition 11.1 of the dual norm $\|\cdot\|_*$ and the strong convexity assumption on ψ. Using the Fenchel-Young inequality, we have

$$\langle g_t, w_t - w_{t+1}\rangle \le \|g_t\|_* \|w_t - w_{t+1}\| \le \frac{\eta_t}{2} \|g_t\|_*^2 + \frac{1}{2\eta_t} \|w_t - w_{t+1}\|^2.$$

Now, we use the strong convexity condition, which gives

$$-\frac{1}{\eta_t} B_\psi(w_{t+1}, w_t) \le -\frac{1}{2\eta_t} \|w_t - w_{t+1}\|^2.$$

Combining the preceding two displays in Lemma 11.12 gives the result of Theorem 11.8.

11.3 Online to batch conversions

Martingales!

11.4 More refined convergence guarantees

It is sometimes possible to give more refined bounds than those we have so far provided. As motivation, let us revisit Example 11.3, but suppose that one of the experts has no loss—that is, it makes perfect predictions. We might expect—accurately!—that we should attain better convergence guarantees using exponentiated weights, as the points w_t be maintain should quickly eliminate non-optimal experts.

To that end, we present a refined regret bound for the mirror descent algorithm 11.3 with the entropic regularization $\psi(w) = \sum_j w_j \log w_j$.

Proposition 11.13. *Let $\psi(w) = \sum_j w_j \log w_j$, and assume that the losses ℓ_t are such that their subgradients have all non-negative entries, that is, $g_t \in \partial \ell_t(w)$ implies $g_t \succeq 0$. For any such sequence of loss functions ℓ_t and any $w^\star \in \mathcal{W} = \Delta_d$,*

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{j=1}^d w_{t,j} g_{t,j}^2.$$

While as stated, the bound of the proposition does not look substantially more powerful than Corollary 11.10, but a few remarks will exhibit its consequences. We prove the proposition in Section 11.4.1 to come.

First, we note that because $w_t \in \Delta_d$, we will *always* have $\sum_j w_{t,j} g_{t,j}^2 \le \|g_t\|_\infty^2$. So certainly the bound of Proposition 11.13 is never worse than that of Corollary 11.10. Sometimes this can be made tighter, however, as exhibited by the next corollary, which applies (for example) to the experts setting of Example 11.3. More specifically, we have d experts, each suffering losses in $[0, 1]$, and we seek to predict with the best of the d experts.

Corollary 11.14. *Consider the linear online convex optimization setting, that is, where $\ell_t(w_t) = \langle g_t, w_t \rangle$ for vectors g_t, and assume that $g_t \in \mathbb{R}_+^d$ with $\|g_t\|_\infty \le 1$. In addition, assume that we know an upper bound $L_n^\star$ on $\sum_{t=1}^n \ell_t(w^\star)$. Then taking the stepsize $\eta = \min\{1, \sqrt{\log d}/\sqrt{L_n^\star}\}$, we have*

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le 3 \max\left\{\log d, \sqrt{L_n^\star \log d}\right\}.$$

Note that when $\ell_t(w^\star) = 0$ for all $w^\star$, which corresponds to a perfect expert in Example 11.3, the upper bound becomes constant in n, yielding $3 \log d$ as a bound on the regret. Unfortunately, in our bound of Corollary 11.14, we had to assume that we *knew* ahead of time a bound on the loss of the best predictor $w^\star$, which is unrealistic in practice. There are a number of techniques for dealing with such issues, including a standard one in the online learning literature known as the *doubling* trick. We explore some in the exercises.

Proof First, we note that $\sum_j w_j g_{t,j}^2 \le \langle w, g_t \rangle$ for any nonnegative vector w, as $g_{t,j} \in [0, 1]$. Thus, Proposition 11.13 gives

$$\sum_{t=1}^n [\ell_t(w_t) - \ell_t(w^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \langle w_t, g_t \rangle = \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \ell_t(w_t).$$

Rearranging via an algebraic manipulation, this is equivalent to

$$\Big(1-\frac{\eta}{2}\Big)\sum_{t=1}^n[\ell_t(w_t)-\ell_t(w^\star)] \le \frac{\log d}{\eta}+\frac{\eta}{2}\sum_{t=1}^n \ell_t(w^\star).$$

Take $\eta = \min\{1, \sqrt{\log d / L_n^\star}\}$. Then if $\sqrt{\log d/L_n^\star} \le 1$, we have that the right hand side of the above inequality becomes $\sqrt{L_n^\star \log d}+\frac{1}{2}\sqrt{L_n^\star \log d}$. On the other hand, if $L_n^\star < \log d$, then the right hand side of the inequality becomes $\log d + \frac{1}{2}L_n^\star \le \frac{3}{2}\log d$. In either case, we obtain the desired result by noting that $1-\frac{\eta}{2} \ge \frac{1}{2}$. □

11.4.1 Proof of Proposition 11.13

Our proof relies on a technical lemma, after which the derivation is a straightforward consequence of Lemma 11.12. We first state the technical lemma, which applies to the update that the exponentiated gradient procedure makes.

Lemma 11.15. *Let $\psi(x) = \sum_j x_j \log x_j$, and let $x, y \in \Delta_d$ be defined by*

$$y_i = \frac{x_i \exp(-\eta g_i)}{\sum_j x_j \exp(-\eta g_j)},$$

where $g \in \mathbb{R}_+^d$ is non-negative. Then

$$-\frac{1}{\eta}B_\psi(y,x) + \langle g, x-y\rangle \le \frac{\eta}{2}\sum_{i=1}^d g_i^2 x_i.$$

Deferring the proof of the lemma, we note that it precisely applies to the setting of Lemma 11.12. Indeed, with a fixed stepsize η, we have

$$\sum_{t=1}^n \ell_t(w_t)-\ell_t(w^\star) \le \frac{1}{\eta}B_\psi(w^\star, w_1) + \sum_{t=1}^n\left[-\frac{1}{\eta}B_\psi(w_{t+1},w_t)+\langle g_t, w_t-w_{t+1}\rangle\right].$$

Earlier, we used the strong convexity of ψ to eliminate the gradient terms $\langle g_t, w_t - w_{t+1}\rangle$ using the bregman divergence B_ψ. This time, we use Lemma 11.12: setting $y = w_{t+1}$ and $x = w_t$ yields the bound

$$\sum_{t=1}^n \ell_t(w_t)-\ell_t(w^\star) \le \frac{1}{\eta}B_\psi(w^\star, w_1) + \sum_{t=1}^n \frac{\eta}{2}\sum_{i=1}^d g_{t,i}^2 w_{t,i}$$

as desired.

Proof of Lemma 11.15 We begin by noting that a direct calculation yields $B_\psi(y,x) = D_{\rm kl}(y\|x) = \sum_i y_i \log\frac{y_i}{x_i}$. Substituting the values for x and y into this expression, we have

$$\sum_i y_i \log\frac{y_i}{x_i} = \sum_i y_i \log\left(\frac{x_i\exp(-\eta g_i)}{x_i(\sum_j \exp(-\eta g_j)x_j)}\right) = -\eta\,\langle g, y\rangle - \sum_i y_i \log\Big(\sum_j x_j e^{-\eta g_j}\Big).$$

Now we use a Taylor expansion of the function $g \mapsto \log(\sum_j x_j e^{-\eta g_j})$ around the point 0. If we define the vector $p(g)$ by $p_i(g) = x_i e^{-\eta g_i} / (\sum_j x_j e^{-\eta g_j})$, then

$$\log\Big(\sum_j x_j e^{-\eta g_j}\Big) = \log(\langle \mathbb{1}, x\rangle) - \eta \langle p(0), g\rangle + \frac{\eta^2}{2} g^\top (\mathrm{diag}(p(\widetilde{g})) - p(\widetilde{g}) p(\widetilde{g})^\top) g,$$

where $\widetilde{g} = \lambda g$ for some $\lambda \in [0, 1]$. Noting that $p(0) = x$ and $\langle \mathbb{1}, x\rangle = \langle \mathbb{1}, y\rangle = 1$, we obtain

$$B_\psi(y, x) = -\eta \langle g, y\rangle + \log(1) + \eta \langle g, x\rangle - \frac{\eta^2}{2} g^\top (\mathrm{diag}(p(\widetilde{g})) - p(\widetilde{g}) p(\widetilde{g})^\top) g,$$

whence

$$-\frac{1}{\eta} B_\psi(y, x) + \langle g, x - y\rangle \le \frac{\eta}{2} \sum_{i=1}^d g_i^2 p_i(\widetilde{g}). \tag{11.4.1}$$

Lastly, we claim that the function

$$s(\lambda) = \sum_{i=1}^d g_i^2 \frac{x_i e^{-\lambda g_i}}{\sum_j x_j e^{-\lambda g_j}}$$

is non-increasing on $\lambda \in [0, 1]$. Indeed, we have

$$s'(\lambda) = \frac{(\sum_i g_i x_i e^{-\lambda g_i})(\sum_i g_i^2 x_i e^{-\lambda g_i})}{(\sum_i x_i e^{-\lambda g_i})^2} - \frac{\sum_i g_i^3 x_i e^{-\lambda g_i}}{\sum_i x_i e^{-\lambda g_i}} = \frac{\sum_{ij} g_i g_j^2 x_i x_j e^{-\lambda g_i - \lambda g_j} - \sum_{ij} g_i^3 x_i x_j e^{-\lambda g_i - \lambda g_j}}{(\sum_i x_i e^{-\lambda g_i})^2}.$$

Using the Fenchel-Young inequality, we have $ab \le \frac{1}{3}|a|^3 + \frac{2}{3}|b|^{3/2}$ for any a, b, so $g_i g_j^2 \le \frac{1}{3} g_i^3 + \frac{2}{3} g_j^3$. This implies that the numerator in our expression for $s'(\lambda)$ is non-positive. Thus we have $s(\lambda) \le s(0) = \sum_{i=1}^d g_i^2 x_i$, which gives the result when combined with inequality (11.4.1). □

Bibliography

[1] S. Boyd, J. Duchi, and L. Vandenberghe. Subgradients. Course notes for Stanford Course EE364b, 2015. URL `http://web.stanford.edu/class/ee364b/lectures/subgradients_notes.pdf`.

[2] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

[3] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001.

Chapter 12

Exploration, exploitation, and bandit problems

Consider the following problem: we have a possible treatment for a population with a disease, but we do not know whether the treatment will have a positive effect or not. We wish to evaluate the treatment to decide whether it is better to apply it or not, and we wish to optimally allocate our resources to attain the best outcome possible. There are challenges here, however, because for each patient, we may only observe the patient's behavior and disease status in one of two possible states—under treatment or under control—and we wish to allocate as few patients to the group with worse outcomes (be they control or treatment) as possible. This balancing act between exploration—observing the effects of treatment or non-treatment—and exploitation—giving treatment or not as we decide which has better palliative outcomes—underpins and is the paridigmatic aspect of the multi-armed bandit problem.[1]

Our main focus in this chapter is a fairly simple variant of the K-armed bandit problem, though we note that there is a substantial literature in statistics, operations research, economics, game theory, and computer science on variants of the problems we consider. In particular, we consider the following sequential decision making scenario. We assume that there are K distributions $P_1, \ldots, P_K$ on $\mathbb{R}$, which we identify (with no loss of generality) with K random variables $Y_1, \ldots, Y_K$. Each random variable Y_i has mean μ_i and is σ^2-sub-Gaussian, meaning that

$$\mathbb{E}\left[\exp\left(\lambda(Y_i - \mu_i)\right)\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right). \tag{12.0.1}$$

The goal is to find the index i with the maximal mean μ_i without evaluating sub-optimal "arms" (or random variables Y_i) too often. At each iteration t of the process, the player takes an action $A_t \in \{1, \ldots, K\}$, then, conditional on $i = A_t$, observes a reward $Y_i(t)$ drawn independently from the distribution P_i. Then the goal is to minimize the the regret after n steps, which is

$$\mathsf{Reg}_n := \sum_{t=1}^{n} \mu_{i^\star} - \mu_{A_t}, \tag{12.0.2}$$

[1] The problem is called the bandit problem in the literature because we imagine a player in a casino, choosing between K different slot machines (hence a K-armed bandit, as this is a casino and the player will surely lose eventually), each with a different unknown reward distribution. The player wishes to put as much of his money as possible into the machine with the greatest expected reward.

where $i^\star \in \operatorname{argmax}_i \mu_i$ so $\mu_{i^\star} = \max_i \mu_i$. The regret Reg_n as defined is a random quantity, so we generally seek to give bounds on its expectation or high-probability guarantees on its value. In this chapter, we generally focus for simplicity on the expected regret,

$$\overline{\mathsf{Reg}}_n := \mathbb{E}\left[\sum_{t=1}^n \mu_{i^\star} - \mu_{A_t}\right], \tag{12.0.3}$$

where the expectation is taken over any randomness in the player's actions A_t and in the repeated observations of the random variables $Y_1, \ldots, Y_K$.

12.1 Confidence-based algorithms

A natural first strategy to consider is one based on confidence intervals with slight optimism. Roughly, if we believe the true mean μ_i for an arm i lies within $[\widehat{\mu}_i - c_i, \widehat{\mu}_i + c_i]$, where c_i is some interval (whose length decreases with time t), then we optimistically "believe" that the value of arm i is $\widehat{\mu}_i + c_i$; then at iteration t, as our action A_t we choose the arm whose optimistic mean is the highest, thus hoping to maximize our received reward.

This strategy lies at the heart of the Upper Confidence Bound (UCB) family of algorithms, due to [3], a simple variant of which we describe here. Before continuing, we recall the standard result on sub-Gaussian random variables of Corollary 3.9 in our context, though we require a somewhat more careful calculation because of the sequential nature of our process. Let $T_i(t) = \operatorname{card}\{\tau \le t : A_\tau = i\}$ denote the number of times that arm i has been pulled by time t of the bandit process. Then if we define

$$\widehat{\mu}_i(t) := \frac{1}{T_i(t)} \sum_{\tau \le t, A_\tau = i} Y_i(\tau),$$

to be the running average of the rewards of arm i at time t (computed only on those instances in which arm i was selected), we claim that for all i and all t,

$$\mathbb{P}\left(\widehat{\mu}_i(t) \ge \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{T_i(t)}}\right) \vee \mathbb{P}\left(\widehat{\mu}_i(t) \le \mu_i - \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{T_i(t)}}\right) \le \delta. \tag{12.1.1}$$

That is, so long as we pull the arms sufficiently many times, we are unlikely to pull the wrong arm. We prove the claim (12.1.1) in the appendix to this chapter.

Here then is the UCB procedure:

Input: Sub-gaussian parameter σ^2 and sequence of deviation probabilities $\delta_1, \delta_2, \ldots$.
Initialization: Play each arm $i = 1, \ldots, K$ once
Repeat: for each iteration t, play the arm maximizing

$$\widehat{\mu}_i(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}}.$$

Figure 12.1: The Upper Confidence Bound (UCB) Algorithm

If we define

$$\Delta_i := \mu_{i^\star} - \mu_i$$

to be the gap in means between the optimal arm and any sub-optimal arm, we then obtain the following guarantee on the expected number of pulls of any sub-optimal arm i after n steps.

Proposition 12.1. *Assume that each of the K arms is σ^2-sub-Gaussian and let the sequence $\delta_1 \ge \delta_2 \ge \cdots$ be non-increasing and positive. Then for any n and any arm $i \neq i^\star$,*

$$\mathbb{E}[T_i(n)] \le \left\lceil \frac{4\sigma^2 \log \frac{1}{\delta_n}}{\Delta_i^2} \right\rceil + 2\sum_{t=2}^{n} \delta_t.$$

Proof Without loss of generality, we assume arm 1 satisfies $\mu_1 = \max_i \mu_i$, and let arm i be any sub-optimal arm. The key insight is to carefully consider what occurs if we play arm i in the UCB procedure of Figure 12.1. In particular, if we play arm i at time t, then we certainly have

$$\widehat{\mu}_i(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}} \ge \widehat{\mu}_1(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_1(t)}}.$$

For this to occur, at least one of the following three events must occur (we suppress the dependence on i for each of them):

$$\mathcal{E}_1(t) := \left\{ \widehat{\mu}_i(t) \ge \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}} \right\}, \quad \mathcal{E}_2(t) := \left\{ \widehat{\mu}_1(t) \le \mu_1 - \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_1(t)}} \right\},$$

$$\mathcal{E}_3(t) := \left\{ \Delta_i \le 2\sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}} \right\}.$$

Indeed, suppose that none of the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ occur at time t. Then we have

$$\widehat{\mu}_i(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}} < \mu_i + 2\sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}} < \mu_i + \Delta_i = \mu_1 < \widehat{\mu}_1(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_1(t)}},$$

the inequalities following by $\mathcal{E}_1$, $\mathcal{E}_3$, and $\mathcal{E}_2$, respectively.

Now, for any $l \in \{1, \ldots, n\}$, we see that

$$\begin{aligned}\mathbb{E}[T_i(n)] = \sum_{t=1}^{n} \mathbb{E}[\mathbf{1}\{A_t = i\}] &= \sum_{t=1}^{n} \mathbb{E}[\mathbf{1}\{A_t = i, T_i(t) > l\} + \mathbf{1}\{A_t = i, T_i(t) \le l\}] \\ &\le l + \sum_{t=l+1}^{n} \mathbb{P}(A_t = i, T_i(t) > l).\end{aligned}$$

Now, we use that δ_t is non-increasing, and see that if we set

$$l^\star = \left\lceil 4\frac{\sigma^2 \log \frac{1}{\delta_n}}{\Delta_i^2} \right\rceil,$$

then to have $T_i(t) > l^\star$ it must be the case that $\mathcal{E}_3(t)$ cannot occur—that is, we would have $2\sqrt{\sigma^2 \log \frac{1}{\delta_t}/T_i(t)} > 2\sqrt{\sigma^2 \log \frac{1}{\delta_t}/l} \ge \Delta_i$. Thus we have

$$\begin{aligned}\mathbb{E}[T_i(n)] = \sum_{t=1}^{n} \mathbb{E}[\mathbf{1}\{A_t = i\}] &\le l^\star + \sum_{t=l^\star+1}^{n} \mathbb{P}(A_t = i, \mathcal{E}_3(t) \text{ fails}) \\ &\le l^\star + \sum_{t=l^\star+1}^{n} \mathbb{P}(\mathcal{E}_1(t) \text{ or } \mathcal{E}_2(t)) \le l^\star + \sum_{t=l^\star+1}^{n} 2\delta_t.\end{aligned}$$

This implies the desired result. □

Naturally, the number of times arm i is selected in the sequential game is related to the regret of a procedure; indeed, we have

$$\mathsf{Reg}_n = \sum_{t=1}^n (\mu_{i^\star} - \mu_{A_t}) = \sum_{i=1}^K (\mu_{i^\star} - \mu_i) T_i(n) = \sum_{i=1}^K \Delta_i T_i(n).$$

Using this identity, we immediately obtain two theorems on the (expected) regret of the UCB algorithm.

Theorem 12.2. *Let $\delta_t = \delta/t^2$ for all t. Then for any $n \in \mathbb{N}$ the UCB algorithm attains*

$$\overline{\mathsf{Reg}}_n \le \sum_{i \neq i^\star} \frac{4\sigma^2[2\log n - \log \delta]}{\Delta_i} + \frac{\pi^2 - 2}{3}\left(\sum_{i=1}^K \Delta_i\right)\delta + \sum_{i=1}^K \Delta_i.$$

Proof First, we note that

$$\mathbb{E}[\Delta_i T_i(n)] \le \Delta_i \left\lceil 4\sigma^2 \log \frac{1}{\delta_n} / \Delta_i^2 \right\rceil + 2\Delta_i \sum_{t=2}^n \frac{\delta}{t^2} \le \frac{4\sigma^2 \log \frac{1}{\delta_n}}{\Delta_i} + \Delta_i + 2\Delta_i \sum_{t=2}^n \frac{\delta}{t^2}$$

by Proposition 12.1. Summing over $i \neq i^\star$ and noting that $\sum_{t \ge 2} t^{-2} = \pi^2/6 - 1$ gives the result. □

Let us unpack the bound of Theorem 12.2 slightly. First, we make the simplifying assumption that $\delta_t = 1/t^2$ for all t, and let $\Delta = \min_{i \neq i^\star} \Delta_i$. In this case, we have expected regret bounded by

$$\overline{\mathsf{Reg}}_n \le 8\frac{K\sigma^2 \log n}{\Delta} + \frac{\pi^2 + 1}{3} \sum_{i=1}^K \Delta_i.$$

So we see that the asymptotic regret with this choice of δ scales as $(K\sigma^2/\Delta)\log n$, roughly linear in the classes, logarithmic in n, and inversely proportional to the gap in means. As a concrete example, if we know that the rewards for each arm Y_i belong to the interval $[0,1]$, then Hoeffding's lemma (recall Example 3.6) states that we may take $\sigma^2 = 1/4$. Thus the mean regret becomes at most $\sum_{i:\Delta_i > 0} \frac{2\log n}{\Delta_i}(1 + o(1))$, where the $o(1)$ term tends to zero as $n \to \infty$.

If we knew a bit more about our problem, then by optimizing over δ and choosing $\delta = \sigma^2/\Delta$, we obtain the upper bound

$$\overline{\mathsf{Reg}}_n \le O(1)\left[\frac{K\sigma^2}{\Delta}\log\frac{n\Delta}{\sigma^2} + K\frac{\max_i \Delta_i}{\min_i \Delta_i}\right], \tag{12.1.2}$$

that is, the expected regret scales asymptotically as $(K\sigma^2/\Delta)\log(\frac{n\Delta}{\sigma^2})$—linearly in the number of classes, logarithmically in n, and inversely proportional to the gap between the largest and other means.

If any of the gaps $\Delta_i \to 0$ in the bound of Theorem 12.2, the bound becomes vacuous—it simply says that the regret is upper bounded by infinity. Intuitively, however, pulling a *slightly* sub-optimal arm should be insignificant for the regret. With that in mind, we present a slight variant of the above bounds, which has a worse scaling with n—the bound scales as $\sqrt{n}$ rather than $\log n$—but is independent of the gaps Δ_i.

Theorem 12.3. *If UCB is run with parameter $\delta_t = 1/t^2$, then*

$$\overline{\mathsf{Reg}}_n \le \sqrt{8K\sigma^2 n \log n} + 4\sum_{i=1}^K \Delta_i.$$

Proof Fix any $\gamma > 0$. Then we may write the regret with the standard identity

$$\mathsf{Reg}_n = \sum_{i \neq i^\star} \Delta_i T_i(n) = \sum_{i:\Delta_i \ge \gamma} \Delta_i T_i(n) + \sum_{i:\Delta_i < \gamma} \Delta_i T_i(n) \le \sum_{i:\Delta_i \ge \gamma} \Delta_i T_i(n) + n\gamma,$$

where the final inequality uses that certainly $\sum_{i=1}^K T_i(n) \le n$. Taking expectations with our UCB procedure and $\delta = 1$, we have by Theorem 12.2 that

$$\overline{\mathsf{Reg}}_n \le \sum_{i:\Delta_i \ge \gamma} \Delta_i \frac{8\sigma^2 \log n}{\Delta_i^2} + \frac{\pi^2 + 1}{3}\sum_{i=1}^K \Delta_i + n\gamma \le K\frac{8\sigma^2 \log n}{\gamma} + n\gamma + \frac{\pi^2+1}{3}\sum_{i=1}^K \Delta_i,$$

Optimizing over γ by taking $\gamma = \frac{\sqrt{8K\sigma^2 \log n}}{\sqrt{n}}$ gives the result. □

Combining the above two theorems, we see that the UCB algorithm with parameters $\delta_t = 1/t^2$ automatically achieves the expected regret guarantee

$$\overline{\mathsf{Reg}}_n \le C \cdot \min\left\{\sum_{i:\Delta_i > 0} \frac{\sigma^2 \log n}{\Delta_i}, \sqrt{K\sigma^2 n \log n}\right\}. \tag{12.1.3}$$

That is, UCB enjoys some adaptive behavior. It is not, however, optimal; there are algorithms, including Audibert and Bubeck's MOSS (Minimax Optimal in the Stochastic Case) bandit procedure [2], which achieve regret

$$\overline{\mathsf{Reg}}_n \le C \cdot \min\left\{\sqrt{Kn}, \frac{K}{\Delta}\log\frac{n\Delta^2}{K}\right\},$$

which is essentially the bound specified by inequality (12.1.2) (which required knowledge of the Δ_is) and an improvement by $\log n$ over the analysis of Theorem 12.3. It is also possible to provie a high-probability guarantee for the UCB algorithms, which follows essentially immediately from the proof techniques of Proposition 12.1, but we leave this to the interested reader.

12.2 Bayesian approaches to bandits

The upper confidence bound procedure, while elegant and straightforward, has a variety of competitors, including online gradient descent approaches and a variety of Bayesian strategies. Bayesian strategies—because they (can) incorporate prior knowledge—have the advantage that they suggest policies for exploration and trading between regret and information; that is, they allow us to quantify a value for information. They often yield very simple procedures, allowing simpler implementations.

In this section, we thus consider the following specialized setting; there is substantially more possible here. We assume that there is a finite set of actions (arms) $\mathcal{A}$ as before, and we have a collection of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ parameterized by a set Θ (often, this is some subset of $\mathbb{R}^K$ when we look at K-armed bandit problems with $\text{card}(\mathcal{A}) = K$, but we stay in this abstract setting temporarily). We also have a loss function $\ell : \mathcal{A} \times \Theta \to \mathbb{R}$ that measure the quality of an action $a \in \mathcal{A}$ for the parameter θ.

Example 12.4 (Classical Bernoulli bandit problem)**:** The classical bandit problem, as in the UCB case of the previous section, has actions (arms) $\mathcal{A} = \{1, \ldots, K\}$, and the parameter space $\Theta = [0, 1]^K$, and we have that P_θ is a distribution on $Y \in \{0, 1\}^K$, where Y has independent coordinates $1, \ldots, K$ with $P(Y_j = 1) = \theta_j$, that is, $Y_j \sim \mathsf{Bernoulli}(\theta_j)$. The goal is to find the arm with highest mean reward, that is, $\text{argmax}_j\, \theta_j$, and thus possible loss functions include $\ell(a, \theta) = -\theta_a$ or, if we wish the loss to be positive, $\ell(a, \theta) = 1 - \theta_a \in [0, 1]$. ♣

Lastly, in this Bayesian setting, we require a prior distribution π on the space Θ, where $\pi(\Theta) = 1$. We then define the Bayesian regret as

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) = \mathbb{E}_\pi \left[\sum_{t=1}^{n} \ell(A_t, \theta) - \ell(A^\star, \theta) \right], \tag{12.2.1}$$

where $A^\star \in \text{argmin}_{a \in \mathcal{A}}\, \ell(a, \theta)$ is the minimizer of the loss, and $A_t \in \mathcal{A}$ is the action the player takes at time t of the process. The expectation (12.2.1) is taken both over the randomness in θ according to the prior π and any randomness in the player's strategy for choosing the actions A_t at each time.

Our approaches in this section build off of those in Chapter 9, except that we no longer fully observe the desired observations Y—we may only observe $Y_{A_t}(t)$ at time t, which may provide less information. The broad algorithmic framework for this section is as follows. We now give several concrete instantiations of this broad procedure, as well as tools (both information-theoretic and otherwise) for its analysis.

Input: Prior distribution π on space Θ, family of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$
Repeat: for each iteration t, choose distribution π_t on space Θ (based on history $Y_{A_1}(1), \ldots, Y_{A_{t-1}}(t-1)$). Draw

$$\theta_t \sim \pi_t.$$

Play action $A_t \in \mathcal{A}$ minimizing

$$\ell(a, \theta_t)$$

over $a \in \mathcal{A}$, observe $Y_{A_t}(t)$.

Figure 12.2: The generic Bayesian algorithm

12.2.1 Posterior (Thompson) sampling

The first strategy we consider is perhaps the simplest; in Algorithm 12.2, it corresponds to using π_t to be the posterior distribution on θ conditional on the history $Y_{A_1}(1), \ldots, Y_{A_{t-}}(t-1)$. That is, we let

$$\mathcal{H}_t := \{A_1, Y_{A_1}(1), A_2, Y_{A_2}(2), \ldots, A_t, Y_{A_t}(t)\}$$

denote the history (or the σ-field thereof) of the procedure and rewards up to time t. Then at iteration t, we use the posterior

$$\pi_t(\theta) = \pi(\theta \mid \mathcal{H}_{t-1}),$$

the distribution on θ conditional on $\mathcal{H}_{t-1}$. This procedure was originally proposed by Thompson [12] in 1933 in the first paper on bandit problems. There are several analyses of Thompson (and related Bayesian) procedures possible; our first analysis proceeds by using confidence bounds, while our later analyses give a more information theoretic analysis.

First, we provide a more concrete specification of Algorithm 12.2 for Thompson (posterior) sampling in the case of Bernoulli rewards.

Example 12.5 (Thompson sampling with Bernoulli penalities)**:** Let us suppose that the vector $\theta \in [0,1]^K$, and we draw $\theta_i \sim \text{Beta}(1,1)$, which corresponds to the uniform distribution on $[0,1]^d$. The actions available are simply to select one of the coordinates, $a \in \mathcal{A} = \{1, \ldots, K\}$, and we observe $Y_a \sim \mathsf{Bernoulli}(\theta_a)$, that is, $\mathbb{P}(Y_a = 1 \mid \theta) = \theta_a$. That is, $\ell(a, \theta) = \theta_a$. Let $T_a^1(t) = \text{card}\{\tau \le t : A_t = a, Y_a(\tau) = 1\}$ be the number of times arm a is pulled and results in a loss of 1 by time t, and similarly let $T_a^0(t) = \text{card}\{\tau \le t : A_t = a, Y_a(\tau) = 0\}$. Then, recalling Example 9.6 on Beta-Bernoulli distributions, Thompson sampling proceeds as follows:

(1) For each arm $a \in \mathcal{A} = \{1, \ldots, K\}$, draw $\theta_a(t) \sim \text{Beta}(1 + T_a^1(t), 1 + T_a^0(t))$.

(2) Play the action $A_t = \text{argmin}_a \, \theta_a(t)$.

(3) Observe the loss $Y_{A_t}(t) \in \{0, 1\}$, and increment the appropriate count.

Thompson sampling is simple in this case, and it is implementable with just a few counters. ♣

We may extend Example 12.5 to the case in which the losses come from any distribution with mean θ_i, so long as the distribution is supported on $[0,1]$. In particular, we have the following example.

Example 12.6 (Thompson sampling with bounded random losses)**:** Let us again consider the setting of Example 12.5, except that the observed losses $Y_a(t) \in [0,1]$ with $\mathbb{E}[Y_a \mid \theta] = \theta_a$. The following modification allows us to perform Thompson sampling in this case, even without knowing the distribution of $Y_a \mid \theta$: instead of observing a loss $Y_a \in \{0, 1\}$, we construct a random observation $\widetilde{Y}_a \in \{0, 1\}$ with the property that $\mathbb{P}(\widetilde{Y}_a = 1 \mid Y_a) = Y_a$. Then the losses $\ell(a, \theta) = \theta_a$ are identical, and the posterior distribution over θ is still a Beta distribution. We simply redefine

$$T_a^0(t) := \text{card}\{\tau \le t : A_t = a, \widetilde{Y}_a(\tau) = 0\} \;\; \text{and} \;\; T_a^1(t) := \text{card}\{\tau \le t : A_t = a, \widetilde{Y}_a(\tau) = 0\}.$$

The Thompson sampling procedure is otherwise identical. ♣

Our first analysis shows that Thompson sampling can guarantee performance similar to (or in some cases, better than) confidence-based procedures, which we do by using a sequence of (potential) lower and upper bounds on the losses of actions. (Recall we wish to minimize our losses, so that we would optimistically play those arms with the lowest estimated loss.) This analysis is based on that of Russo and Van Roy [9]. Let $L_t : \mathcal{A} \to \mathbb{R}$ and $U_t : \mathcal{A} \to \mathbb{R}$ be an arbitrary sequence of (random) functions that are measurable with respect to $\mathcal{H}_{t-1}$, that is, they are constructed based only on $\{A_1, Y_{A_1}(1), \ldots, A_{t-1}, Y_{A_{t-1}}(t-1)\}$. Then we can decompose the

Bayesian regret (12.2.1) as

$$\begin{aligned}\mathsf{Reg}_n(\mathcal{A}, \ell, \pi) &= \mathbb{E}_\pi\left[\sum_{t=1}^n \ell(A_t, \theta) - \ell(A^\star, \theta)\right] \qquad (12.2.2)\\ &= \sum_{t=1}^n \mathbb{E}_\pi[U_t(A_t) - L_t(A_t)] + \sum_{t=1}^n \mathbb{E}_\pi[\ell(A_t, \theta) - U_t(A_t)] + \sum_{t=1}^n \mathbb{E}_\pi[L_t(A_t) - \ell(A^\star, \theta)]\\ &\overset{(i)}{=} \sum_{t=1}^n \mathbb{E}_\pi[U_t(A_t) - L_t(A_t)] + \sum_{t=1}^n \mathbb{E}_\pi[\ell(A_t, \theta) - U_t(A_t)] + \sum_{t=1}^n \mathbb{E}_\pi[L_t(A_t^\star) - \ell(A_t^\star, \theta)],\end{aligned}$$

where in equality (i) we used that conditional on $\mathcal{H}_{t-1}$, A_t and $A_t^\star = A^\star$ have the same distribution, as we sample from the posterior $\pi(\theta \mid \mathcal{H}_{t-1})$, and L_t is a function of $\mathcal{H}_{t-1}$. With the decomposition (12.2.2) at hand, we may now provide an expected regret bound for Thompson (or posterior) sampling. We remark that the behavior of Thompson sampling is independent of these upper and lower bounds U_t, L_t we have chosen—they are simply an artifact to make analysis easier.

Theorem 12.7. *Suppose that conditional on the choice of action $A_t = a$, the received loss $Y_a(t)$ is σ^2-sub-Gaussian with mean $\ell(a, \theta)$, that is,*

$$\mathbb{E}\left[\exp\left(\lambda(Y_a(t) - \ell(a, \theta))\right) \mid \mathcal{H}_{t-1}\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \quad \text{for all } a \in \mathcal{A}.$$

Then for all $\delta \ge 0$ we have

$$\mathsf{Reg}_n(\mathcal{A}, \ell, \pi) \le 4\sqrt{2\sigma^2 \log\frac{1}{\delta}}\sqrt{|\mathcal{A}|n} + 3n\delta\sigma|\mathcal{A}|.$$

In particular, choosing $\delta = \frac{1}{n}$ gives

$$\mathsf{Reg}_n(\mathcal{A}, \ell, \pi) \le 6\sigma\sqrt{|\mathcal{A}|n \log n} + 3\sigma|\mathcal{A}|.$$

Proof We choose the upper and lower bound functions somewhat carefully so as to get a fairly sharp regret guarantee. In particular, we (as in our analysis of the UCB algorithm) let $\delta \in (0, 1)$ and define $T_a(t) := \text{card}\{\tau \le t : A_t = a\}$ to be the number of times that action a has been chosen by iteration t. Then we define the mean loss for action a at time t by

$$\widehat{\ell}_a(t) := \frac{1}{T_a(t)} \sum_{\tau \le t, A_\tau = a} Y_a(\tau)$$

and our bounds for the analysis by

$$U_t(a) := \widehat{\ell}_a(t) + \sqrt{\frac{2\sigma^2 \log\frac{1}{\delta}}{T_a(t)}} \quad \text{and} \quad L_t(a) := \widehat{\ell}_a(t) - \sqrt{\frac{2\sigma^2 \log\frac{1}{\delta}}{T_a(t)}}.$$

With these choices, we see that by the extension of the sub-Gaussian concentration bound (12.1.1) and the equality (12.A.1) showing that the sum $\sum_{\tau \le t, A_\tau = a} Y_a(\tau)$ is equal in distribution to the sum $\sum_{\tau \le t, A_\tau = a} Y_a'(\tau)$, where $Y_a'(\tau)$ are independent and identically distributed copies of $Y_a(\tau)$, we have for any $\epsilon \ge 0$ that

$$\mathbb{P}(U_t(a) \le \ell(a, \theta) - \epsilon \mid T_a(t)) \le \exp\left(-\frac{T_a(t)}{2\sigma^2}\left(\sqrt{\frac{2\sigma^2 \log\frac{1}{\delta}}{T_a(t)}} + \epsilon\right)^2\right) \le \exp\left(-\log\frac{1}{\delta} - \frac{T_a(t)\epsilon^2}{2\sigma^2}\right), \tag{12.2.3}$$

where the final inequality uses that $(a+b)^2 \ge a^2 + b^2$ for $ab \ge 0$. We have an identical bound for $\mathbb{P}(L_t(a) \ge \ell(a,\theta) + \epsilon \mid T_a(t))$.

We may now bound the final two sums in the regret expansion (12.2.2) using inequality (12.2.3). First, however, we make the observation that for any nonnegative random variable Z, we have $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \ge \epsilon) d\epsilon$. Using this, we have

$$\begin{aligned}
\sum_{t=1}^n \mathbb{E}_\pi\left[\ell(A_t,\theta) - U_t(A_t)\right] &\le \sum_{t=1}^n \sum_{a\in\mathcal{A}} \mathbb{E}_\pi\left[\left[\ell(a,\theta) - U_t(a)\right]_+\right] \\
&= \sum_{t=1}^n \sum_{a\in\mathcal{A}} \mathbb{E}_\pi\left[\int_0^\infty \mathbb{P}(U_t(a) \ge \ell(a,\theta) + \epsilon \mid T_a(t)) d\epsilon\right] \\
&\overset{(i)}{\le} \sum_{t=1}^n \sum_{a\in\mathcal{A}} \delta \mathbb{E}_\pi\left[\int_0^\infty \exp\left(-\frac{T_a(t)\epsilon^2}{2\sigma^2}\right) d\epsilon\right] \overset{(ii)}{=} \sum_{t=1}^n \delta \sum_{a\in\mathcal{A}} \mathbb{E}_\pi\left[\sqrt{\frac{\pi\sigma^2}{2T_a(t)}}\right],
\end{aligned}$$

where inequality (i) uses the bound (12.2.3) and equality (ii) uses that this is the integral of half of a normal density. Substituting this bound, as well as the identical one for the terms involving $L_t(A_t^\star)$, into the decomposition (12.2.2) yields

$$\mathsf{Reg}_n(\mathcal{A},\ell,\pi) \le \sum_{t=1}^n \mathbb{E}_\pi[U_t(A_t) - L_t(A_t)] + \sum_{t=1}^n \delta \sum_{a\in\mathcal{A}} \mathbb{E}_\pi\left[\sqrt{\frac{2\pi\sigma^2}{T_a(t)}}\right].$$

Using that $T_a(t) \ge 1$ for each action a, we have $\sum_{a\in\mathcal{A}} \mathbb{E}_\pi[\sqrt{2\pi\sigma^2/T_a(t)}] < 3\sigma|\mathcal{A}|$. Lastly, we use that

$$U_t(A_t) - L_t(A_t) = 2\sqrt{\frac{2\sigma^2 \log\frac{1}{\delta}}{T_{A_t}(t)}}.$$

Thus we have

$$\sum_{t=1}^n \mathbb{E}_\pi[U_t(A_t) - L_t(A_t)] = 2\sqrt{2\sigma^2 \log\frac{1}{\delta}} \sum_{a\in\mathcal{A}} \mathbb{E}_\pi\left[\sum_{t:A_t=a} \frac{1}{\sqrt{T_a(t)}}\right].$$

Once we see that $\sum_{t=1}^T t^{-\frac{1}{2}} \le \int_0^T t^{-\frac{1}{2}} dt = 2\sqrt{T}$, we have the upper bound

$$\mathsf{Reg}_n(\mathcal{A},\ell,\pi) \le 4\sqrt{2\sigma^2 \log\frac{1}{\delta}} \sum_{a\in\mathcal{A}} \mathbb{E}_\pi[\sqrt{T_a(n)}] + 3n\delta\sigma|\mathcal{A}|.$$

As $\sum_{a\in\mathcal{A}} T_a(n) = n$, the Cauchy-Scwharz inequality implies $\sum_{a\in\mathcal{A}} \sqrt{T_a(n)} \le \sqrt{|\mathcal{A}|n}$, which gives the result. □

An immediate Corollary of Theorem 12.7 is the following result, which applies in the case of bounded losses Y_a as in Examples 12.5 and 12.6.

Corollary 12.8. *Let the losses $Y_a \in [0,1]$ with $\mathbb{E}[Y_a \mid \theta] = \theta_a$, where $\theta_i \overset{\text{i.i.d.}}{\sim}$ Beta$(1,1)$ for $i = 1,\ldots,K$. Then Thompson sampling satisfies*

$$\mathsf{Reg}_n(\mathcal{A},\ell,\pi) \le 3\sqrt{Kn\log n} + \frac{3}{2}K.$$

12.2.2 An information-theoretic analysis

12.2.3 Information and exploration

12.3 Online gradient descent approaches

It is also possible to use online gradient descent approaches to minimize regret in the more standard multi-armed bandit setting. In this scenario, our goal is to minimize a sequentially (partially) observed loss, as in the previous section. In this case, as usual we have K arms with non-negative means $\mu_1, \ldots, \mu_K$, and we wish to find the arm with lowest mean loss. We build off of the online convex optimization procedures of Chapter 11 to achieve good regret guarantees. In particular, at each step of the bandit procedure, we play a distribution $w_t \in \Delta_K$ on the arms, and then we select one arm j at random, each with probability $w_{t,j}$. The *expected* loss we suffer is then $\ell_t(w_t) = \langle w_t, \mu \rangle$, though we observe only a random realization of the loss for the arm a that we play.

Because of its natural connections with estimation of probability distributions, we would like to use the exponentiated gradient algorithm, Example 11.6, to play this game. We face one main difficulty: we must estimate the gradient of the losses, $\nabla \ell_t(w_t) = \mu$, even though we only observe a random variable $Y_a(t) \in \mathbb{R}_+$, conditional on selecting action $A_t = a$ at time t, with the property that $\mathbb{E}[Y_a(t)] = \mu_a$. Happily, we can construct such an estimate without too much additional variance.

Lemma 12.9. *Let $Y \in \mathbb{R}^K$ be a random variable with $\mathbb{E}[Y] = \mu$ and $w \in \Delta_K$ be a probability vector. Choose a coordinate a with probability w_a and define the random vector*

$$\widetilde{Y}_j = \begin{cases} Y_j / w_j & \text{if } j = a \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathbb{E}[\widetilde{Y} \mid Y] = Y$.

Proof The proof is immediate: for each coordinate j of $\widetilde{Y}$, we have $\mathbb{E}[\widetilde{Y}_j \mid Y] = w_j Y_j / w_j = Y_j$. □

Lemma 12.9 suggests the following procedure, which gives rise to (a variant of) Auer et al.'s EXP3 (Exponentiated gradient for Exploration and Exploitation) algorithm [4]. We can prove

```
Input: stepsize parameter $\eta$, initial vector $w_1 = [\frac{1}{K} \; \cdots \; \frac{1}{K}]^\top$
Repeat: for each iteration $t$, choose random action $A_t = a$ with probability $w_{t,a}$
Receive non-negative loss $Y_a(t)$, and define
    $g_{t,j} = \begin{cases} Y_j(t)/w_j & \text{if } A_t = j \\ 0 & \text{otherwise.} \end{cases}$
Update for each $i = 1, \ldots, K$
    $w_{t+1,i} = \frac{w_{t,i} \exp(-\eta g_{t,i})}{\sum_j w_{t,j} \exp(-\eta g_{t,j})}.$
```

Figure 12.3: Exponentiated gradient for bandit problems.

the following bound on the expected regret of the EXP3 Algorithm 12.3 by leveraging our refined analysis of exponentiated gradients in Proposition 11.13.

Proposition 12.10. *Assume that for each j, we have $\mathbb{E}[Y_j^2] \le \sigma^2$ and the observed loss $Y_j \ge 0$. Then Alg. 12.3 attains expected regret (we are minimizing)*

$$\mathsf{Reg}_n = \sum_{t=1}^n \mathbb{E}[\mu_{A_t} - \mu_{i^\star}] \le \frac{\log K}{\eta} + \frac{\eta}{2}\sigma^2 K n.$$

In particular, choosing $\eta = \sqrt{\log K/(K\sigma^2 n)}$ gives

$$\mathsf{Reg}_n = \sum_{t=1}^n \mathbb{E}[\mu_{A_t} - \mu_{i^\star}] \le \frac{3}{2}\sigma\sqrt{Kn\log K}.$$

Proof With Lemma 12.9 in place, we recall the refined regret bound of Proposition 11.13. We have that for $w^\star \in \Delta_K$ and any sequence of vectors $g_1, g_2, \ldots$ with $g_t \in \mathbb{R}_+^K$, then exponentiated gradient descent achieves

$$\sum_{t=1}^n \langle g_t, w_t - w^\star \rangle \le \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^n \sum_{j=1}^k w_{t,j} g_{t,j}^2.$$

To transform this into a useful bound, we take expectations. Indeed, we have

$$\mathbb{E}[g_t \mid w_t] = \mathbb{E}[Y] = \mu$$

by construction, and we also have

$$\mathbb{E}\left[\sum_{j=1}^k w_{t,j} g_{t,j}^2 \mid w_t\right] = \sum_{j=1}^K w_{t,j}^2 \mathbb{E}[Y_j(t)^2/w_{t,j}^2 \mid w_t] = \sum_{j=1}^K \mathbb{E}[Y_j^2] = \mathbb{E}[\|Y\|_2^2].$$

This careful normalizing, allowed by Proposition 11.13, is essential to our analysis (and fails for more naive applications of online convex optimization bounds). In particular, we have

$$\mathsf{Reg}_n = \sum_{t=1}^n \mathbb{E}[\langle \mu, w_t - w^\star \rangle] = \sum_{t=1}^n \mathbb{E}[\langle g_t, w_t - w^\star \rangle] \le \frac{\log K}{\eta} + \frac{\eta}{2} n \mathbb{E}[\|Y\|_2^2].$$

Taking expectations gives the result. □

When the random observed losses $Y_a(t)$ are bounded in $[0,1]$, then we have the mean regret bound $\frac{3}{2}\sqrt{Kn\log K}$, which is as sharp as any of our other bounds.

12.4 Further notes and references

An extraordinarily abbreviated bibliography follows.

The golden oldies: Thompson [12], Robbins [8], and Lai and Robbins [7].

More recent work in machine learning (there are far too many references to list): the books Cesa-Bianchi and Lugosi [6] and Bubeck and Cesa-Bianchi [5] are good references. The papers of Auer et al. [4] and Auer et al. [3] introduced UCB and EXP3.

Our approach to Bayesian bandits follows Russo and Van Roy [9, 10, 11]. More advanced techniques allow Thompson sampling to apply even when the prior is unknown (e.g. Agrawal and Goyal [1]).

12.A Technical proofs

12.A.1 Proof of Claim (12.1.1)

We let $Y_i'(\tau)$, for $\tau = 1, 2, \ldots$, be independent and identically distributed copies of the random variables $Y_i(\tau)$, so that $Y_i'(\tau)$ is also independent of $T_i(t)$ for all t and τ. We claim that the pairs

$$(\widehat{\mu}_i(t), T_i(t)) \overset{\text{dist}}{=} \left(\widehat{\mu}_i'(t), T_i(t)\right), \tag{12.A.1}$$

where $\widehat{\mu}_i'(t) = \frac{1}{T_i(t)} \sum_{\tau : A_\tau = i} Y_i'(\tau)$ is the empirical mean of the copies $Y_i'(\tau)$ for those steps when arm i is selected. To see this, we use the standard fact that the characteristic function of a random variable completely characterizes the random variable. Let $\varphi_{Y_i}(\lambda) = \mathbb{E}[e^{\iota\lambda Y_i}]$, where $\iota = \sqrt{-1}$ is the imaginary unit, denote the characteristic function of Y_i, noting that by construction we have $\varphi_{Y_i} = \varphi_{Y_i'}$. Then writing the joint characteristic function of $T_i(t)\widehat{\mu}_i(t)$ and $T_i(t)$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\iota\lambda_1 \sum_{\tau=1}^t \mathbf{1}\left\{A_\tau = i\right\} Y_i(\tau) + \iota\lambda_2 T_i(t)\right)\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\prod_{\tau=1}^t \mathbb{E}\left[\exp\left(\iota\lambda_1 \mathbf{1}\left\{A_\tau = i\right\} Y_i(\tau) + \iota\lambda_2 \mathbf{1}\left\{A_\tau = i\right\}\right) \mid \mathcal{H}_{\tau-1}\right]\right] \\
&\overset{(ii)}{=} \mathbb{E}\left[\prod_{\tau=1}^t \Big(\mathbf{1}\left\{A_\tau = i\right\} e^{\iota\lambda_2} \mathbb{E}\left[\exp(\iota\lambda_1 Y_i(\tau)) \mid \mathcal{H}_{\tau-1}\right] + \mathbf{1}\left\{A_\tau \neq i\right\}\Big)\right] \\
&\overset{(iii)}{=} \mathbb{E}\left[\prod_{\tau=1}^t \Big(\mathbf{1}\left\{A_\tau = i\right\} e^{\lambda_2\iota} \varphi_{Y_i}(\lambda_1) + \mathbf{1}\left\{A_\tau \neq i\right\}\Big)\right] \\
&\overset{(iv)}{=} \mathbb{E}\left[\prod_{\tau=1}^t \Big(\mathbf{1}\left\{A_\tau = i\right\} e^{\lambda_2\iota} \varphi_{Y_i'}(\lambda_1) + \mathbf{1}\left\{A_\tau \neq i\right\}\Big)\right] \\
&= \mathbb{E}\left[\exp\left(\iota\lambda_1 \sum_{\tau=1}^t \mathbf{1}\left\{A_\tau = i\right\} Y_i'(\tau) + \iota\lambda_2 T_i(t)\right)\right],
\end{aligned}$$

where equality (i) is the usual tower property of conditional expectations, where $\mathcal{H}_{\tau-1}$ denotes the history to time $\tau - 1$, equality (ii) because $A_\tau \in \mathcal{H}_{\tau-1}$ (that is, it is a function of the history), equality (iii) follows because $Y_i(\tau)$ is independent of $\mathcal{H}_{\tau-1}$, and equality (iv) follows because Y_i' and Y_i have identical distributions. The final step is simply reversing the steps.

With the distributional equality (12.A.1) in place, we see that for any $\delta \in [0, 1]$, we have

$$\begin{aligned}
\mathbb{P}\left(\widehat{\mu}_i(t) \ge \mu_i + \sqrt{\frac{\sigma^2 \log\frac{1}{\delta}}{T_i(t)}}\right) = \mathbb{P}\left(\widehat{\mu}_i'(t) \ge \mu_i + \sqrt{\frac{\sigma^2 \log\frac{1}{\delta}}{T_i(t)}}\right) &= \mathbb{P}\left(\widehat{\mu}_i'(t) \ge \mu_i + \sqrt{\frac{\sigma^2 \log\frac{1}{\delta}}{T_i(t)}}\right) \\
&= \sum_{s=1}^t \mathbb{P}\left(\widehat{\mu}_i'(t) \ge \mu_i + \sqrt{\frac{\sigma^2 \log\frac{1}{\delta}}{s}} \mid T_i(t) = s\right) \mathbb{P}(T_i(t) = s) \\
&\le \sum_{s=1}^t \delta \mathbb{P}(T_i(t) = s) = \delta.
\end{aligned}$$

The proof for the lower tail is similar.

Bibliography

[1] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the Twenty Fifth Annual Conference on Computational Learning Theory*, 2012.

[2] J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. In *Journal of Machine Learning Research*, pages 2635–2686, 2010.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002. ISSN 0885-6125.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

[5] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[6] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[7] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[8] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, 55:527–535, 1952.

[9] D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, page To appear, 2014.

[10] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems 27*, 2014.

[11] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[12] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Chapter 13

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."

13.1 Basic framework and minimax risk

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

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

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

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

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

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

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

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

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

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\Phi\big(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta(P))\big)\right], \tag{13.1.1}$$

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

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

$$R_P(\theta) := \mathbb{E}_P[\ell(\theta; X)] = \int_{\mathcal{X}} \ell(\theta; x) dP(x) \tag{13.1.2}$$

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

[1] 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 13.1 (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 $\mathrm{sign}(\langle \theta, x \rangle) = y$. A convex loss surrogate for this problem is the hinge loss $\ell(\theta; z) = [1 - y \langle \theta, x \rangle]_+$; minimizing the associated risk functional (13.1.2) over a set $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq r\}$ gives the support vector machine [5]. ♣

Example 13.2 (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 (13.1.2) 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 $\widehat{\theta} : \mathcal{X}^n \to \Theta$ given access to a sample $X_1, \ldots, X_n$, we may define the associated maximum *excess risk* for the family $\mathcal{P}$ by

$$\sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[R_P(\widehat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} R(\theta) \right\},$$

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

$$\mathfrak{M}_n(\Theta, \mathcal{P}, \ell) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[R_P(\widehat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} R_P(\theta) \right\}, \tag{13.1.3}$$

where the infimum is taken over all estimators $\widehat{\theta} : \mathcal{X}^n \to \Theta$ and the risk R_P is implicitly defined in terms of the loss ℓ. The techniques for providing lower bounds for the minimax risk (13.1.1) or the excess risk (13.1.3) are essentially identical; we focus for the remainder of this section on techniques for providing lower bounds on the minimax risk.

13.2 Preliminaries on methods for lower bounds

There are a variety of techniques for providing lower bounds on the minimax risk (13.1.1). 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 π be any probability mass function over v. Then for any estimator $\widehat{\theta}$, the maximum risk has lower bound

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi(\rho(\widehat{\theta}(X_1^n), \theta(P))) \right] \ge \sum_v \pi(v) \mathbb{E}_{P_v} \left[\Phi(\rho(\widehat{\theta}(X_1^n), \theta(P_v))) \right].$$

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

13.2.1 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 $\mathcal{V}$ of finite cardinality, consider a family of distributions $\{P_v\}_{v \in \mathcal{V}}$ contained within $\mathcal{P}$. This family induces a collection of parameters $\{\theta(P_v)\}_{v \in \mathcal{V}}$; we call the family a 2δ-packing in the ρ-semimetric if

$$\rho(\theta(P_v), \theta(P_{v'})) \ge 2\delta \quad \text{for all } v \ne v'.$$

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

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

Given the observed sample X, the goal is to determine the value of the underlying index v. We refer to any measurable mapping $\Psi : \mathcal{X}^n \to \mathcal{V}$ as a test function. Its associated error probability is $\mathbb{P}(\Psi(X_1^n) \ne V)$, where $\mathbb{P}$ denotes the joint distribution over the random index V and X. In particular, if we set $\overline{P} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v$ to be the mixture distribution, then the sample X is drawn (marginally) from $\overline{P}$, and our hypothesis testing problem is to determine the randomly chosen index V given a sample from this mixture $\overline{P}$.

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

Proposition 13.3. *The minimax error* (13.1.1) *has lower bound*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \ge \Phi(\delta) \inf_{\Psi} \mathbb{P}(\Psi(X_1, \ldots, X_n) \ne V), \tag{13.2.1}$$

where the infimum ranges over all testing functions.

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

$$\mathbb{E}[\Phi(\rho(\widehat{\theta}, \theta))] \ge \mathbb{E}\left[\Phi(\delta) \mathbf{1}\left\{\rho(\widehat{\theta}, \theta) \ge \delta\right\}\right] = \Phi(\delta) \mathbb{P}(\rho(\widehat{\theta}, \theta) \ge \delta),$$

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

$$\Psi(\widehat{\theta}) := \operatorname*{argmin}_{v \in \mathcal{V}} \{\rho(\widehat{\theta}, \theta_v)\},$$

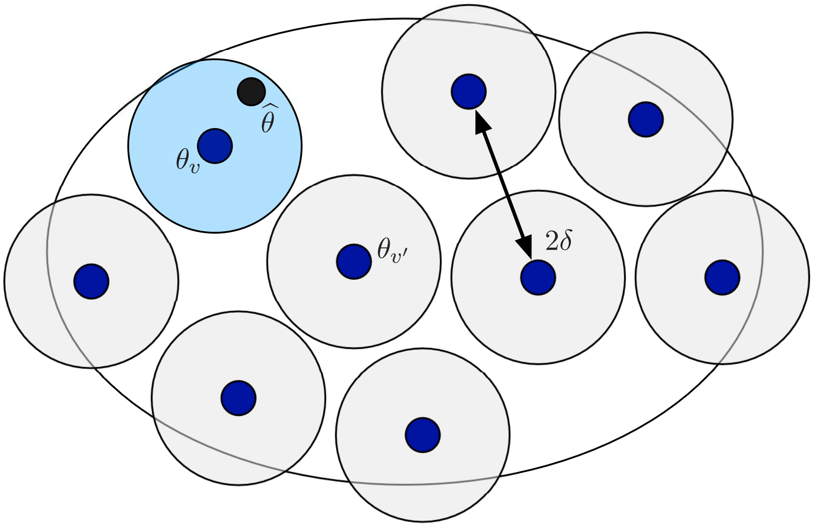

Figure 13.1. Example of a 2δ-packing of a set. The estimate $\widehat{\theta}$ is contained in at most one of the δ-balls around the points θ_v.

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

$$\rho(\widehat{\theta}, \theta_{v'}) \geq \rho(\theta_v, \theta_{v'}) - \rho(\widehat{\theta}, \theta_v) > 2\delta - \delta = \delta.$$

The test must thus return v as claimed. Equivalently, for $v \in \mathcal{V}$, the inequality $\Psi(\widehat{\theta}) \neq v$ implies $\rho(\widehat{\theta}, \theta_v) \geq \delta$. (See Figure 13.1.) By averaging over $\mathcal{V}$, we find that

$$\sup_P \mathbb{P}(\rho(\widehat{\theta}, \theta(P)) \geq \delta) \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{P}(\rho(\widehat{\theta}, \theta(P_v)) \geq \delta \mid V = v) \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{P}(\Psi(\widehat{\theta}) \neq v \mid V = v).$$

Taking an infimum over all tests $\Psi : \mathcal{X}^n \to V$ gives inequality (13.2.1). □

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

13.2.2 Inequalities between divergences and product distributions

We now present a few inequalities, and their consequences when applied to product distributions, that will be quite useful for proving our lower bounds. The three divergences we relate are the total variation distance, Kullback-Leibler divergence, and Hellinger distance, all of which are instances of f-divergences (recall Section 2.2.3). We first recall the definitions of the three when applied to distributions P, Q on a set $\mathcal{X}$, which we assume have densities p, q with respect to a base measure μ. Then we recall the total variation distance (2.2.4) is

$$\|P - Q\|_{\rm TV} := \sup_{A \subset \mathcal{X}} |P(A) - Q(A)| = \frac{1}{2} \int |p(x) - q(x)| d\mu(x),$$

which is the f-divergence $D_f(P\|Q)$ generated by $f(t) = \frac{1}{2}|t-1|$. The Hellinger distance (2.2.6) is

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

which is the f-divergence $D_f(P\|Q)$ generated by $f(t) = (\sqrt{t} - 1)^2$. We also recall the Kullback-Leibler (KL) divergence

$$D_{\rm kl}(P\|Q) := \int p(x) \log \frac{p(x)}{q(x)} d\mu(x), \tag{13.2.2}$$

which is the f-divergence $D_f(P\|Q)$ generated by $f(t) = t \log t$. As noted in Section 2.2.3, Proposition 2.10, these divergences have the following relationships.

Proposition (Proposition 2.10, restated)**.** *The total variation distance satisfies the following relationships:*

(a) For the Hellinger distance,

$$\frac{1}{2} d_{\rm hel}(P, Q)^2 \le \|P - Q\|_{\rm TV} \le d_{\rm hel}(P, Q) \sqrt{1 - d_{\rm hel}(P, Q)^2/4}.$$

(b) Pinsker's inequality: for any distributions P, Q,

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

We now show how Proposition 2.10 is useful, because KL-divergence and Hellinger distance both are easier to manipulate on product distributions than is total variation. Specifically, consider the product distributions $P = P_1 \times \cdots \times P_n$ and $Q = Q_1 \times \cdots \times Q_n$. Then the KL-divergence satisfies the decoupling equality

$$D_{\rm kl}(P\|Q) = \sum_{i=1}^n D_{\rm kl}(P_i\|Q_i), \tag{13.2.3}$$

while the Hellinger distance satisfies

$$\begin{aligned}
d_{\rm hel}(P, Q)^2 &= \int \left(\sqrt{p_1(x_1) \cdots p_n(x_n)} - \sqrt{q_1(x_1) \cdots q_n(x_n)} \right)^2 d\mu(x_1^n) \\
&= \int \left(\prod_{i=1}^n p_i(x_i) + \prod_{i=1}^n q_i(x_i) - 2\sqrt{p_1(x_1) \cdots p_n(x_n) q_1(x_n) \cdots q_n(x_n)} \right) d\mu(x_1^n) \\
&= 2 - 2 \prod_{i=1}^n \int \sqrt{p_i(x) q_i(x)} d\mu(x) = 2 - 2 \prod_{i=1}^n \left(1 - \frac{1}{2} d_{\rm hel}(P_i, Q_i)^2 \right). \qquad (13.2.4)
\end{aligned}$$

In particular, we see that for product distributions P^n and Q^n, Proposition 2.10 implies that

$$\|P^n - Q^n\|_{\mathrm{TV}}^2 \le \frac{1}{2} D_{\mathrm{kl}}(P^n \| Q^n) = \frac{n}{2} D_{\mathrm{kl}}(P \| Q)$$

and

$$\|P^n - Q^n\|_{\mathrm{TV}} \le d_{\mathrm{hel}}(P^n, Q^n) \le \sqrt{2 - 2(1 - d_{\mathrm{hel}}(P, Q)^2)^n}.$$

As a consequence, if we can guarantee that $D_{\mathrm{kl}}(P \| Q) \le 1/n$ or $d_{\mathrm{hel}}(P, Q) \le 1/\sqrt{n}$, then we guarantee the strict inequality $\|P^n - Q^n\|_{\mathrm{TV}} \le 1 - c$ for a fixed constant $c > 0$, for any n. We will see how this type of guarantee can be used to prove minimax lower bounds in the following sections.

13.2.3 Metric entropy and packing numbers

The second part of proving our lower bounds involves the construction of the packing set in Section 13.2.1. The size of the space Θ of parameters associated with our estimation problem—and consequently, how many parameters we can pack into it—is strongly coupled with the difficulty of estimation. Given a non-empty set Θ with associated (semi)metric ρ, 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 13.1 (Covering number). *Let Θ be a set with (semi)metric ρ. A* δ-cover *of the set Θ with respect to ρ is a set $\{\theta_1, \ldots, \theta_N\}$ such that for any point $\theta \in \Theta$, there exists some $v \in \{1, \ldots, N\}$ such that $\rho(\theta, \theta_v) \le \delta$. The* δ-covering number *of Θ is*

$$N(\delta, \Theta, \rho) := \inf \left\{ N \in \mathbb{N} \;:\; \textit{there exists a } \delta\textit{-cover } \theta_1, \ldots, \theta_N \textit{ of } \Theta \right\}.$$

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

Definition 13.2 (Packing number). *A* δ-packing *of the set Θ with respect to ρ is a set $\{\theta_1, \ldots, \theta_M\}$ such that for all distinct $v, v' \in \{1, \ldots, M\}$, we have $\rho(\theta_v, \theta_{v'}) \ge \delta$. The* δ-packing number *of Θ is*

$$M(\delta, \Theta, \rho) := \sup \left\{ M \in \mathbb{N} \;:\; \textit{there exists a } \delta\textit{-packing } \theta_1, \ldots, \theta_M \textit{ of } \Theta \right\}.$$

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

Lemma 13.4. *The packing and covering numbers satisfy the following inequalities:*

$$M(2\delta, \Theta, \rho) \le N(\delta, \Theta, \rho) \le M(\delta, \Theta, \rho).$$

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 δ. As a simple example, we see for any interval $[a, b]$ on the real line that in the usual absolute distance metric, $N(\delta, [a, b], |\cdot|) \asymp (b - a)/\delta$.

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 13.5 (Gilbert-Varshamov bound)**.** *Let $d \ge 1$. There is a subset $\mathcal{V}$ of the d-dimensional hypercube $\mathcal{H}_d = \{-1, 1\}^d$ of size $|\mathcal{V}| \ge \exp(d/8)$ such that the ℓ_1-distance*

$$\left\|v - v'\right\|_1 = 2 \sum_{j=1}^{d} \mathbf{1}\left\{v_j \neq v'{}_j\right\} \ge \frac{d}{2}$$

for all $v \neq v'$ with $v, v' \in \mathcal{V}$.

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

$$\left\|v - v'\right\|_1 \ge d/2 \;\text{ for all distinct } v, v' \in \mathcal{V}. \tag{13.2.5}$$

That is, the addition of any vector $w \in \mathcal{H}_d, w \notin \mathcal{V}$ to $\mathcal{V}$ will break the constraint (13.2.5). This means that if we construct the closed balls $B(v, d/2) := \{w \in \mathcal{H}_d : \|v - w\|_1 \le d/2\}$, we must have

$$\bigcup_{v \in \mathcal{V}} B(v, d/2) = \mathcal{H}_d \quad \text{so} \quad |\mathcal{V}||B(0, d/2)| = \sum_{v \in \mathcal{V}} |B(v, d/2)| \ge 2^d. \tag{13.2.6}$$

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

$$\begin{aligned} 2^{-d}|B(v, d/2)| &= \mathbb{P}(S_1 + S_2 + \ldots + S_d \le d/4) = \mathbb{P}(S_1 + S_2 + \ldots + S_d \ge 3d/4) \\ &\le \mathbb{E}\left[\exp(\lambda S_1 + \ldots + \lambda S_d)\right] \exp(-3\lambda d/4) \end{aligned}$$

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

$$2^{-d}|B(v, d/2)| \le \inf_{\lambda \ge 0} \left\{2^{-d}(1 + e^\lambda)^d \exp(-3\lambda d/4)\right\}$$

Choosing $\lambda = \log 3$, we have

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

Recalling inequality (13.2.6), we have

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

as claimed. □

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

Lemma 13.6. *Let $\mathbb{B}$ denote the unit $\|\cdot\|$-ball in $\mathbb{R}^d$. Then*

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

As a consequence of Lemma 13.6, we see that for any $\delta < 1$, there is a packing $\mathcal{V}$ of $\mathbb{B}$ such that $\|v - v'\| \ge \delta$ for all distinct $v, v' \in \mathcal{V}$ and $|\mathcal{V}| \ge (1/\delta)^d$, because we know $M(\delta, \mathbb{B}, \|\cdot\|) \ge N(\delta, \mathbb{B}, \|\cdot\|)$ as in Lemma 13.4. 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 δ-cover of $\mathbb{B}$, then

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

In particular, $N \ge \delta^{-d}$. For the upper bound on $N(\delta, \mathbb{B}, \|\cdot\|)$, let $\mathcal{V}$ be a δ-packing of $\mathbb{B}$ with maximal cardinality, so that $|\mathcal{V}| = M(\delta, \mathbb{B}, \|\cdot\|) \ge N(\delta, \mathbb{B}, \|\cdot\|)$ (recall Lemma 13.4). Notably, the collection of δ-balls $\{\delta\mathbb{B} + v_i\}_{i=1}^M$ cover the ball $\mathbb{B}$ (as otherwise, we could put an additional element in the packing $\mathcal{V}$), and moreover, the balls $\{\frac{\delta}{2}\mathbb{B} + v_i\}$ are all disjoint by definition of a packing. Consequently, we find that

$$M\left(\frac{\delta}{2}\right)^d \mathrm{Vol}(\mathbb{B}) = M \,\mathrm{Vol}\left(\frac{\delta}{2}\mathbb{B}\right) \le \mathrm{Vol}\left(\mathbb{B} + \frac{\delta}{2}\mathbb{B}\right) = \left(1 + \frac{\delta}{2}\right)^d \mathrm{Vol}(\mathbb{B}).$$

Rewriting, we obtain

$$M(\delta, \mathbb{B}, \|\cdot\|) \le \left(\frac{2}{\delta}\right)^d \left(1 + \frac{\delta}{2}\right)^d \frac{\mathrm{Vol}(\mathbb{B})}{\mathrm{Vol}(\mathbb{B})} = \left(1 + \frac{2}{\delta}\right)^d,$$

completing the proof. □

13.3 Le Cam's method

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

In the first homework, we considered several representations of the total variation distance, including a question showing its relation to optimal testing. We begin again with this strand of thought, recalling the general testing problem discussed in Section 13.2.1. 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

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

Recalling Section 13.2.1, we note that Proposition 2.11 gives an exact representation of the testing error using total variation distance. In particular, we have

Proposition (Proposition 2.11, restated)**.** *For any distributions P_1 and P_2 on $\mathcal{X}$, we have*

$$\inf_{\Psi} \left\{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\right\} = 1 - \|P_1 - P_2\|_{\rm TV}, \tag{13.3.1}$$

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

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

$$\inf_{\Psi} \mathbb{P}\left(\Psi(X_1, \ldots, X_n) \neq V\right) = \frac{1}{2} - \frac{1}{2}\|P_1^n - P_2^n\|_{\rm TV}. \tag{13.3.2}$$

The representations (13.3.1) and (13.3.2), in conjunction with our reduction of estimation to testing in Proposition 13.3, 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

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta)\left[\frac{1}{2} - \frac{1}{2}\|P_1^n - P_2^n\|_{\rm TV}\right]. \tag{13.3.3}$$

The lower bound (13.3.3) suggests the following strategy: we find distributions P_1 and P_2, which we choose as a function of δ, that guarantee $\|P_1^n - P_2^n\|_{\rm TV} \leq \frac{1}{2}$. In this case, so long as $\rho(\theta(P_1), \theta(P_2)) \geq 2\delta$, we have the lower bound

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

We now give an example illustrating this idea.

> **Example 13.7** (Bernoulli mean estimation)**:** Consider the problem of estimating the mean $\theta \in [-1, 1]$ of a $\{\pm 1\}$-valued Bernoulli distribution under the squared error loss $(\theta - \widehat{\theta})^2$, where $X_i \in \{-1, 1\}$. In this case, by fixing some $\delta > 0$, we set $\mathcal{V} = \{-1, 1\}$, and we define P_v so that
>
> $$P_v(X = 1) = \frac{1 + v\delta}{2} \quad \text{and} \quad P_v(X = -1) = \frac{1 - v\delta}{2},$$
>
> whence we see that the mean $\theta(P_v) = \delta v$. Using the metric $\rho(\theta, \theta') = |\theta - \theta'|$ and loss $\Phi(\delta) = \delta^2$, we have separation 2δ of $\theta(P_{-1})$ and $\theta(P_1)$. Thus, via Le Cam's method (13.3.3), we have that
>
> $$\mathfrak{M}_n(\mathsf{Bernoulli}([-1, 1]), (\cdot)^2) \geq \frac{1}{2}\delta^2\left(1 - \left\|P_{-1}^n - P_1^n\right\|_{\rm TV}\right).$$
>
> We would thus like to upper bound $\|P_{-1}^n - P_1^n\|_{\rm TV}$ as a function of the separation δ and sample size n; here we use Pinsker's inequality (Proposition 2.10(b)) and the tensorization identity (13.2.3) that makes KL-divergence so useful. Indeed, we have
>
> $$\left\|P_{-1}^n - P_1^n\right\|_{\rm TV}^2 \leq \frac{1}{2}D_{\rm kl}\left(P_{-1}^n \| P_1^n\right) = \frac{n}{2}D_{\rm kl}\left(P_{-1} \| P_1\right) = \frac{n}{2}\delta \log\frac{1 + \delta}{1 - \delta}.$$
>
> Noting that $\delta \log \frac{1+\delta}{1-\delta} \leq 3\delta^2$ for $\delta \in [0, 1/2]$, we obtain that $\|P_{-1}^n - P_1^n\|_{\rm TV} \leq \delta\sqrt{3n/2}$ for $\delta \leq 1/2$. In particular, we can guarantee a high probability of error in the associated hypothesis testing problem (recall inequality (13.3.2)) by taking $\delta = 1/\sqrt{6n}$; this guarantees $\|P_{-1}^n - P_1^n\|_{\rm TV} \leq \frac{1}{2}$. We thus have the minimax lower bound
>
> $$\mathfrak{M}_n(\mathsf{Bernoulli}([-1, 1]), (\cdot)^2) \geq \frac{1}{2}\delta^2\left(1 - \frac{1}{2}\right) = \frac{1}{24n}.$$

While the factor $1/24$ is smaller than necessary, this bound is optimal to within constant factors; the sample mean $(1/n)\sum_{i=1}^n X_i$ achieves mean-squared error $(1-\theta^2)/n$.
As an alternative proof, we may use the Hellinger distance and its associated decoupling identity (13.2.4). We sketch the idea, ignoring lower order terms when convenient. In this case, Proposition 2.10(a) implies

$$\|P_1^n - P_2^n\|_{\rm TV} \le d_{\rm hel}(P_1^n, P_2^n) = \sqrt{2 - 2(1 - d_{\rm hel}(P_1, P_2)^2)^n}.$$

Noting that

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

and noting that $(1-\delta^2) \approx e^{-\delta^2}$, we have (up to lower order terms in δ) that $\|P_1^n - P_2^n\|_{\rm TV} \le \sqrt{2 - 2\exp(-\delta^2 n/2)}$. Choosing $\delta^2 = 1/(4n)$, we have $\sqrt{2 - 2\exp(-\delta^2 n/2)} \le 1/2$, thus giving the lower bound

$$\mathfrak{M}_n(\mathsf{Bernoulli}([-1,1]), (\cdot)^2) \text{ “}\ge\text{” } \frac{1}{2}\delta^2\left(1 - \frac{1}{2}\right) = \frac{1}{16n},$$

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

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

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

13.4 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 $\mathcal{V}$ has cardinality larger than two. The method has played a central role in minimax theory, beginning with the pioneering work of Has’minskii and Ibragimov [10, 11]. More recent work following this initial push continues to the present day (e.g. [1, 21, 20, 2, 16, 9, 4]).

13.4.1 The classical (local) Fano method

We begin by stating Fano’s inequality, which provides a lower bound on the error in a multi-way hypothesis testing problem. Let V be a random variable taking values in a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| \ge 2$. If we let the function $h_2(p) = -p\log p - (1-p)\log(1-p)$ denote the entropy of the Bernoulli random variable with parameter p, Fano’s inequality (Proposition 2.13 from Chapter 2) takes the following form [e.g. 6, Chapter 2]:

Proposition 13.8 (Fano inequality)**.** *For any Markov chain $V \to X \to \widehat{V}$, we have*

$$h_2(\mathbb{P}(\widehat{V} \neq V)) + \mathbb{P}(\widehat{V} \neq V) \log(|\mathcal{V}| - 1) \geq H(V \mid \widehat{V}). \tag{13.4.1}$$

Restating the results in Chapter 2, we also have the following convenient rewriting of Fano's inequality when V is uniform in $\mathcal{V}$ (recall Corollary 2.14).

Corollary 13.9. *Assume that V is uniform on $\mathcal{V}$. For any Markov chain $V \to X \to \widehat{V}$,*

$$\mathbb{P}(\widehat{V} \neq V) \geq 1 - \frac{I(V;X) + \log 2}{\log(|\mathcal{V}|)}. \tag{13.4.2}$$

In particular, Corollary 13.9 shows that we have

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

where the infimum is taken over all testing procedures Ψ. By combining Corollary 13.9 with the reduction from estimation to testing in Proposition 13.3, we obtain the following result.

Proposition 13.10. *Let $\{\theta(P_v)\}_{v \in \mathcal{V}}$ be a 2δ-packing in the ρ-semimetric. Assume that V is uniform on the set $\mathcal{V}$, and conditional on $V = v$, we draw a sample $X \sim P_v$. Then the minimax risk has lower bound*

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

To gain some intuition for Proposition 13.10, we think of the lower bound as a function of the separation $\delta > 0$. Roughly, as $\delta \downarrow 0$, the separation condition between the distributions P_v is relaxed and we expect the distributions P_v to be closer to one another. In this case—as will be made more explicity presently—the hypothesis testing problem of distinguishing the P_v becomes more challenging, and the information $I(V;X)$ shrinks. Thus, what we roughly attempt to do is to choose our packing $\theta(P_v)$ as a function of δ, and find the largest $\delta > 0$ making the mutual information small enough that

$$\frac{I(V;X) + \log 2}{\log |\mathcal{V}|} \leq \frac{1}{2}. \tag{13.4.3}$$

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

$$I(V;X) = D_{\rm kl}\left(\mathbb{P}_{X,V} \| \mathbb{P}_X \mathbb{P}_V\right),$$

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 π (this may be a discrete or arbitrary distribution, though for simplicity we focus on the case when

π is discrete). Let P_v denote the distribution of X conditional on $V = v$, as in Proposition 13.10. Then marginally, we know that X is drawn from the mixture distribution

$$\overline{P} := \sum_v \pi(v) P_v.$$

With this definition of the mixture distribution, via algebraic manipulations, we have

$$I(V; X) = \sum_v \pi(v) D_{\rm kl}\left(P_v \| \overline{P}\right), \tag{13.4.4}$$

a representation that plays an important role in our subsequent derivations. To see equality (13.4.4), let μ be a base measure over $\mathcal{X}$ (assume w.l.o.g. that X has density $p(\cdot \mid v) = p_v(\cdot)$ conditional on $V = v$), and note that

$$I(V; X) = \sum_v \int_{\mathcal{X}} p(x \mid v)\pi(v) \log \frac{p(x \mid v)}{\sum_{v'} p(x \mid v')\pi(v')} d\mu(x) = \sum_v \pi(v) \int_{\mathcal{X}} p(x \mid v) \log \frac{p(x \mid v)}{\overline{p}(x)} d\mu(x).$$

Representation (13.4.4) 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 13.10, we have in this uniform setting that

$$\overline{P} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \quad \text{and} \quad I(V; X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\rm kl}\left(P_v \| \overline{P}\right).$$

The mutual information is small if the typical conditional distribution P_v is difficult to distinguish—has small KL-divergence—from $\overline{P}$.

The local Fano method

The local Fano method is based on a weakening of the mixture representation of mutual information (13.4.4), then giving a uniform upper bound on divergences between all pairs of the conditional distributions P_v and $P_{v'}$. (This method is known in the statistics literature as the "generalied Fano method," a poor name, as it is based on a weak upper bound on mutual information.) In particular (focusing on the case when V is uniform), the convexity of $-\log$ implies that

$$I(V; X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\rm kl}\left(P_v \| \overline{P}\right) \le \frac{1}{|\mathcal{V}|^2} \sum_{v, v'} D_{\rm kl}\left(P_v \| P_{v'}\right). \tag{13.4.5}$$

(In fact, the KL-divergence is *jointly* convex in its arguments; see Appendix 13.7 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 2δ-packing (recall Section 13.2.1), meaning that $\rho(\theta(P_v), \theta(P_{v'})) \ge 2\delta$ for all $v \neq v'$, but which additionally satisfy the uniform upper bound

$$D_{\rm kl}\left(P_v \| P_{v'}\right) \le \kappa^2 \delta^2 \quad \text{for all } v, v' \in \mathcal{V}, \tag{13.4.6}$$

where $\kappa > 0$ is a fixed problem-dependent constant. If we have the inequality (13.4.6), then so long as we can find a *local* packing $\mathcal{V}$ such that

$$\log |\mathcal{V}| \ge 2(\kappa^2 \delta^2 + \log 2),$$

we are guaranteed the testing error condition (13.4.3), and hence the minimax lower bound

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta).$$

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

Constructing local packings As mentioned above, the main difficulty in using Fano's method is in the construction of so-called "local" packings. In these problems, the idea is to construct a packing $\mathcal{V}$ of a fixed set (in a vector space, say $\mathbb{R}^d$) with constant radius and constant distance. Then we scale elements of the packing by $\delta > 0$, which leaves the cardinality $|\mathcal{V}|$ identical, but allows us to scale δ in the separation in the packing and the uniform divergence bound (13.4.6). In particular, Lemmas 13.5 and 13.6 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 13.11 (Normal mean estimation)**:** Consider the d-dimensional normal location family $\mathcal{N}_d = \{\mathsf{N}(\theta, \sigma^2 I_{d\times d}) \mid \theta \in \mathbb{R}^d\}$; we wish to estimate the mean $\theta = \theta(P)$ of a given distribution $P \in \mathcal{N}_d$ in mean-squared error, that is, with loss $\|\widehat{\theta} - \theta\|_2^2$. Let $\mathcal{V}$ be a 1/2-packing of the unit ℓ_2-ball with cardinality at least 2^d, as guaranteed by Lemma 13.6. (We assume for simplicity that $d \geq 2$.)

Now we construct our local packing. Fix $\delta > 0$, and for each $v \in \mathcal{V}$, set $\theta_v = \delta v \in \mathbb{R}^d$. Then we have

$$\|\theta_v - \theta_{v'}\|_2 = \delta \left\|v - v'\right\|_2 \geq \frac{\delta}{2}$$

for each distinct pair $v, v' \in \mathcal{V}$, and moreover, we note that $\|\theta_v - \theta_{v'}\|_2 \leq \delta$ for such pairs as well. By applying the Fano minimax bound of Proposition 13.10, we see that (given n normal observations $X_i \stackrel{\text{i.i.d.}}{\sim} P$)

$$\mathfrak{M}_n(\theta(\mathcal{N}_d), \|\cdot\|_2^2) \geq \left(\frac{1}{2} \cdot \frac{\delta}{2}\right)^2 \left(1 - \frac{I(V; X_1^n) + \log 2}{\log|\mathcal{V}|}\right) = \frac{\delta^2}{16}\left(1 - \frac{I(V; X_1^n) + \log 2}{d \log 2}\right).$$

Now note that for any pair v, v', if P_v is the normal distribution $\mathsf{N}(\theta_v, \sigma^2 I_{d\times d})$ we have

$$D_{\rm kl}\left(P_v^n \| P_{v'}^n\right) = n \cdot D_{\rm kl}\left(\mathsf{N}(\delta v, \sigma^2 I_{d\times d}) \| \mathsf{N}(\delta v', \sigma^2 I_{d\times d})\right) = n \cdot \frac{\delta^2}{2\sigma^2}\left\|v - v'\right\|_2^2,$$

as the KL-divergence between two normal distributions with identical covariance is

$$D_{\rm kl}\left(\mathsf{N}(\theta_1, \Sigma) \| \mathsf{N}(\theta_2, \Sigma)\right) = \frac{1}{2}(\theta_1 - \theta_2)^\top \Sigma^{-1}(\theta_1 - \theta_2)$$

as in Example 2.7. As $\|v - v'\|_2 \leq 1$, we have the KL-divergence bound (13.4.6) with $\kappa^2 = n/2\sigma^2$.

Combining our derivations, we have the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}_d), \|\cdot\|_2^2) \geq \frac{\delta^2}{16}\left(1 - \frac{n\delta^2/2\sigma^2 + \log 2}{d \log 2}\right). \tag{13.4.7}$$

Then by taking $\delta^2 = d\sigma^2 \log 2/(2n)$, we see that

$$1 - \frac{n\delta^2/2\sigma^2 + \log 2}{d \log 2} = 1 - \frac{1}{d} - \frac{1}{4} \ge \frac{1}{4}$$

by assumption that $d \ge 2$, and inequality (13.4.7) implies the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}_d), \|\cdot\|_2^2) \ge \frac{d\sigma^2 \log 2}{32n} \cdot \frac{1}{4} \ge \frac{1}{185} \cdot \frac{d\sigma^2}{n}.$$

While the constant 1/185 is not sharp, we do obtain the right scaling in d, n, and the variance σ^2; the sample mean attains the same risk. ♣

Example 13.12 (Linear regression)**:** In this example, we show how local packings can give (up to some constant factors) sharp minimax rates for standard linear regression problems. In particular, for fixed matrix $X \in \mathbb{R}^{n \times d}$, we observe

$$Y = X\theta + \varepsilon,$$

where $\varepsilon \in \mathbb{R}^n$ consists of independent random variables ε_i with variance bounded by $\mathrm{Var}(\varepsilon_i) \le \sigma^2$, and $\theta \in \mathbb{R}^d$ is allowed to vary over $\mathbb{R}^d$. For the purposes of our lower bound, we may assume that $\varepsilon \sim \mathsf{N}(0, \sigma^2 I_{n\times n})$. Let $\mathcal{P}$ denote the family of such normally distributed linear regression problems, and assume for simplicity that $d \ge 32$.

In this case, we use the Gilbert-Varshamov bound (Lemma 13.5) to construct a local packing and attain minimax rates. Indeed, let $\mathcal{V}$ be a packing of $\{-1, 1\}^d$ such that $\|v - v'\|_1 \ge d/2$ for distinct elements of $\mathcal{V}$, and let $|\mathcal{V}| \ge \exp(d/8)$ as guaranteed by the Gilbert-Varshamov bound. For fixed $\delta > 0$, if we set $\theta_v = \delta v$, then we have the packing guarantee for distinct elements v, v' that

$$\|\theta_v - \theta_{v'}\|_2^2 = \delta^2 \sum_{j=1}^d (v_j - {v'}_j)^2 = 4\delta^2 \left\|v - v'\right\|_1 \ge 2d\delta^2.$$

Moreover, we have the upper bound

$$\begin{aligned} D_{\rm kl}\left(\mathsf{N}(X\theta_v, \sigma^2 I_{n\times n}) \| \mathsf{N}(X\theta_{v'}, \sigma^2 I_{n\times n})\right) &= \frac{1}{2\sigma^2} \|X(\theta_v - \theta_{v'})\|_2^2 \\ &\le \frac{\delta^2}{2\sigma^2} \gamma_{\max}^2(X) \left\|v - v'\right\|_2^2 \le \frac{2d}{\sigma^2} \gamma_{\max}^2(X) \delta^2, \end{aligned}$$

where $\gamma_{\max}(X)$ denotes the maximum singular value of X. Consequently, the bound (13.4.6) holds with $\kappa^2 \le 2d\gamma_{\max}^2(X)/\sigma^2$, and we have the minimax lower bound

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|_2^2) \ge \frac{d\delta^2}{2}\left(1 - \frac{I(V;Y) + \log 2}{\log |\mathcal{V}|}\right) \ge \frac{d\delta^2}{2}\left(1 - \frac{\frac{2d\gamma_{\max}^2(X)}{\sigma^2}\delta^2 + \log 2}{d/8}\right).$$

Now, if we choose

$$\delta^2 = \frac{\sigma^2}{64\gamma_{\max}^2(X)}, \quad \text{then} \quad 1 - \frac{8\log 2}{d} - \frac{16d\gamma_{\max}^2(X)\delta^2}{d} \ge 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2},$$

by assumption that $d \ge 32$. In particular, we obtain the lower bound

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|_2^2) \ge \frac{1}{256} \frac{\sigma^2 d}{\gamma_{\max}^2(X)} = \frac{1}{256} \frac{\sigma^2 d}{n} \frac{1}{\gamma_{\max}^2(\frac{1}{\sqrt{n}}X)},$$

for a convergence rate (roughly) of $\sigma^2 d/n$ after rescaling the singular values of X by $1/\sqrt{n}$. This bound is sharp in terms of the dimension, dependence on n, and the variance σ^2, but it does not fully capture the dependence on X, as it depends only on the maximum singular value. Indeed, in this case, an exact calculation (cf. [14]) shows that the minimax value of the problem is exactly $\sigma^2 \operatorname{tr}((X^\top X)^{-1})$. Letting $\lambda_j(A)$ be the jth eigenvalue of a matrix A, we have

$$\begin{aligned}\sigma^2 \operatorname{tr}((X^\top X)^{-1}) = \frac{\sigma^2}{n} \operatorname{tr}((n^{-1} X^\top X)^{-1}) &= \frac{\sigma^2}{n} \sum_{j=1}^d \frac{1}{\lambda_j(\frac{1}{n} X^\top X)} \\ &\ge \frac{\sigma^2 d}{n} \min_j \frac{1}{\lambda_j(\frac{1}{n} X^\top X)} = \frac{\sigma^2 d}{n} \frac{1}{\gamma_{\max}^2(\frac{1}{\sqrt{n}} X)}.\end{aligned}$$

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

13.4.2 A distance-based Fano method

While the testing lower bound (13.4.2) 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 (13.4.1) and (13.4.2) 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 $\rho_{\mathcal{V}}$ is defined on $\mathcal{V}$, and we are interested in bounding the estimation error $\rho_{\mathcal{V}}(\widehat{V}, V)$. We begin by providing analogues of the lower bounds (13.4.1) and (13.4.2) that replace the testing error with the tail probability $\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t)$. By Markov's inequality, such control directly yields bounds on the expectation $\mathbb{E}[\rho_{\mathcal{V}}(\widehat{V}, V)]$. As we show in the sequel and in chapters to come, these distance-based Fano inequalities allow more direct proofs of a variety of minimax bounds without the need for careful construction of packing sets or metric entropy calculations as in other arguments.

We begin with the distance-based analogue of the usual discrete Fano inequality in Proposition 13.8. Let V be a random variable supported on a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| \ge 2$, and let $\rho : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a function defined on $\mathcal{V} \times \mathcal{V}$. In the usual setting, the function ρ is a metric on the space $\mathcal{V}$, but our theory applies to general functions. For a given scalar $t \ge 0$, the maximum and minimum *neighborhood sizes at radius* t are given by

$$N_t^{\max} := \max_{v \in \mathcal{V}} \left\{ \operatorname{card}\{v' \in \mathcal{V} \mid \rho(v, v') \le t\} \right\} \quad \text{and} \quad N_t^{\min} := \min_{v \in \mathcal{V}} \left\{ \operatorname{card}\{v' \in \mathcal{V} \mid \rho(v, v') \le t\} \right\}. \tag{13.4.8}$$

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

Proposition 13.13. *For any Markov chain $V \to X \to \widehat{V}$, we have*

$$h_2(P_t) + P_t \log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}} + \log N_t^{\max} \ge H(V \mid \widehat{V}). \tag{13.4.9}$$

Before proving the proposition, which we do in Section 13.5.1, it is informative to note that it reduces to the standard form of Fano's inequality (13.4.1) in a special case. Suppose that we take $\rho_{\mathcal{V}}$ to be the 0-1 metric, meaning that $\rho_{\mathcal{V}}(v, v') = 0$ if $v = v'$ and 1 otherwise. Setting $t = 0$ in Proposition 13.13, we have $P_0 = \mathbb{P}[\widehat{V} \ne V]$ and $N_0^{\min} = N_0^{\max} = 1$, whence inequality (13.4.9) reduces to inequality (13.4.1). Other weakenings allow somewhat clearer statements (see Section 13.5.2 for a proof):

Corollary 13.14. *If V is uniform on $\mathcal{V}$ and $(|\mathcal{V}| - N_t^{\min}) > N_t^{\max}$, then*

$$\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t) \ge 1 - \frac{I(V; X) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}. \tag{13.4.10}$$

Inequality (13.4.10) is the natural analogue of the classical mutual-information based form of Fano's inequality (13.4.2), and it provides a qualitatively similar bound. The main difference is that the usual cardinality $|\mathcal{V}|$ is replaced by the ratio $|\mathcal{V}|/N_t^{\max}$. This quantity serves as a rough measure of the number of possible "regions" in the space $\mathcal{V}$ that are distinguishable—that is, the number of subsets of $\mathcal{V}$ for which $\rho_{\mathcal{V}}(v, v') > t$ when v and v' belong to different regions. While this construction is similar in spirit to the usual construction of packing sets in the standard reduction from testing to estimation (cf. Section 13.2.1), our bound allows us to skip the packing set construction. We can directly compute $I(V; X)$ where V takes values over the full space, as opposed to computing the mutual information $I(V'; X)$ for a random variable V' uniformly distributed over a packing set contained within $\mathcal{V}$. In some cases, the former calculation can be much simpler, as illustrated in examples and chapters to follow.

We now turn to providing a few consequences of Proposition 13.13 and Corollary 13.14, showing how they can be used to derive lower bounds on the minimax risk. Proposition 13.13 is a generalization of the classical Fano inequality (13.4.1), so it leads naturally to a generalization of the classical Fano lower bound on minimax risk, which we describe here. This reduction from estimation to testing is somewhat more general than the classical reductions, since we do not map the original estimation problem to a strict test, but rather a test that allows errors. Consider as in the standard reduction of estimation to testing in Section 13.2.1 a family of distributions $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ indexed by a finite set $\mathcal{V}$. This family induces an associated collection of parameters $\{\theta_v := \theta(P_v)\}_{v \in \mathcal{V}}$. Given a function $\rho_{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and a scalar t, we define the separation $\delta(t)$ of this set relative to the metric ρ on Θ via

$$\delta(t) := \sup \left\{ \delta \mid \rho(\theta_v, \theta_{v'}) \ge \delta \text{ for all } v, v' \in \mathcal{V} \text{ such that } \rho_{\mathcal{V}}(v, v') > t \right\}. \tag{13.4.11}$$

As a special case, when $t = 0$ and $\rho_{\mathcal{V}}$ is the discrete metric, this definition reduces to that of a packing set: we are guaranteed that $\rho(\theta_v, \theta_{v'}) \ge \delta(0)$ for all distinct pairs $v \neq v'$, as in the classical approach to minimax lower bounds. On the other hand, allowing for $t > 0$ lends greater flexibility to the construction, since only certain pairs θ_v and $\theta_{v'}$ are required to be well-separated.

Given a set $\mathcal{V}$ and associated separation function (13.4.11), we assume the canonical estimation setting: nature chooses $V \in \mathcal{V}$ uniformly at random, and conditioned on this choice $V = v$, a sample X is drawn from the distribution P_v. We then have the following corollary of Proposition 13.13, whose argument is completely identical to that for inequality (13.2.1):

Corollary 13.15. *Given V uniformly distributed over $\mathcal{V}$ with separation function $\delta(t)$, we have*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \ge \Phi\Big(\frac{\delta(t)}{2}\Big) \left[1 - \frac{I(X; V) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}\right] \qquad \textit{for all } t. \tag{13.4.12}$$

Notably, using the discrete metric $\rho_{\mathcal{V}}(v, v') = \mathbf{1}\{v \neq v'\}$ and taking $t = 0$ in the lower bound (13.4.12) 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 13.15 in providing a minimax lower bound on the performance of regression estimators.

Example: Normal regression model Consider the d-dimensional linear regression model $Y = X\theta + \varepsilon$, where $\varepsilon \in \mathbb{R}^n$ is i.i.d. $\mathsf{N}(0, \sigma^2)$ and $X \in \mathbb{R}^{n \times d}$ is known, but θ is not. In this case, our family of distributions is

$$\mathcal{P}_X := \left\{ Y \sim \mathsf{N}(X\theta, \sigma^2 I_{n\times n}) \mid \theta \in \mathbb{R}^d \right\} = \left\{ Y = X\theta + \varepsilon \mid \varepsilon \sim \mathsf{N}(0, \sigma^2 I_{n \times n}), \theta \in \mathbb{R}^d \right\}.$$

We then obtain the following minimax lower bound on the minimax error in squared ℓ_2-norm: there is a universal (numerical) constant $c > 0$ such that

$$\mathfrak{M}_n(\theta(\mathcal{P}_X, \|\cdot\|_2^2) \ge c \frac{\sigma^2 d^2}{\|X\|_{\mathrm{Fr}}^2} \ge \frac{c}{\gamma_{\max}(X/\sqrt{n})^2} \cdot \frac{\sigma^2 d}{n}, \tag{13.4.13}$$

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

To see inequality (13.4.13), let the set $\mathcal{V} = \{-1, 1\}^d$ be the d-dimensional hypercube, and define $\theta_v = \delta v$ for some fixed $\delta > 0$. Then letting $\rho_{\mathcal{V}}$ be the Hamming metric on $\mathcal{V}$ and ρ be the usual ℓ_2-norm, the associated separation function (13.4.11) satisfies $\delta(t) > \max\{\sqrt{t}, 1\}\delta$. Now, for any $t \le \lceil d/3 \rceil$, the neighborhood size satisfies

$$N_t^{\max} = \sum_{\tau=0}^{t} \binom{d}{\tau} \le 2 \binom{d}{t} \le 2 \left(\frac{de}{t} \right)^t.$$

Consequently, for $t \le d/6$, the ratio $|\mathcal{V}| / N_t^{\max}$ satisfies

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

for $d \ge 12$. (The case $2 \le d < 12$ can be checked directly). In particular, by taking $t = \lfloor d/6 \rfloor$ we obtain via Corollary 13.15 that

$$\mathfrak{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \ge \frac{\max\{\lfloor d/6 \rfloor, 2\}\delta^2}{4} \left(1 - \frac{I(Y; V) + \log 2}{\max\{d/6, 2\log 2\}} \right).$$

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

$$\begin{aligned} I(Y; V) &\le n \frac{1}{|\mathcal{V}|^2} \sum_{v \in \mathcal{V}} \sum_{v' \in \mathcal{V}} D_{\mathrm{kl}}\left(\mathsf{N}(X\theta_v, \sigma^2 I_{n\times n}) \| \mathsf{N}(X\theta_{v'}, \sigma^2 I_{n \times n}) \right) \\ &= \frac{\delta^2}{2\sigma^2} \mathbb{E}\left[\left\| XV - XV' \right\|_2^2 \right] = \frac{\delta^2}{\sigma^2} \|X\|_{\mathrm{Fr}}^2. \end{aligned}$$

Substituting this into the preceding minimax bound, we obtain

$$\mathfrak{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \ge \frac{\max\{\lfloor d/6 \rfloor, 2\}\delta^2}{4} \left(1 - \frac{\delta^2 \|X\|_{\mathrm{Fr}}^2 / \sigma^2 + \log 2}{\max\{d/6, 2\log 2\}} \right).$$

Choosing $\delta^2 \asymp d\sigma^2 / \|X\|_{\mathrm{Fr}}^2$ gives the result (13.4.13).

13.5 Proofs of results

13.5.1 Proof of Proposition 13.13

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 $\rho(\widehat{V}, V) \le t$, we compute the entropy $H(E, V \mid \widehat{V})$ in two different ways. On one hand, by the chain rule for entropy, we have

$$H(E, V \mid \widehat{V}) = H(V \mid \widehat{V}) + \underbrace{H(E \mid V, \widehat{V})}_{=0}, \tag{13.5.1}$$

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

$$H(E, V \mid \widehat{V}) = H(E \mid \widehat{V}) + H(V \mid E, \widehat{V}) \le H(E) + H(V \mid E, \widehat{V}),$$

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

$$H(V \mid E, \widehat{V}) = \mathbb{P}(E = 0) H(V \mid E = 0, \widehat{V}) + \mathbb{P}(E = 1) H(V \mid E = 1, \widehat{V}),$$

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

$$H(V \mid E = 0, \widehat{V}) \le \log(|\mathcal{V}| - N_t^{\min}),$$

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

$$H(V \mid E = 1, \widehat{V}) \le \log N_t^{\max},$$

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

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

$$H(E, V \mid \widehat{V}) \le H(E) + P_t \log\left(|\mathcal{V}| - N^{\min}\right) + (1 - P_t) \log N_t^{\max}.$$

Combining this inequality with our earlier equality (13.5.1), we see that

$$H(V \mid \widehat{V}) \le H(E) + P_t \log(|\mathcal{V}| - N_t^{\min}) + (1 - P_t) \log N_t^{\max}.$$

Since $H(E) = h_2(P_t)$, the claim (13.4.9) follows.

13.5.2 Proof of Corollary 13.14

First, by the information-processing inequality [e.g. 6, Chapter 2], we have $I(V; \widehat{V}) \le I(V; X)$, and hence $H(V \mid X) \le H(V \mid \widehat{V})$. Since $h_2(P_t) \le \log 2$, inequality (13.4.9) implies that

$$H(V \mid X) - \log N_t^{\max} \le H(V \mid \widehat{V}) - \log N_t^{\max} \le \mathbb{P}(\rho(\widehat{V}, V) > t) \log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}} + \log 2.$$

Rearranging the preceding equations yields

$$\mathbb{P}(\rho(\widehat{V}, V) > t) \ge \frac{H(V \mid X) - \log N_t^{\max} - \log 2}{\log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}}}. \tag{13.5.2}$$

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 (13.5.2) yields the inequality

$$\mathbb{P}(\rho(\widehat{V}, V) > t) \ge \frac{\log \frac{|\mathcal{V}|}{N_t^{\max}}}{\log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}}} - \frac{I(V;X) + \log 2}{\log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}}} \ge 1 - \frac{I(V;X) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}.$$

13.6 Deferred proofs

13.7 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 $x/t \in \operatorname{dom} f$, is jointly convex in the arguments x, t (see Chapter 3.2.6 of Boyd and Vandenberghe [3]). Then we have

Proposition 13.16. *Let P_1, P_2, Q_1, Q_2 be distributions on a set $\mathcal{X}$ and $f : \mathbb{R}_+ \to \mathbb{R}$ be convex. Then for any $\lambda \in [0,1]$,*

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

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 μ. Define the perspective $g(x,t) = tf(x/t)$. Then

$$\begin{aligned}
D_f\left(\lambda P_1 + (1-\lambda)P_2 \| \lambda Q_1 + (1-\lambda)Q_2\right) &= \int (\lambda q_1 + (1-\lambda)q_2) f\left(\frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2}\right) d\mu \\
&= \int g(\lambda p_1 + (1-\lambda)p_2, \lambda q_1 + (1-\lambda)q_2) d\mu \\
&\le \lambda \int g(p_1, q_1) d\mu + (1-\lambda) \int g(p_2, q_2) d\mu \\
&= \lambda D_f\left(P_1 \| Q_1\right) + (1-\lambda) D_f\left(P_2 \| Q_2\right),
\end{aligned}$$

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

Bibliography

[1] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwebte Gebiet*, 65:181–238, 1983.

[2] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Transactions on Information Theory*, 51(4):1611–1614, 2005.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[4] E. J. Candès and M. A. Davenport. How well can we estimate a sparse vector. *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition.* Wiley, 2006.

[7] J. C. Duchi and M. J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. *arXiv:1311.2669 [cs.IT]*, 2013.

[8] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual Symposium on Foundations of Computer Science*, 2013.

[9] A. Guntuboyina. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

[10] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory of Probability and Applications*, 23:794–798, 1978.

[11] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory.* Springer-Verlag, 1981.

[12] A. Kolmogorov and V. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *Uspekhi Matematischeskikh Nauk*, 14(2):3–86, 1959.

[13] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, 1986.

[14] E. L. Lehmann and G. Casella. *Theory of Point Estimation, Second Edition.* Springer, 1998.

[15] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[16] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976—6994, 2011.

[17] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Programming Society, 2009.

[18] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[19] A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10(4):299–326, 1939.

[20] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[21] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

Chapter 14

Assouad's method

Assouad's method provides a somewhat different technique for proving lower bounds. Instead of reducing the estimation problem to a multiple hypothesis test or simpler estimation problem, as with Le Cam's method and Fano's method from the preceding lectures, here we transform the original estimation problem into multiple binary hypothesis testing problems, using the structure of the problem in an essential way. Assouad's method applies only problems where the loss we care about is naturally related to identification of individual points on a hypercube.

14.1 The method

14.1.1 Well-separated problems

To describe the method, we begin by encoding a notion of separation and loss, similar to what we did in the classical reduction of estimation to testing. For some $d \in \mathbb{N}$, let $\mathcal{V} = \{-1, 1\}^d$, and let us consider a family $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ indexed by the hypercube. We say that the the family P_v induces a 2δ-*Hamming separation* for the loss $\Phi \circ \rho$ if there exists a function $\widehat{\mathsf{v}} : \theta(\mathcal{P}) \to \{-1, 1\}^d$ satisfying

$$\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^{d} \mathbf{1}\left\{[\widehat{\mathsf{v}}(\theta)]_j \neq v_j\right\}. \tag{14.1.1}$$

That is, we can take the parameter θ and test the individual indices via $\widehat{\mathsf{v}}$.

Example 14.1 (Estimation in ℓ_1-error)**:** Suppose we have a family of multivariate Laplace distributions on $\mathbb{R}^d$—distributions with density proportional to $p(x) \propto \exp(-\|x - \mu\|_1)$—and we wish to estimate the mean in ℓ_1-distance. For $v \in \{-1, 1\}^d$ and some fixed $\delta > 0$ let p_v be the density

$$p_v(x) = \frac{1}{2} \exp\left(-\|x - \delta v\|_1\right),$$

which has mean $\theta(P_v) = \delta v$. Under the ℓ_1-loss, we have for any $\theta \in \mathbb{R}^d$ that

$$\|\theta - \theta(P_v)\|_1 = \sum_{j=1}^{d} |\theta_j - \delta v_j| \geq \delta \sum_{j=1}^{d} \mathbf{1}\left\{\operatorname{sign}(\theta_j) \neq v_j\right\},$$

so that this family induces a δ-Hamming separation for the ℓ_1-loss. ♣

14.1.2 From estimation to multiple binary tests

As in the standard reduction from estimation to testing, we consider the following random process: nature chooses a vector $V \in \{-1, 1\}^d$ uniformly at random, after which the sample X is drawn from the distribution P_v conditional on $V = v$. Then, if we let $\mathbb{P}_{\pm j}$ denote the joint distribution over the random index V and X conditional on the jth coordinate $V_j = \pm 1$, we obtain the following sharper version of Assouad's lemma [2] (see also the paper [1]); we provide a proof in Section 14.1.3 to follow.

Lemma 14.2. *Under the conditions of the previous paragraph, we have*

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \ge \delta \sum_{j=1}^d \inf_\Psi \left[\mathbb{P}_{+j}(\Psi(X) \neq +1) + \mathbb{P}_{-j}(\Psi(X) \neq -1)\right].$$

While Lemma 14.2 requires conditions on the loss Φ and metric ρ for the separation condition (14.1.1) to hold, it is sometimes easier to apply than Fano's method. Moreover, while we will not address this in class, several researchers [1, 3] have noted that it appears to allow easier application in so-called "interactive" settings—those for which the sampling of the X_i may not be precisely i.i.d. It is closely related to Le Cam's method, discussed previously, as we see that if we define $P_{+j} = 2^{1-d} \sum_{v:v_j=1} P_v$ (and similarly for $-j$), Lemma 14.2 is equivalent to

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \ge \delta \sum_{j=1}^d \left[1 - \|P_{+j} - P_{-j}\|_{\rm TV}\right]. \tag{14.1.2}$$

There are standard weakenings of the lower bound (14.1.2) (and Lemma 14.2). We give one such weakening. First, we note that the total variation is convex, so that if we define $P_{v,+j}$ to be the distribution P_v where coordinate j takes the value $v_j = 1$ (and similarly for $P - v, -j$), we have

$$P_{+j} = \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} P_{v,+j} \quad \text{and} \quad P_{-j} = \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} P_{v,+j}.$$

Thus, by the triangle inequality, we have

$$\begin{aligned}\|P_{+j} - P_{-j}\|_{\rm TV} &= \left\| \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} P_{v,+j} - P_{v,-j} \right\|_{\rm TV} \\ &\le \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} \|P_{v,+j} - P_{v,-j}\|_{\rm TV} \le \max_{v,j} \|P_{v,+j} - P_{v,-j}\|_{\rm TV}.\end{aligned}$$

Then as long as the loss satisfies the per-coordinate separation (14.1.1), we obtain the following:

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \ge d\delta \left(1 - \max_{v,j} \|P_{v,+j} - P_{v,-j}\|_{\rm TV}\right). \tag{14.1.3}$$

This is the version of Assouad's lemma most frequently presented.

We also note that by the Cauchy-Schwarz inequality and convexity of the variation-distance, we have

$$\sum_{j=1}^d \|P_{+j} - P_{-j}\|_{\rm TV} \le \sqrt{d} \bigg(\sum_{j=1}^d \|P_{+j} - P_{-j}\|_{\rm TV}^2 \bigg)^{1/2} \le \sqrt{d} \bigg(\sum_{j=1}^d \frac{1}{2^d} \sum_v \|P_{v,+j} - P_{v,-j}\|_{\rm TV}^2 \bigg)^{\frac{1}{2}},$$

and consequently we have a not quite so terribly weak version of inequality (14.1.2):

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \ge \delta d \left[1 - \left(\frac{1}{d} \sum_{j=1}^{d} \sum_{v \in \{-1,1\}^d} \|P_{v,+j} - P_{v,-j}\|_{\rm TV}^2\right)^{\frac{1}{2}}\right]. \tag{14.1.4}$$

Regardless of whether we use the sharper version (14.1.2) or weakened versions (14.1.3) or (14.1.4), the technique is essentially the same. We simply seek a setting of the distributions P_v so that the probability of making a mistake in the hypothesis test of Lemma 14.2 is high enough—say 1/2—or the variation distance is small enough—such as $\|P_{+j} - P_{-j}\|_{\rm TV} \le 1/2$ for all j. Once this is satisfied, we obtain a minimax lower bound of the form

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \ge \delta \sum_{j=1}^{d} \left[1 - \frac{1}{2}\right] = \frac{d\delta}{2}.$$

14.1.3 Proof of Lemma 14.2

Fix an (arbitrary) estimator $\widehat{\theta}$. By assumption (14.1.1), we have

$$\Phi(\rho(\theta, \theta(P_v))) \ge 2\delta \sum_{j=1}^{d} \mathbf{1}\left\{[\widehat{\mathsf{v}}(\theta)]_j \neq v_j\right\}.$$

Taking expectations, we see that

$$\begin{aligned}\sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\Phi(\rho(\widehat{\theta}(X), \theta(P)))\right] &\ge \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{P_v}\left[\Phi(\rho(\widehat{\theta}(X), \theta_v))\right] \\ &\ge \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} 2\delta \sum_{j=1}^{d} \mathbb{E}_{P_v}\left[\mathbf{1}\left\{[\psi(\widehat{\theta})]_j \neq v_j\right\}\right]\end{aligned}$$

as the average is smaller than the maximum of a set and using the separation assumption (14.1.1). Recalling the definition of the mixtures $\mathbb{P}_{\pm j}$ as the joint distribution of V and X conditional on $V_j = \pm 1$, we swap the summation orders to see that

$$\begin{aligned}\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v\left([\widehat{\mathsf{v}}(\widehat{\theta})]_j \neq v_j\right) &= \frac{1}{|\mathcal{V}|} \sum_{v : v_j = 1} P_v\left([\widehat{\mathsf{v}}(\widehat{\theta})]_j \neq v_j\right) + \frac{1}{|\mathcal{V}|} \sum_{v : v_j = -1} P_v\left([\widehat{\mathsf{v}}(\widehat{\theta})]_j \neq v_j\right) \\ &= \frac{1}{2}\mathbb{P}_{+j}\left([\widehat{\mathsf{v}}(\widehat{\theta})]_j \neq v_j\right) + \frac{1}{2}\mathbb{P}_{-j}\left([\widehat{\mathsf{v}}(\widehat{\theta})]_j \neq v_j\right).\end{aligned}$$

This gives the statement claimed in the lemma, while taking an infimum over all testing procedures $\Psi : \mathcal{X} \to \{-1, +1\}$ gives the claim (14.1.2).

14.2 Example applications of Assouad's method

We now provide two example applications of Assouad's method. The first is a standard finite-dimensional lower bound, where we provide a lower bound in a normal mean estimation problem. For the second, we consider estimation in a logistic regression problem, showing a similar lower bound. In Chapter 15 to follow, we show how to use Assouad's method to prove strong lower bounds in a standard nonparametric problem.

Example 14.3 (Normal mean estimation)**:** For some $\sigma^2 > 0$ and $d \in \mathbb{N}$, we consider estimation of mean parameter for the normal location family

$$\mathcal{N} := \left\{ \mathsf{N}(\theta, \sigma^2 I_{d\times d}) : \theta \in \mathbb{R}^d \right\}$$

in squared Euclidean distance. We now show how for this family, the sharp Assouad's method implies the lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \ge \frac{d\sigma^2}{8n}. \tag{14.2.1}$$

Up to constant factors, this bound is sharp; the sample mean has mean squared error $d\sigma^2/n$. We proceed in (essentially) the usual way we have set up. Fix some $\delta > 0$ and define $\theta_v = \delta v$, taking $P_v = \mathsf{N}(\theta_v, \sigma^2 I_{d\times d})$ to be the normal distribution with mean θ_v. In this case, we see that the hypercube structure is natural, as our loss function decomposes on coordinates: we have $\|\theta - \theta_v\|_2^2 \ge \delta^2 \sum_{j=1}^d \mathbf{1}\{\mathrm{sign}(\theta_j) \neq v_j\}$. The family P_v thus induces a δ^2-Hamming separation for the loss $\|\cdot\|_2^2$, and by Assouad's method (14.1.2), we have

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \ge \frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \left\| P_{+j}^n - P_{-j}^n \right\|_{\mathrm{TV}} \right],$$

where $P_{\pm j}^n = 2^{1-d} \sum_{v : v_j = \pm 1} P_v^n$. It remains to provide upper bounds on $\|P_{+j}^n - P_{-j}^n\|_{\mathrm{TV}}$. By the convexity of $\|\cdot\|_{\mathrm{TV}}^2$ and Pinsker's inequality, we have

$$\left\| P_{+j}^n - P_{-j}^n \right\|_{\mathrm{TV}}^2 \le \max_{d_{\rm ham}(v,v') \le 1} \|P_v^n - P_{v'}^n\|_{\mathrm{TV}}^2 \le \frac{1}{2} \max_{d_{\rm ham}(v,v') \le 1} D_{\rm kl}\left(P_v^n \| P_{v'}^n\right).$$

But of course, for any v and v' differing in only 1 coordinate,

$$D_{\rm kl}\left(P_v^n \| P_{v'}^n\right) = \frac{n}{2\sigma^2} \left\| \theta_v - \theta_{v'} \right\|_2^2 = \frac{2n}{\sigma^2} \delta^2,$$

giving the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \ge 2\delta^2 \sum_{j=1}^d \left[1 - \sqrt{2n\delta^2/\sigma^2} \right].$$

Choosing $\delta^2 = \sigma^2/8n$ gives the claimed lower bound (14.2.1). ♣

Example 14.4 (Logistic regression)**:** In this example, consider the logistic regression model, where we have known (fixed) regressors $X_i \in \mathbb{R}^d$ and an unknown parameter $\theta \in \mathbb{R}^d$; the goal is to infer θ after observing a sequence of $Y_i \in \{-1, 1\}$, where for $y \in \{-1, 1\}$ we have

$$P(Y_i = y \mid X_i, \theta) = \frac{1}{1 + \exp(-y X_i^\top \theta)}.$$

Denote this family by $\mathcal{P}_{\mathsf{log}}$, and for $P \in \mathcal{P}_{\mathsf{log}}$, let $\theta(P)$ be the predictor vector θ. We would like to estimate the vector θ in squared ℓ_2 error. As in Example 14.3, if we choose some $\delta > 0$ and for each $v \in \{-1, 1\}^d$, we set $\theta_v = \delta v$, then we have the δ^2-separation in Hamming metric $\|\theta - \theta_v\|_2^2 \ge \delta^2 \sum_{j=1}^d \mathbf{1}\{\mathrm{sign}(\theta_j) \neq v_j\}$. Let P_v^n denote the distribution of the n independent observations Y_i when $\theta = \theta_v$. Then we have by Assouad's lemma (and the weakening (14.1.4)) that

$$\begin{aligned}\mathfrak{M}_n(\theta(\mathcal{P}_{\mathsf{log}}), \|\cdot\|_2^2) &\ge \frac{\delta^2}{2}\sum_{j=1}^d \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\rm TV}\right] \\ &\ge \frac{d\delta^2}{2}\left[1 - \left(\frac{1}{d}\sum_{j=1}^d \frac{1}{2^d}\sum_{v\in\{-1,1\}^d}\left\|P_{v,+j}^n - P_{v,-j}^n\right\|_{\rm TV}^2\right)^{\frac{1}{2}}\right]. \qquad (14.2.2)\end{aligned}$$

It remains to bound $\|P_{v,+j}^n - P_{v,-j}^n\|_{\rm TV}^2$ to find our desired lower bound. To that end, use the shorthands $p_v(x) = 1/(1+\exp(\delta x^\top v))$ and let $D_{\rm kl}(p\|q)$ be the binary KL-divergence between $\mathsf{Bernoulli}(p)$ and $\mathsf{Bernoulli}(q)$ distributions. Then we have by Pinsker's inequality (recall Proposition 2.10) that for any v, v',

$$\|P_v^n - P_{v'}^n\|_{\rm TV} \le \frac{1}{4}[D_{\rm kl}(P_v^n\|P_{v'}^n) + D_{\rm kl}(P_{v'}^n\|P_v^n)] = \frac{1}{4}\sum_{i=1}^n\left[D_{\rm kl}(p_v(X_i)\|p_{v'}(X_i)) + D_{\rm kl}(p_{v'}(X_i)\|p_v(X_i))\right].$$

Let us upper bound the final KL-divergence. Let $p_a = 1/(1+e^a)$ and $p_b = 1/(1+e^b)$. We claim that

$$D_{\rm kl}(p_a\|p_b) + D_{\rm kl}(p_b\|p_a) \le (a-b)^2. \qquad (14.2.3)$$

Deferring the proof of claim (14.2.3), we immediately see that

$$\|P_v^n - P_{v'}^n\|_{\rm TV} \le \frac{\delta^2}{4}\sum_{i=1}^n \left(X_i^\top(v - v')\right)^2.$$

Now we recall inequality (14.2.2) for motivation, and we see that the preceding display implies

$$\frac{1}{2^d d}\sum_{j=1}^d\sum_{v\in\{-1,1\}^d}\left\|P_{v,+j}^n - P_{v,-j}^n\right\|_{\rm TV}^2 \le \frac{\delta^2}{4d}\frac{1}{2^d}\sum_{v\in\{-1,1\}^d}\sum_{j=1}^d\sum_{i=1}^n (2X_{ij})^2 = \frac{\delta^2}{d}\sum_{i=1}^n\sum_{j=1}^d X_{ij}^2.$$

Replacing the final double sum with $\|X\|_{\rm Fr}^2$, where X is the matrix of the X_i, we have

$$\mathfrak{M}_n(\theta(\mathcal{P}_{\mathsf{log}}), \|\cdot\|_2^2) \ge \frac{d\delta^2}{2}\left[1 - \left(\frac{\delta^2}{d}\|X\|_{\rm Fr}^2\right)^{\frac{1}{2}}\right].$$

Setting $\delta^2 = d/4\|X\|_{\rm Fr}^2$, we obtain

$$\mathfrak{M}_n(\theta(\mathcal{P}_{\mathsf{log}}), \|\cdot\|_2^2) \ge \frac{d\delta^2}{4} = \frac{d^2}{16\|X\|_{\rm Fr}^2} = \frac{d}{n}\cdot\frac{1}{16\frac{1}{dn}\sum_{i=1}^n\|X_i\|_2^2}.$$

That is, we have a minimax lower bound scaling roughly as d/n for logistic regression, where "large" X_i (in ℓ_2-norm) suggest that we may obtain better performance in estimation. This is intuitive, as a larger X_i gives a better signal to noise ratio.

We now return to prove the claim (14.2.3). Indeed, by a straightforward expansion, we have

$$\begin{aligned}D_{\rm kl}(p_a\|p_b) + D_{\rm kl}(p_b\|p_a) &= p_a\log\frac{p_a}{p_b} + (1-p_a)\log\frac{1-p_a}{1-p_b} + p_b\log\frac{p_b}{p_a} + (1-p_b)\log\frac{1-p_b}{1-p_a} \\ &= (p_a - p_b)\log\frac{p_a}{p_b} + (p_b - p_a)\log\frac{1-p_a}{1-p_b} = (p_a - p_b)\log\left(\frac{p_a}{1-p_a}\frac{1-p_b}{p_b}\right).\end{aligned}$$

Now note that $p_a/(1-p_a) = e^{-a}$ and $(1-p_b)/p_b = e^b$. Thus we obtain

$$D_{\rm kl}\left(p_a \| p_b\right) + D_{\rm kl}\left(p_b \| p_a\right) = \left(\frac{1}{1+e^a} - \frac{1}{1+e^b}\right) \log\left(e^{b-a}\right) = (b-a)\left(\frac{1}{1+e^a} - \frac{1}{1+e^b}\right)$$

Now assume without loss of generality that $b \ge a$. Noting that $e^x \ge 1 + x$ by convexity, we have

$$\frac{1}{1+e^a} - \frac{1}{1+e^b} = \frac{e^b - e^a}{(1+e^a)(1+e^b)} \le \frac{e^b - e^a}{e^b} = 1 - e^{a-b} \le 1 - (1 + (a-b)) = b - a,$$

yielding claim (14.2.3). ♣

Bibliography

[1] E. Arias-Castro, E. Candés, and M. Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2013.

[2] P. Assouad. Deux remarques sur l'estimation. *C. R. Academy Scientifique Paris Séries I Mathematics*, 296(23):1021–1024, 1983.

[3] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy, data processing inequalities, and minimax rates. *arXiv:1302.3203 [math.ST]*, 2013. URL `http://arxiv.org/abs/1302.3203`.

Chapter 15

Nonparametric regression: minimax upper and lower bounds

15.1 Introduction

We consider one of the two the most classical non-parametric problems in this example: estimating a regression function on a subset of the real line (the most classical problem being estimation of a density). In non-parametric regression, we assume there is an unknown function $f : \mathbb{R} \to \mathbb{R}$, where f belongs to a pre-determined class of functions $\mathcal{F}$; usually this class is parameterized by some type of smoothness guarantee. To make our problems concrete, we will assume that the unknown function f is L-Lipschitz and defined on $[0, 1]$. Let $\mathcal{F}$ denote this class. (For a fuller technical introduction into nonparametric estimation, see the book by Tsybakov [2].)

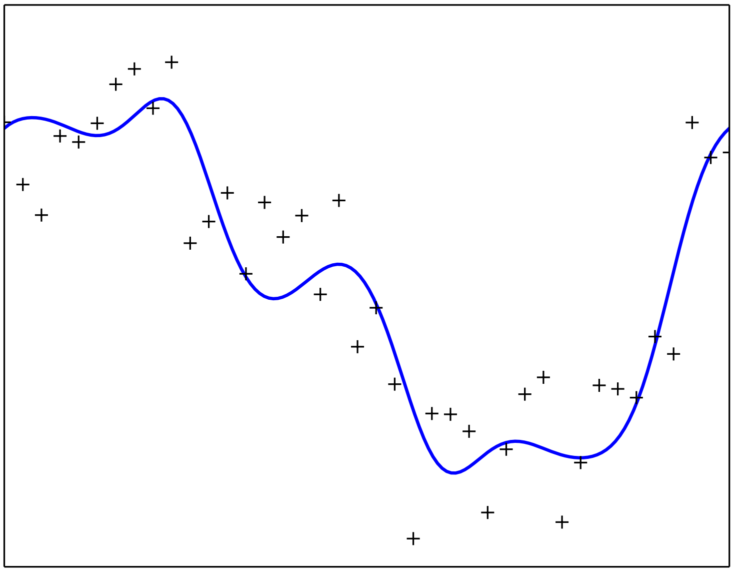

Figure 15.1. Observations in a non-parametric regression problem, with function f plotted. (Here $f(x) = \sin(2x + \cos^2(3x))$.)

In the standard non-parametric regression problem, we obtain observations of the form

$$Y_i = f(X_i) + \varepsilon_i \tag{15.1.1}$$

where ε_i are independent, mean zero conditional on X_i, and $\mathbb{E}[\varepsilon_i^2] \le \sigma^2$. See Figure 15.1 for an example. We also assume that we fix the locations of the X_i as $X_i = i/n \in [0, 1]$, that is, the X_i are evenly spaced in $[0, 1]$. Given n observations Y_i, we ask two questions: (1) how can we estimate f? and (2) what are the optimal rates at which it is possible to estimate f?

15.2 Kernel estimates of the function

A natural strategy is to place small "bumps" around the observed points, and estimate f in a neighborhood of a point x by weighted averages of the Y values for other points near x. We now formalize a strategy for doing this. Suppose we have a kernel function $K : \mathbb{R} \to \mathbb{R}_+$, which is continuous, not identically zero, has support $\operatorname{supp} K = [-1, 1]$, and satisfies the technical condition

$$\lambda_0 \sup_x K(x) \le \inf_{|x| \le 1/2} K(x), \tag{15.2.1}$$

where $\lambda_0 > 0$ (this says the kernel has some width to it). A natural example is the "tent" function given by $K_{\text{tent}}(x) = [1 - |x|]_+$, which satisfies inequality (15.2.1) with $\lambda_0 = 1/2$. See Fig. 15.2 for two examples, one the tent function and the other the function

$$K(x) = \mathbf{1}\left\{|x| < 1\right\} \exp\left(-\frac{1}{(x-1)^2}\right) \exp\left(-\frac{1}{(x+1)^2}\right),$$

which is infinitely differentiable and supported on $[-1, 1]$.

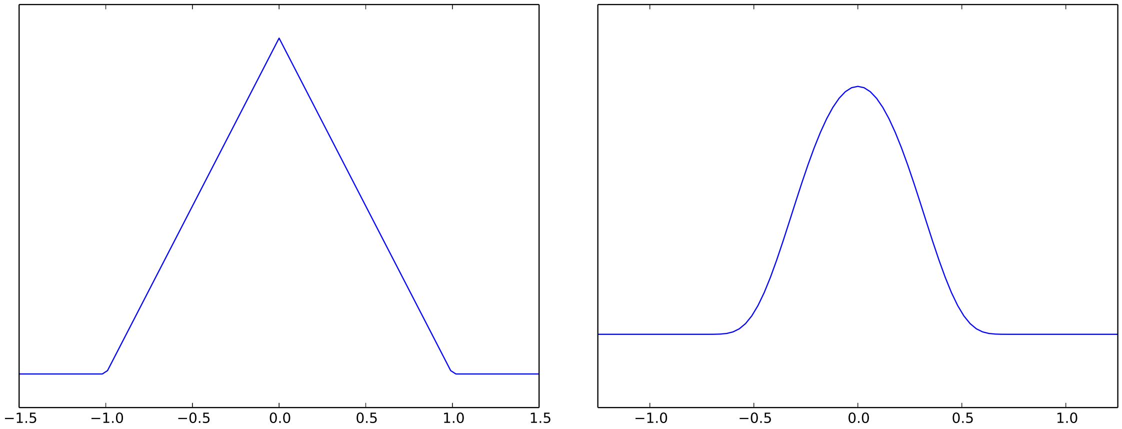

Figure 15.2: Left: "tent" kernel. Right: infinitely differentiable compactly supported kernel.

Now we consider a natural estimator of the function f based on observations (15.2.1) known as the Nadaraya-Watson estimator. Fix a bandwidth h, which we will see later smooths the estimated functions f. For all x, define weights

$$W_{ni}(x) := \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)}$$

and define the estimated function

$$\widehat{f}_n(x) := \sum_{i=1}^n Y_i W_{ni}(x).$$

The intuition here is that we have a locally weighted regression function, where points X_i in the neighborhood of x are given higher weight than further points. Using this function $\widehat{f}_n$ as our estimator, it is possible to provide a guarantee on the bias and variance of the estimated function at each point $x \in [0, 1]$.

Proposition 15.1. *Let the observation model* (15.1.1) *hold and assume condition* (15.2.1)*. In addition assume the bandwidth is suitably large that $h \geq 2/n$ and that the X_i are evenly spaced on* $[0, 1]$*. Then for any $x \in [0, 1]$, we have*

$$|\mathbb{E}[\widehat{f}_n(x)] - f(x)| \leq Lh \quad \textit{and} \quad \mathrm{Var}(\widehat{f}_n(x)) \leq \frac{2\sigma^2}{\lambda_0 nh}.$$

Proof To bound the bias, we note that (conditioning implicitly on X_i)

$$\mathbb{E}[\widehat{f}_n(x)] = \sum_{i=1}^n \mathbb{E}[Y_i W_{ni}(x)] = \sum_{i=1}^n \mathbb{E}[f(X_i)W_{ni}(x) + \varepsilon_i W_{ni}(x)] = \sum_{i=1}^n f(X_i)W_{ni}(x).$$

Thus we have that the bias is bounded as

$$\begin{aligned}\left|\mathbb{E}[\widehat{f}_n(x)] - f(x)\right| &\leq \sum_{i=1}^n |f(X_i) - f(x)|W_{ni}(x) \\ &\leq \sum_{i:|X_i - x| \leq h} |f(X_i) - f(x)|W_{ni}(x) \leq Lh \sum_{i=1}^n W_{ni}(x) = Lh.\end{aligned}$$

To bound the variance, we claim that

$$W_{ni}(x) \leq \min\left\{\frac{2}{\lambda_0 nh}, 1\right\}. \tag{15.2.2}$$

Indeed, we have that

$$W_{ni}(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)} = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j:|X_j - x| \leq h/2} K\left(\frac{X_j - x}{h}\right)} \leq \frac{K\left(\frac{X_i - x}{h}\right)}{\lambda_0 \sup_x K(x)|\{j : |X_j - x| \leq h/2\}|},$$

and because there are at least $nh/2$ indices satisfying $|X_j - x| \leq h$, we obtain the claim (15.2.2). Using the claim, we have

$$\begin{aligned}\mathrm{Var}(\widehat{f}_n(x)) = \mathbb{E}\left[\left(\sum_{i=1}^n (Y_i - f(X_i))W_{ni}(x)\right)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n \varepsilon_i W_{ni}(x)\right)^2\right] \\ &= \sum_{i=1}^n W_{ni}(x)^2 \mathbb{E}[\varepsilon_i^2] \leq \sum_{i=1}^n \sigma^2 W_{ni}(x)^2.\end{aligned}$$

Noting that $W_{ni}(x) \le 2/\lambda_0 nh$ and $\sum_{i=1}^n W_{ni}(x) = 1$, we have

$$\sum_{i=1}^n \sigma^2 W_{ni}(x)^2 \le \sigma^2 \max_i W_{ni}(x) \underbrace{\sum_{i=1}^n W_{ni}(x)}_{=1} \le \sigma^2 \frac{2}{\lambda_0 nh},$$

completing the proof. □

With the proposition in place, we can then provide a theorem bounding the worst case pointwise mean squared error for estimation of a function $f \in \mathcal{F}$.

Theorem 15.2. *Under the conditions of Proposition 15.1, choose $h = (\sigma^2/L^2\lambda_0)^{1/3} n^{-1/3}$. Then there exists a universal (numerical) constant $C < \infty$ such that for any $f \in \mathcal{F}$,*

$$\sup_{x \in [0,1]} \mathbb{E}[(\widehat{f}_n(x) - f(x))^2] \le C \left(\frac{L\sigma^2}{\lambda_0}\right)^{2/3} n^{-\frac{2}{3}}.$$

Proof Using Proposition 15.1, we have for any $x \in [0,1]$ that

$$\mathbb{E}[(\widehat{f}_n(x) - f(x))^2] = \Big(\mathbb{E}[\widehat{f}_n(x)] - f(x)\Big)^2 + \mathbb{E}[(\widehat{f}_n(x) - \mathbb{E}[\widehat{f}_n(x)])^2] \le \frac{2\sigma^2}{\lambda_0 nh} + L^2h^2.$$

Choosing h to balance the above bias/variance tradeoff, we obtain the thoerem. □

By integrating the result in Theorem 15.2 over the interval $[0,1]$, we immediately obtain the following corollary.

Corollary 15.3. *Under the conditions of Theorem 15.2, if we use the tent kernel K_{tent}, we have*

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f[\|\widehat{f}_n - f\|_2^2] \le C \left(\frac{L\sigma^2}{n}\right)^{2/3},$$

where C is a universal constant.

In Proposition 15.1, it is possible to show that a more clever choice of kernels—ones that are not always positive—can attain bias $\mathbb{E}[\widehat{f}_n(x)] - f(x) = \mathcal{O}(h^\beta)$ if f has Lipschitz $(\beta - 1)$th derivative. In this case, we immediately obtain that the rate can be improved to

$$\sup_x \mathbb{E}[(\widehat{f}_n(x) - f(x))^2] \le Cn^{-\frac{2\beta}{2\beta+1}},$$

and every additional degree of smoothness gives a corresponding improvement in convergence rate. We also remark that rates of this form, which are much larger than n^{-1}, are characteristic of non-parametric problems; essentially, we must adaptively choose a dimension that balances the sample size, so that rates of $1/n$ are difficult or impossible to achieve.

15.3 Minimax lower bounds on estimation with Assouad's method

Now we can ask whether the results we have given are in fact sharp; do there exist estimators attaining a faster rate of convergence than our kernel-based (locally weighted) estimator? Using Assouad's method, we show that, in fact, these results are all tight. In particular, we prove the following result on minimax estimation of a regression function $f \in \mathcal{F}$, where $\mathcal{F}$ consists of 1-Lipschitz functions defined on $[0, 1]$, in the $\|\cdot\|_2^2$ error, that is, $\|f - g\|_2^2 = \int_0^1 (f(t) - g(t))^2 dt$.

Theorem 15.4. *Let the observation points X_i be spaced evenly on $[0, 1]$, and assume the observation model* (15.1.1)*. Then there exists a universal constant $c > 0$ such that*

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) := \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[\|\widehat{f}_n - f\|_2^2\right] \ge c \left(\frac{\sigma^2}{n}\right)^{\frac{2}{3}}.$$

Deferring the proof of the theorem temporarily, we make a few remarks. It is in fact possible to show—using a completely identical technique—that if $\mathcal{F}_\beta$ denotes the class of functions with $\beta - 1$ derivatives, where the $(\beta - 1)$th derivative is Lipschitz, then

$$\mathfrak{M}_n(\mathcal{F}_\beta, \|\cdot\|_2^2) \ge c \left(\frac{\sigma^2}{n}\right)^{\frac{2\beta}{2\beta+1}}.$$

So for any smoothness class, we can never achieve the parametric σ^2/n rate, but we can come arbitrarily close. As another remark, which we do not prove, in dimensions $d \ge 1$, the minimax rate for estimation of functions f with Lipschitz $(\beta - 1)$th derivative scales as

$$\mathfrak{M}_n(\mathcal{F}_\beta, \|\cdot\|_2^2) \ge c \left(\frac{\sigma^2}{n}\right)^{\frac{2\beta}{2\beta+d}}.$$

This result can, similarly, be proved using a variant of Assouad's method; see, for example, the book of Györfi et al. [1, Chapter 3], which is available online. This is a striking example of the curse of dimensionality: the penalty for increasing dimension results in worse rates of convergence. For example, suppose that $\beta = 1$. In 1 dimension, we require $n \ge 90 \approx (.05)^{-3/2}$ observations to achieve accuracy .05 in estimation of f, while we require $n \ge 8000 = (.05)^{-(2+d)/2}$ even when the dimension $d = 4$, and $n \ge 64 \cdot 10^6$ observations even in 10 dimensions, which is a relatively small problem. That is, the problem is made exponentially more difficult by dimension increases.

We now turn to proving Theorem 15.4. To establish the result, we show how to construct a family of problems—indexed by binary vectors $v \in \{-1, 1\}^k$—so that our estimation problem satisfies the separation (14.1.1), then we show that information based on observing noisy versions of the functions we have defined is small. We then choose k to make our resulting lower bound as high as possible.

Construction of a separated family of functions To construct our separation in Hamming metric, as required by Eq. (14.1.1), fix some $k \in \mathbb{N}$; we will choose k later. This approach is somewhat different from our standard approach of using a fixed dimensionality and scaling the separation directly; in non-parametric problems, we scale the "dimension" itself to adjust the difficulty of the estimation problem. Define the function $g(x) = [1/2 - |x - 1/2|]_+$, so that g is

1-Lipschitz and is 0 outside of the interval $[0, 1]$. Then for any $v \in \{-1, 1\}^k$, define the "bump" functions

$$g_j(x) := \frac{1}{k} g\left(k\left(x - \frac{j-1}{k}\right)\right) \quad \text{and} \quad f_v(x) := \sum_{j=1}^k v_j g_j(x),$$

which we see is 1-Lipschitz. Now, consider any function $f : [0, 1] \to \mathbb{R}$, and let E_j be shorthand for the intervals $E_j = [(j-1)/k, j/k]$ for $j = 1, \ldots, k$. We must find a mapping identifying a function f with points in the hypercube $\{-1, 1\}^k$. To that end, we may define a vector $\widehat{\mathsf{v}}(f) \in \{-1, 1\}^k$ by

$$\widehat{\mathsf{v}}_j(f) = \operatorname*{argmin}_{s \in \{-1,1\}} \int_{E_j} (f(t) - s g_j(t))^2 \, dt.$$

We claim that for any function f,

$$\left(\int_{E_j} (f(t) - f_v(t))^2 dt\right)^{\frac{1}{2}} \ge \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\} \left(\int_{E_j} f_v(t)^2 dt\right)^{\frac{1}{2}}. \tag{15.3.1}$$

Indeed, on the set E_j, we have $v_j g_j(t) = f_v(t)$, and thus $\int_{E_j} g_j(t)^2 dt = \int_{E_j} f_v(t)^2 dt$. Then by the triangle inequality, we have

$$\begin{aligned} 2 \cdot \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\} \left(\int_{E_j} g_j(t)^2 dt\right)^{\frac{1}{2}} &= \left(\int_{E_j} \left((\widehat{\mathsf{v}}_j(f) - v_j) g_j(t)\right)^2 dt\right)^{\frac{1}{2}} \\ &\le \left(\int_{E_j} (f(t) - v_j g_j(t))^2 \, dt\right)^{\frac{1}{2}} + \left(\int_{E_j} (f(t) - \widehat{\mathsf{v}}_j(f) g_j(t))^2 \, dt\right)^{\frac{1}{2}} \\ &\le 2 \left(\int_{E_j} (f(t) - f_v(t))^2 \, dt\right)^{\frac{1}{2}}, \end{aligned}$$

by definition of the sign $\widehat{\mathsf{v}}_j(f)$.

With the definition of $\widehat{\mathsf{v}}$ and inequality (15.3.1), we see that for any vector $v \in \{-1, 1\}^k$, we have

$$\|f - f_v\|_2^2 = \sum_{j=1}^k \int_{E_j} (f(t) - f_v(t))^2 \, dt \ge \sum_{j=1}^k \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\} \int_{E_j} f_v(t)^2 dt.$$

In particular, we know that

$$\int_{E_j} f_v(t)^2 dt = \frac{1}{k^2} \int_0^{1/k} g(kt)^2 dt = \frac{1}{k^3} \int_0^1 g(u)^2 du \ge \frac{c}{k^3},$$

where c is a numerical constant. In particular, we have the desired separation

$$\|f - f_v\|_2^2 \ge \frac{c}{k^3} \sum_{j=1}^k \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\}. \tag{15.3.2}$$

Bounding the binary testing error Let P_v^n denote the distribution of the n observations $Y_i = f_v(X_i) + \varepsilon_i$ when f_v is the true regression function. Then inequality (15.3.2) implies via Assouad's lemma that

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) \ge \frac{c}{k^3} \sum_{j=1}^k \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\rm TV}\right]. \tag{15.3.3}$$

Now, we use convexity and Pinsker's inequality to note that

$$\left\|P_{+j}^n - P_{-j}^n\right\|_{\rm TV}^2 \le \max_v \left\|P_{v,+j}^n - P_{v,-j}^n\right\|_{\rm TV}^2 \le \max_v \frac{1}{2} D_{\rm kl}\left(P_{v,+j}^n \| P_{v,-j}^n\right).$$

For any two functions f_v and $f_{v'}$, we have that the observations Y_i are independent and normal with means $f_v(X_i)$ or $f_{v'}(X_i)$, respectively. Thus

$$\begin{aligned} D_{\rm kl}\left(P_v^n \| P_{v'}^n\right) &= \sum_{i=1}^n D_{\rm kl}\left(\mathsf{N}(f_v(X_i), \sigma^2) \| \mathsf{N}(f_{v'}(X_i), \sigma^2)\right) \\ &= \sum_{i=1}^n \frac{1}{2\sigma^2}(f_v(X_i) - f_{v'}(X_i))^2. \end{aligned} \tag{15.3.4}$$

Now we must show that the expression (15.3.4) scales more slowly than n, which we will see must be the case as whenever $d_{\rm ham}(v, v') \le 1$. Intuitively, most of the observations have the same distribution by our construction of the f_v as bump functions; let us make this rigorous.

We may assume without loss of generality that $v_j = {v'}_j$ for $j > 1$. As the $X_i = i/n$, we thus have that only X_i for i near 1 can have non-zero values in the tensorization (15.3.4). In particular,

$$f_v(i/n) = f_{v'}(i/n) \text{ for all } i \text{ s.t. } \frac{i}{n} \ge \frac{2}{k}, \text{ i.e. } i \ge \frac{2n}{k}.$$

Rewriting expression (15.3.4), then, and noting that $f_v(x) \in [-1/k, 1/k]$ for all x by construction, we have

$$\sum_{i=1}^n \frac{1}{2\sigma^2}(f_v(X_i) - f_{v'}(X_i))^2 \le \sum_{i=1}^{2n/k} \frac{1}{2\sigma^2}(f_v(X_i) - f_{v'}(X_i))^2 \le \frac{1}{2\sigma^2}\frac{2n}{k}\frac{1}{k^2} = \frac{n}{k^3\sigma^2}.$$

Combining this with inequality (15.3.4) and the minimax bound (15.3.3), we obtain

$$\left\|P_{+j}^n - P_{-j}^n\right\|_{\rm TV} \le \sqrt{\frac{n}{2k^3\sigma^2}},$$

so

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) \ge \frac{c}{k^3}\sum_{j=1}^k \left[1 - \sqrt{\frac{n}{2k^3\sigma^2}}\right].$$

Choosing k for optimal tradeoffs Now we simply choose k; in particular, setting

$$k = \left\lceil \left(\frac{n}{2\sigma^2}\right)^{1/3} \right\rceil \quad \text{then} \quad 1 - \sqrt{\frac{n}{2k^3\sigma^2}} \ge 1 - \sqrt{1/4} = \frac{1}{2},$$

and we arrive at

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) \ge \frac{c}{k^3}\sum_{j=1}^k \frac{1}{2} = \frac{c}{2k^2} \ge c'\left(\frac{\sigma^2}{n}\right)^{2/3},$$

where $c' > 0$ is a universal constant. Theorem 15.4 is proved.

Bibliography

[1] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[2] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Chapter 16

Global Fano Method

In this chapter, we extend the techniques of Chapter 13.4 on Fano's method (the local Fano method) to a more global construction. In particular, we show that, rather than constructing a local packing, choosing a scaling $\delta > 0$, and then optimizing over this δ, it is actually, in many cases, possible to prove lower bounds on minimax error directly using packing and covering numbers (metric entropy and packing entropy). The material in this chapter is based on a paper of Yang and Barron [5].

16.1 A mutual information bound based on metric entropy

To begin, we recall the classical Fano inequality, which says that for any Markov chain $V \to X \to \widehat{V}$, where V is uniform on the finite set $\mathcal{V}$, we have

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

(Recall Corollary 13.9.) Thus, there are two ingredients in proving lower bounds on the error in a hypothesis test: upper bounding the mutual information and lower bounding the size $|\mathcal{V}|$. Here, we state a proposition doing the former.

Before stating our result, we require a bit of notation. First, we assume that V is drawn from a distribution μ, and conditional on $V = v$, assume the sample $X \sim P_v$. Then a standard calculation (or simply the definition of mutual information; recall equation (13.4.4)) gives that

$$I(V;X) = \int D_{\rm kl}\left(P_v \| \overline{P}\right) d\mu(v), \quad \text{where} \quad \overline{P} = \int P_v d\mu(v). \tag{16.1.1}$$

Now, we show how to connect this mutual information quantity to a covering number of a set of distributions.

Assume that for all v, we have $P_v \in \mathcal{P}$, where $\mathcal{P}$ is a collection of distributions. In analogy with Definition 13.1, we say that the collection of distributions $\{Q_i\}_{i=1}^N$ form an ϵ-cover of $\mathcal{P}$ in KL-divergence if for all $P \in \mathcal{P}$, there exists some i such that $D_{\rm kl}\left(P \| Q_i\right) \leq \epsilon^2$. With this, we may define the KL-covering number of the set $\mathcal{P}$ as

$$N_{\rm kl}\left(\epsilon, \mathcal{P}\right) := \inf\left\{N \in \mathbb{N} \mid \exists\, Q_i, i = 1, \ldots, N, \ \sup_{P \in \mathcal{P}} \min_i D_{\rm kl}\left(P \| Q_i\right) \leq \epsilon^2\right\}, \tag{16.1.2}$$

where $N_{\rm kl}\left(\epsilon, \mathcal{P}\right) = +\infty$ if no such cover exists. With definition (16.1.2) in place, we have the following proposition.

Proposition 16.1. *Under conditions of the preceding paragraphs, we have*

$$I(V;X) \le \inf_{\epsilon>0} \left\{ \epsilon^2 + \log N_{\rm kl}(\epsilon, \mathcal{P}) \right\}. \tag{16.1.3}$$

Proof First, we claim that

$$\int D_{\rm kl}\left(P_v \| \overline{P}\right) d\mu(v) \le \int D_{\rm kl}\left(P_v \| Q\right) d\mu(v) \tag{16.1.4}$$

for any distribution Q. Indeed, briefly, we have

$$\begin{aligned} \int D_{\rm kl}\left(P_v \| \overline{P}\right) d\mu(v) &= \int_{\mathcal{V}} \int_{\mathcal{X}} dP_v \log \frac{dP_v}{d\overline{P}} d\mu(v) = \int_{\mathcal{V}} \int_{\mathcal{X}} dP_v \left[\log \frac{dP_v}{Q} + \log \frac{dQ}{d\overline{P}}\right] d\mu(v) \\ &= \int_{\mathcal{V}} D_{\rm kl}\left(P_v \| Q\right) d\mu(v) + \int_{\mathcal{X}} \underbrace{\int_{\mathcal{V}} d\mu(v) dP_v}_{=d\overline{P}} \log \frac{dQ}{d\overline{P}} \\ &= \int D_{\rm kl}\left(P_v \| Q\right) d\mu(v) - D_{\rm kl}\left(\overline{P} \| Q\right) \le \int D_{\rm kl}\left(P_v \| Q\right) d\mu(v), \end{aligned}$$

so that inequality (16.1.4) holds. By carefully choosing the distribution Q in the upper bound (16.1.4), we obtain the proposition.

Now, assume that the distributions Q_i, $i = 1, \ldots, N$ form an ϵ^2-cover of the family $\mathcal{P}$, meaning that

$$\min_{i \in [N]} D_{\rm kl}\left(P \| Q_i\right) \le \epsilon^2 \quad \text{for all } P \in \mathcal{P}.$$

Let p_v and q_i denote the densities of P_v and Q_i with respect to some fixed base measure on $\mathcal{X}$ (the choice of based measure does not matter). Then definining the distribution $Q = (1/N) \sum_{i=1}^N Q_i$, we obtain for any v that in expectation over $X \sim P_v$,

$$\begin{aligned} D_{\rm kl}\left(P_v \| Q\right) &= \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{q(X)}\right] = \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{N^{-1} \sum_{i=1}^n q_i(X)}\right] \\ &= \log N + \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{\sum_{i=1}^N q_i(X)}\right] \le \log N + \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{\max_i q_i(X)}\right] \\ &\le \log N + \min_i \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{q_i(X)}\right] = \log N + \min_i D_{\rm kl}\left(P_v \| Q_i\right). \end{aligned}$$

By our assumption that the Q_i form a cover, this gives the desired result, as $\epsilon \ge 0$ was arbitrary, as was our choice of the cover. □

By a completely parallel proof, we also immediately obtain the following corollary.

Corollary 16.2. *Assume that $X_1, \ldots, X_n$ are drawn i.i.d. from P_v conditional on $V = v$. Let $N_{\rm kl}(\epsilon, \mathcal{P})$ denote the KL-covering number of a collection $\mathcal{P}$ containing the distributions (over a single observation) P_v for all $v \in \mathcal{V}$. Then*

$$I(V; X_1, \ldots, X_n) \le \inf_{\epsilon \ge 0} \left\{ n\epsilon^2 + \log N_{\rm kl}(\epsilon, \mathcal{P}) \right\}.$$

With Corollary 16.2 and Proposition 16.1 in place, we thus see that the global covering numbers in KL-divergence govern the behavior of information.

We remark in passing that the quantity (16.1.3), and its i.i.d. analogue in Corollary 16.2, is known as the *index of resolvability*, and it controls estimation rates and redundancy of coding schemes for unknown distributions in a variety of scenarios; see, for example, Barron [1] and Barron and Cover [2]. It is also similar to notions of complexity in Dudley's entropy integral (cf. Dudley [3]) in empirical process theory, where the fluctuations of an empirical process are governed by a tradeoff between covering number and approximation of individual terms in the process.

16.2 Minimax bounds using global packings

There is now a four step process to proving minimax lower bounds using the global Fano method. Our starting point is to recall the Fano minimax lower bound in Proposition 13.10, which begins with the construction of a set of points $\{\theta(P_v)\}_{v\in\mathcal{V}}$ that form a 2δ-packing of a set Θ in some ρ-semimetric. With this inequality in mind, we perform the following four steps:

(i) *Bound the packing entropy.* Give a lower bound on the packing number of the set Θ with 2δ-separation (call this lower bound $M(\delta)$).

(ii) *Bound the metric entropy.* Give an upper bound on the KL-metric entropy of the class $\mathcal{P}$ of distributions containing all the distributions P_v, that is, an upper bound on $\log N_{\rm kl}(\epsilon, \mathcal{P})$.

(iii) *Find the critical radius.* Noting as in Corollary 16.2 that with n i.i.d. observations, we have

$$I(V; X_1, \ldots, X_n) \le \inf_{\epsilon\ge 0}\left\{n\epsilon^2 + \log N_{\rm kl}(\epsilon, \mathcal{P})\right\},$$

we now balance the information $I(V; X_1^n)$ and the packing entropy $\log M(\delta)$. To that end, we choose ϵ_n and $\delta > 0$ at the *critical radius*, defined as follows: choose the any ϵ_n such that

$$n\epsilon_n^2 \ge \log N_{\rm kl}(\epsilon_n, \mathcal{P}),$$

and choose the largest $\delta_n > 0$ such that

$$\log M(\delta_n) \ge 4n\epsilon_n^2 + 2\log 2 \ge 2N_{\rm kl}(\epsilon_n, \mathcal{P}) + 2n\epsilon_n^2 + 2\log 2 \ge 2\left(I(V; X_1^n) + \log 2\right).$$

(We could have chosen the ϵ_n attaining the infimum in the mutual information, but this way we need only an upper bound on $\log N_{\rm kl}(\epsilon, \mathcal{P})$.)

(iv) *Apply the Fano minimax bound.* Having chosen δ_n and ϵ_n as above, we immediately obtain that for the Markov chain $V \to X_1^n \to \widehat{V}$,

$$\mathbb{P}(V \ne \widehat{V}) \ge 1 - \frac{I(V; X_1, \ldots, X_n) + \log 2}{\log M(\delta_n)} \ge 1 - \frac{1}{2} = \frac{1}{2},$$

and thus, applying the Fano minimax bound in Proposition 13.10, we obtain

$$\mathfrak{M}_n(\theta(\mathcal{P}); \Phi\circ\rho) \ge \frac{1}{2}\Phi(\delta_n).$$

16.3 Example: non-parametric regression

In this section, we flesh out the outline in the prequel to show how to obtain a minimax lower bound for a non-parametric regression problem directly with packing and metric entropies. In this example, we sketch the result, leaving explicit constant calculations to the dedicated reader. Nonetheless, we recover an analogue of Theorem 15.4 on minimax risks for estimation of 1-Lipschitz functions on $[0, 1]$.

We use the standard non-parametric regression setting, where our observations Y_i follow the independent noise model (15.1.1), that is, $Y_i = f(X_i) + \varepsilon_i$. Letting

$$\mathcal{F} := \{f : [0, 1] \to \mathbb{R},\ f(0) = 0,\ f \text{ is Lipschitz}\}$$

be the family of 1-Lipschitz functions with $f(0) = 0$, we have

Proposition 16.3. *There exists a universal constant $c > 0$ such that*

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_\infty) := \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[\|\widehat{f}_n - f\|_\infty\right] \ge c \left(\frac{\sigma^2}{n}\right)^{1/3},$$

where $\widehat{f}_n$ is constructed based on the n independent observations $f(X_i) + \varepsilon_i$.

The rate in Proposition 16.3 is sharp to within factors logarithmic in n; a more precise analysis of the upper and lower bounds on the minimax rate yields

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_\infty) := \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[\|\widehat{f}_n - f\|_\infty\right] \asymp \left(\frac{\sigma^2 \log n}{n}\right)^{1/3}.$$

See, for example, Tsybakov [4] for a proof of this fact.

Proof Our first step is to note that the covering and packing numbers of the set $\mathcal{F}$ in the ℓ_∞ metric satisfy

$$\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \log M(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \frac{1}{\delta}. \tag{16.3.1}$$

To see this, fix some $\delta \in (0, 1)$ and assume for simplicity that $1/\delta$ is an integer. Define the sets $E_j = [\delta(j-1), \delta j)$, and for each $v \in \{-1, 1\}^{1/\delta}$ define $h_v(x) = \sum_{j=1}^{1/\delta} v_j \mathbf{1}\{x \in E_j\}$. Then define the function $f_v(t) = \int_0^t h_v(t)dt$, which increases or decreases linearly on each interval of width δ in $[0, 1]$. Then these f_v form a 2δ-packing and a 2δ-cover of $\mathcal{F}$, and there are $2^{1/\delta}$ such f_v. Thus the asymptotic approximation (16.3.1) holds. **TODO: Draw a picture**

Now, if for some fixed $x \in [0, 1]$ and $f, g \in \mathcal{F}$ we define P_f and P_g to be the distributions of the observations $f(x) + \varepsilon$ or $g(x) + \varepsilon$, we have that

$$D_{\rm kl}\left(P_f \| P_g\right) = \frac{1}{2\sigma^2}(f(X_i) - g(X_i))^2 \le \frac{\|f - g\|_\infty^2}{2\sigma^2},$$

and if P_f^n is the distribution of the n observations $f(X_i) + \varepsilon_i$, $i = 1, \ldots, n$, we also have

$$D_{\rm kl}\left(P_f^n \| P_g^n\right) = \sum_{i=1}^n \frac{1}{2\sigma^2}(f(X_i) - g(X_i))^2 \le \frac{n}{2\sigma^2} \|f - g\|_\infty^2.$$

In particular, this implies the upper bound

$$\log N_{\rm kl}\left(\epsilon, \mathcal{P}\right) \lesssim \frac{1}{\sigma\epsilon}$$

on the KL-metric entropy of the class $\mathcal{P} = \{P_f : f \in \mathcal{F}\}$, as $\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \delta^{-1}$. Thus we have completed steps (i) and (ii) in our program above.

It remains to choose the critical radius in step (iii), but this is now relatively straightforward: by choosing $\epsilon_n \asymp (1/\sigma n)^{1/3}$, and whence $n\epsilon_n^2 \asymp (n/\sigma^2)^{1/3}$, we find that taking $\delta \asymp (\sigma^2/n)^{1/3}$ is sufficient to ensure that $\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \gtrsim \delta^{-1} \geq 4n\epsilon_n^2 + 2\log 2$. Thus we have

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_\infty) \gtrsim \delta_n \cdot \frac{1}{2} \gtrsim \left(\frac{\sigma^2}{n}\right)^{1/3}$$

as desired. □

Bibliography

[1] A. R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics*, pages 561–576. Kluwer Academic, 1991.

[2] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[3] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

[4] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[5] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

Appendix A

Review of Convex Analysis

In this appendix, we review several results in convex analysis that are useful for our purposes. We give only a cursory study here, identifying the basic results and those that will be of most use to us; the field of convex analysis as a whole is vast. The study of convex analysis and optimization has become very important practically in the last fourty to fifty years for a few reasons, the most important of which is probably that convex optimization problems—those optimization problems in which the objective and constraints are convex—are tractable, while many others are not. We do not focus on optimization ideas here, however, building only some analytic tools that we will find useful. We borrow most of our results from Hiriart-Urruty and Lemaréchal [5], focusing mostly on the finite-dimensional case (though we present results that apply in infinite dimensional cases with proofs that extend straightforwardly, and we do not specify the domains of our functions unless necessary), as we require no results from infinite-dimensional analysis.

In addition, we abuse notation and assume that the range of any function is the *extended real line*, meaning that if $f : C \to \mathbb{R}$ we mean that $f(x) \in \mathbb{R} \cup \{-\infty, +\infty\}$, where $-\infty$ and $+\infty$ are infinite and satisfy $a + \infty = +\infty$ and $a - \infty = -\infty$ for any $a \in \mathbb{R}$. However, we assume that our functions are *proper*, meaning that $f(x) > -\infty$ for all x, as this allows us to avoid annoying pathologies.

A.1 Convex sets

We begin with the simplest and most important object in convex analysis, a convex set.

Definition A.1. *A set C is* convex *if for all $\lambda \in [0, 1]$ and all $x, y \in C$, we have*

$$\lambda x + (1 - \lambda)y \in C.$$

An important restriction of convex sets is to *closed* convex sets, those convex sets that are, well, closed.

TODO: Picture

We now consider two operations that extend sets, convexifying them in nice ways.

Definition A.2. *The* affine hull *of a set C is the smallest affine set containing C. That is,*

$$\mathrm{aff}(C) := \left\{ \sum_{i=1}^{k} \lambda_i x_i : k \in \mathbb{N}, x_i \in C, \lambda \in \mathbb{R}^k, \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

Associated with any set is also its convex hull:

Definition A.3. *The* convex hull *of a set $C \subset \mathbb{R}^d$, denoted* $\mathrm{Conv}(C)$*, is the intersection of all convex sets containing C.*

TODO: picture

An almost immediate associated result is that the convex hull of a set is equal to the set of all convex combinations of points in the set.

Proposition A.1. *Let C be an arbitrary set. Then*

$$\mathrm{Conv}(C) = \left\{ \sum_{i=1}^k \lambda_i x_i : k \in \mathbb{N}, x_i \in C, \lambda \in \mathbb{R}_+^k, \sum_{i=1}^k \lambda_i = 1 \right\}.$$

Proof Call T the set on the right hand side of the equality in the proposition. Then $T \supset C$ is clear, as we may simply take $\lambda_1 = 1$ and vary $x \in C$. Moreover, the set $T \subset \mathrm{Conv}(C)$, as any convex set containing C must contain all convex combinations of its elements; similarly, any convex set $S \supset C$ must have $S \supset T$.

Thus if we show that T is convex, then we are done. Take any two points $x, y \in T$. Then $x = \sum_{i=1}^k \alpha_i x_i$ and $y = \sum_{i=1}^l \beta_i y_i$ for $x_i, y_i \in C$. Fix $\lambda \in [0,1]$. Then $(1-\lambda)\beta_i \ge 0$ and $\lambda\alpha_i \ge 0$ for all i,

$$\lambda \sum_{i=1}^k \alpha_i + (1-\lambda) \sum_{i=1}^l \beta_i = \lambda + (1-\lambda) = 1,$$

and $\lambda x + (1-\lambda) y$ is a convex combination of the points x_i and y_i weighted by $\lambda\alpha_i$ and $(1-\lambda)\beta_i$, respectively. So $\lambda x + (1-\lambda) y \in T$ and T is convex. □

We also give one more definition, which is useful for dealing with some pathalogical cases in convex analysis, as it allows us to assume many sets are full-dimensional.

Definition A.4. *The* relative interior *of a set C is the interior of C relative to its affine hull, that is,*

$$\mathrm{relint}(C) := \{x \in C : B(x, \epsilon) \cap \mathrm{aff}(C) \subset C \text{ for some } \epsilon > 0\},$$

where $B(x, \epsilon) := \{y : \|y - x\| < \epsilon\}$ denotes the open ball of radius ϵ centered at x.

An example may make Definition A.4 clearer.

Example A.2 (Relative interior of a disc)**:** Consider the (convex) set

$$C = \left\{ x \in \mathbb{R}^d : x_1^2 + x_2^2 \le 1, \ x_j = 0 \text{ for } j \in \{3, \ldots, d\} \right\}.$$

The affine hull $\mathrm{aff}(C) = \mathbb{R}^2 \times \{0\} = \{(x_1, x_2, 0, \ldots, 0) : x_1, x_2 \in \mathbb{R}\}$ is simply the (x_1, x_2)-plane in $\mathbb{R}^d$, while the relative interior $\mathrm{relint}(C) = \{x \in \mathbb{R}^d : x_1^2 + x_2^2 < 1\} \cap \mathrm{aff}(C)$ is the "interior" of the 2-dimensional disc in $\mathbb{R}^d$. ♣

In finite dimensions, we may actually restrict the definition of the convex hull of a set C to convex combinations of a bounded number (the dimension plus one) of the points in C, rather than arbitrary convex combinations as required by Proposition A.1. This result is known as *Carathéodory's theorem.*

Theorem A.3. *Let $C \subset \mathbb{R}^d$. Then $x \in \mathrm{Conv}(C)$ if and only if there exist points $x_1, \ldots, x_{d+1} \in C$ and $\lambda \in \mathbb{R}_+^{d+1}$ with $\sum_{i=1}^{d+1} \lambda_i = 1$ such that*

$$x = \sum_{i=1}^{d+1} \lambda_i x_i.$$

Proof It is clear that if x can be represetned as such a sum, then $x \in \mathrm{Conv}(C)$. Conversely, Proposition A.1 implies that for any $x \in \mathrm{Conv}(C)$ we have

$$x = \sum_{i=1}^{k} \lambda_i x_i, \quad \lambda_i \geq 0, \ \sum_{i=1}^{k} \lambda_i = 1, \ x_i \in C$$

for some λ_i, x_i. Assume that $k > d+1$ and $\lambda_i > 0$ for each i, as otherwise, there is nothing to prove. Then we know that the points $x_i - x_1$ are certainly linearly dependent (as there are $k-1 > d$ of them), and we can find (not identically zero) values $\alpha_2, \ldots, \alpha_k$ such that $\sum_{i=2}^k \alpha_i(x_i - x_1) = 0$. Let $\alpha_1 = -\sum_{i=2}^k \alpha_i$ to obtain that we have both

$$\sum_{i=1}^{k} \alpha_i x_i = 0 \ \text{ and } \ \sum_{i=1}^{k} \alpha_i = 0. \tag{A.1.1}$$

Notably, the equalities (A.1.1) imply that at least one $\alpha_i > 0$, and if we define $\lambda^* = \min_{i:\alpha_i > 0} \frac{\lambda_i}{\alpha_i} > 0$, then setting $\lambda_i' = \lambda_i - \lambda^* \alpha_i$ we have

$$\lambda_i' \geq 0 \text{ for all } i, \quad \sum_{i=1}^{k} \lambda_i' = \sum_{i=1}^{k} \lambda_i - \lambda^* \sum_{i=1}^{k} \alpha_i = 1, \quad \text{and} \quad \sum_{i=1}^{k} \lambda_i' x_i = \sum_{i=1}^{k} \lambda_i x_i - \lambda^* \sum_{i=1}^{k} \alpha_i x_i = x.$$

But we know that at least one of the $\lambda_i' = 0$, so that we could write x as a convex combination of $k-1$ elements. Repeating this strategy until $k = d+1$ gives the theorem. □

A.1.1 Operations preserving convexity

We now touch on a few simple results about operations that preserve convexity of convex sets. First, we make the following simple observation.

Observation A.4. *Let C be a convex set. Then $C = \mathrm{Conv}(C)$.*

Observation A.4 is clear, as we have $C \subset \mathrm{Conv}(C)$, while any other convex $S \supset C$ clearly satisfies $S \supset \mathrm{Conv}(C)$. Secondly, we note that intersections preserve convexity.

Observation A.5. *Let $\{C_\alpha\}_{\alpha \in \mathcal{A}}$ be an arbitrary collection of convex sets. Then*

$$C = \bigcap_{\alpha \in \mathcal{A}} C_\alpha$$

is convex. Moreover, if C_α is closed for each α, then C is closed as well.

The convexity property follows because if $x_1 \in C$ and $x_2 \in C$, then clearly $x_1, x_2 \in C_\alpha$ for all $\alpha \in \mathcal{A}$, and moreover $\lambda x_1 + (1-\lambda)x_2 \in C_\alpha$ for all α and any $\lambda \in [0,1]$. The closure property is standard. In addition, we note that closing a convex set maintains convexity.

Observation A.6. *Let C be convex. Then $\mathrm{cl}(C)$ is convex.*

To see this, we note that if $x, y \in \mathrm{cl}(C)$ and $x_n \to x$ and $y_n \to y$ (where $x_n, y_n \in C$), then for any $\lambda \in [0,1]$, we have $\lambda x_n + (1-\lambda)y_n \in C$ and $\lambda x_n + (1-\lambda)y_n \to \lambda x + (1-\lambda)y$. Thus we have $\lambda x + (1-\lambda)y \in \mathrm{cl}(C)$ as desired.

Observation A.6 also implies the following result.

Observation A.7. *Let D be an arbitrary set. Then*

$$\bigcap \{C : C \supset D,\ C \text{ is convex}\} = \mathrm{cl}\,\mathrm{Conv}(D).$$

Proof Let T denote the leftmost set. It is clear that $T \subset \mathrm{cl}\,\mathrm{Conv}(D)$ as $\mathrm{cl}\,\mathrm{Conv}(D)$ is a closed convex set (by Observation A.6) containing D. On the other hand, if $C \supset D$ is a closed convex set, then $C \supset \mathrm{Conv}(D)$, while the closedness of C implies it also contains the closure of $\mathrm{Conv}(D)$. Thus $T \supset \mathrm{cl}\,\mathrm{Conv}(D)$ as well. □

TODO: picture

As our last consideration of operations that preserve convexity, we consider what is known as the perspective of a set. To define this set, we need to define the perspective function, which, given a point $(x,t) \in \mathbb{R}^d \times \mathbb{R}_{++}$ (here $\mathbb{R}_{++} = \{t : t > 0\}$ denotes strictly positive points), is defined as

$$\mathrm{pers}(x,t) = \frac{x}{t}.$$

We have the following definition.

Definition A.5. *Let $C \subset \mathbb{R}^d \times \mathbb{R}_+$ be a set. The* perspective transform *of C, denoted by* $\mathrm{pers}(C)$, *is*

$$\mathrm{pers}(C) := \left\{ \frac{x}{t} : (x,t) \in C \text{ and } t > 0 \right\}.$$

This corresponds to taking all the points $z \in C$, normalizing them so their last coordinate is 1, and then removing the last coordinate. (For more on perspective functions, see Boyd and Vandenberghe [4, Chapter 2.3.3].)

It is interesting to note that the perspective of a convex set is convex. First, we note the following.

Lemma A.8. *Let $C \subset \mathbb{R}^{d+1}$ be a compact line segment, meaning that $C = \{\lambda x + (1-\lambda)y : \lambda \in [0,1]\}$, where $x_{d+1} > 0$ and $y_{d+1} > 0$. Then $\mathrm{pers}(C) = \{\lambda\,\mathrm{pers}(x) + (1-\lambda)\,\mathrm{pers}(y) : \lambda \in [0,1]\}$.*

Proof Let $\lambda \in [0,1]$. Then

$$\begin{aligned}
\mathrm{pers}(\lambda x + (1-\lambda)y) &= \frac{\lambda x_{1:d} + (1-\lambda)y_{1:d}}{\lambda x_{d+1} + (1-\lambda)y_{d+1}} \\
&= \frac{\lambda x_{d+1}}{\lambda x_{d+1} + (1-\lambda)y_{d+1}} \frac{x_{1:d}}{x_{d+1}} + \frac{(1-\lambda)y_{d+1}}{\lambda x_{d+1} + (1-\lambda)y_{d+1}} \frac{y_{1:d}}{y_{d+1}} \\
&= \theta\,\mathrm{pers}(x) + (1-\theta)\,\mathrm{pers}(y),
\end{aligned}$$

where $x_{1:d}$ and $y_{1:d}$ denote the vectors of the first d components of x and y, respectively, and

$$\theta = \frac{\lambda x_{d+1}}{\lambda x_{d+1} + (1 - \lambda) y_{d+1}} \in [0, 1].$$

Sweeping λ from 0 to 1 sweeps $\theta \in [0, 1]$, giving the result. □

Based on Lemma A.8, we immediately obtain the following proposition.

Proposition A.9. *Let $C \subset \mathbb{R}^d \times \mathbb{R}_{++}$ be a convex set. Then* $\mathrm{pers}(C)$ *is convex.*

Proof Let $x, y \in C$ and define $L = \{\lambda x + (1 - \lambda) y : \lambda \in [0, 1]\}$ to be the line segment between them. By Lemma A.8, $\mathrm{pers}(L) = \{\lambda \,\mathrm{pers}(x) + (1 - \lambda) \,\mathrm{pers}(y) : \lambda \in [0, 1]\}$ is also a (convex) line segment, and we have $\mathrm{pers}(L) \subset \mathrm{pers}(C)$ as necessary. □

A.1.2 Representation and separation of convex sets

We now consider some properties of convex sets, showing that (1) they have nice separation properties—we can put hyperplanes between them—and (2) this allows several interesting representations of convex sets. We begin with the separation properties, developing them via the existence of projections. Interestingly, this existence of projections does not rely on any finite-dimensional structure, and can even be shown to hold in arbitrary Banach spaces (assuming the axiom of choice) [6]. We provide the results in a *Hilbert space*, meaning a complete vector space for which there exists an inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|$ given by $\|x\|^2 = \langle x, x \rangle$. We first note that projections exist.

Theorem A.10 (Projections)**.** *Let C be a closed convex set. Then for any x, there exists a unique point $\pi_C(x)$ minimizing $\|y - x\|$ over $y \in C$. Moreover, this point is characterized by the inequality*

$$\langle \pi_C(x) - x, y - \pi_C(x) \rangle \ge 0 \;\; \text{for all } y \in C. \tag{A.1.2}$$

Proof The existence and uniqueness of the projection follows from the parallelogram identity, that is, that for any x, y we have $\|x - y\|^2 + \|x + y\|^2 = 2(\|x\|^2 + \|y\|^2)$, which follows by noting that $\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2 \langle x, y \rangle$. Indeed, let $\{y_n\} \subset C$ be a sequence such that

$$\|y_n - x\| \to \inf_{y \in C} \|y - x\| =: p_\star$$

as $n \to \infty$, where $p_\star$ is the infimal value. We show that y_n is Cauchy, so that there exists a (unique) limit point of the sequence. Fix $\epsilon > 0$ and let N be such that $n \ge N$ implies $\|y_n - x\|^2 \le p_\star^2 + \epsilon^2$. Let $m, n \ge N$. Then by the parallelogram identity,

$$\|y_n - y_m\|^2 = \|(x - y_n) - (x - y_m)\|^2 = 2 \left[\|x - y_n\|^2 + \|x - y_m\|^2 \right] - \|(x - y_n) + (x - y_m)\|^2 .$$

Noting that

$$(x - y_n) + (x - y_m) = 2 \left[x - \frac{y_n + y_m}{2} \right] \quad \text{and} \quad \frac{y_n + y_m}{2} \in C \text{ (by convexity of } C\text{)},$$

we have

$$\|x - y_n\|^2 \le p_\star^2 + \epsilon^2, \quad \|x - y_m\|^2 \le p_\star^2 + \epsilon^2, \text{ and } \|(x - y_n) + (x - y_m)\|^2 = 4\left\|x - \frac{y_n + y_m}{2}\right\|^2 \ge 4p_\star^2.$$

In particular, we have

$$\|y_n - y_m\|^2 \le 2\left[p_\star^2 + \epsilon^2 + p_\star^2 + \epsilon^2\right] - 4p_\star^2 = 4\epsilon^2.$$

As $\epsilon > 0$ was arbitrary, this completes the proof of the first statement of the theorem.

To see the second result, assume that z is a point satisfying inequality (A.1.2), that is, such that

$$\langle z - x, y - z\rangle \ge 0 \text{ for all } y \in C.$$

Then we have

$$\|z - x\|^2 = \langle z - x, z - x\rangle = \underbrace{\langle z - x, z - y\rangle}_{\le 0} + \langle z - x, y - x\rangle \le \|z - x\| \, \|y - x\|$$

by the Cauchy-Schwarz inequality. Dividing both sides by $\|z - x\|$ yields $\|z - x\| \le \|y - x\|$ for any $y \in C$, giving the result. Conversely, let $t \in [0, 1]$. Then for any $y \in C$,

$$\begin{aligned}\|\pi_C(x) - x\|^2 &\le \|(1-t)\pi_C(x) + ty - x\|^2 = \|\pi_C(x) - x + t(y - \pi_C(x))\|^2 \\ &= \|\pi_C(x) - x\|^2 + 2t\langle \pi_C(x) - x, y - \pi_C(x)\rangle + t^2\|y - \pi_C(x)\|^2.\end{aligned}$$

Subtracting the projection value $\|\pi_C(x) - x\|^2$ from both sides and dividing by $t > 0$, we have

$$0 \le 2\langle \pi_C(x) - x, y - \pi_C(x)\rangle + t\|y - \pi_C(x)\|^2.$$

Taking $t \to 0$ gives inequality (A.1.2). □

As an immediate consequence of Theorem A.10, we obtain several separation properties of convex sets, as well as a theorem stating that a closed convex set (not equal to the entire space in which it lies) can be represented as the intersection of all the half-spaces containing it.

Corollary A.11. *Let C be closed convex and $x \notin C$. Then there is a vector v strictly separating x from C, that is,*

$$\langle v, x\rangle > \sup_{y \in C} \langle v, y\rangle.$$

Moreover, we can take $v = x - \pi_C(x)$.

Proof By Theorem A.10, we know that taking $v = x - \pi_C(x)$ we have

$$0 \le \langle y - \pi_C(x), \pi_C(x) - x\rangle = \langle y - \pi_C(x), -v\rangle = \langle y - x + v, -v\rangle = -\langle y, v\rangle + \langle x, v\rangle - \|v\|^2.$$

That is, we have $\langle v, y\rangle \le \langle v, x\rangle - \|v\|^2$ for all $y \in C$ and $v \neq 0$. □

In addition, we can show the existence of supporting hyperplanes, that is, hyperplanes "separating" the boundary of a convex set from itself.

Theorem A.12. *Let C be a convex set and $x \in \operatorname{bd}(C)$, where $\operatorname{bd}(C) = \operatorname{cl}(C) \setminus \operatorname{int} C$. Then there exists a non-zero vector v such that $\langle v, x\rangle \ge \sup_{y \in C} \langle v, y\rangle$.*

Proof Let $D = \operatorname{cl}(C)$ be the closure of C and let $x_n \notin D$ be a sequence of points such that $x_n \to x$. Let us define the sequence of separating vectors $s_n = x_n - \pi_D(x_n)$ and the normalized version $v_n = s_n / \|s_n\|$. Notably, we have $\langle v_n, x_n\rangle > \sup_{y \in C} \langle v_n, y\rangle$ for all n. Now, the sequence $\{v_n\} \subset \{v : \|v\| = 1\}$ belongs to a compact set.[1] Passing to a subsequence if necessary, let us assume w.l.o.g. that $v_n \to v$ with $\|v\| = 1$. Then by a standard limiting argument for the $x_n \to x$, we have

$$\langle v, x\rangle \ge \langle v, y\rangle \quad \text{for all } y \in C,$$

which was our desired result. □

TODO: Picture of supporting hyperplanes and representations

Theorem A.12 gives us an important result. In particular, let D be an arbitrary set, and let $C = \operatorname{cl}\operatorname{Conv}(D)$ be the closure of the convex hull of D, which is the smallest closed convex set containing D. Then we can write C as the intersection of all the closed half-spaces containing D; this is, in some sense, the most useful "convexification" of D. Recall that a closed half-space H is defined with respect to a vector v and real $a \in \mathbb{R}$ as

$$H := \{x : \langle v, x\rangle \le r\}.$$

Before stating the theorem, we remark that by Observation A.6, the intersection of all the closed convex sets containing a set D is equal to the closure of the convex hull of D.

Theorem A.13. *Let D be an arbitrary set. If $C = \operatorname{cl}\operatorname{Conv}(D)$, then*

$$C = \bigcap_{H \supset D} H, \tag{A.1.3}$$

where H denotes a closed half-space containing D. Moreover, for any closed convex set C,

$$C = \bigcap_{x \in \operatorname{bd}(C)} H_x, \tag{A.1.4}$$

where H_x denotes the intersection of halfspaces supporting C at x.

Proof We begin with the proof of the second result (A.1.4). Indeed, by Theorem A.12, we know that at each point x on the boundary of C, there exists a non-zero supporting hyperplane v, so that the half-space

$$H_{x,v} := \{y : \langle v, y\rangle \le \langle v, x\rangle\} \supset C$$

is closed, convex, and contains C. We clearly have the containment $C \subset \cap_{x \in \operatorname{bd}(C)} H_x$. Now let $x_0 \notin C$; we show that $x_0 \notin \cap_{x \in \operatorname{bd}(C)} H_x$. As $x_0 \notin C$, the projection $\pi_C(x_0)$ of x_0 onto C satisfies $\langle x_0 - \pi_C(x_0), x_0\rangle > \sup_{y \in C} \langle x_0 - \pi_C(x_0), y\rangle$ by Corollary A.11. Moreover, letting $v = x_0 - \pi_C(x_0)$, the hyperplane

$$h_{x_0,v} := \{y : \langle y, v\rangle = \langle \pi_C(x_0), v\rangle\}$$

[1] In infinite dimensions, this may not be the case. But we can apply the Banach-Alaoglu theorem, which states that, as v_n are linear operators, the sequence is weak-* compact, so that there is a vector v with $\|v\| \le 1$ and a subequence $m(n) \subset \mathbb{N}$ such that $\langle v_{m(n)}, x\rangle \to \langle v, x\rangle$ for all x.

is clearly supporting to C at the point $\pi_C(x_0)$. The half-space $\{y : \langle y, v\rangle \le \langle \pi_C(x_0), v\rangle\}$ thus contains C and does not contain x_0, implying that $x_0 \notin \cap_{x\in \mathrm{bd}(C)} H_x$.

Now we show the first result (A.1.3). Let C be the closed convex hull of D and $T = \cap_{H\supset D} H$. By a trivial extension of the representation (A.1.4), we have that $C = \cap_{H\supset C} H$, where H denotes any halfspace containing C. As $C \supset D$, we have that $H \supset C$ implies $H \supset D$, so that

$$T = \bigcap_{H\supset D} H \subset \bigcap_{H\supset C} H = C.$$

On the other hand, as $C = \mathrm{cl}\,\mathrm{Conv}(D)$, Observation A.7 implies that any closed set containing D contains C. As a closed halfspace is convex and closed, we have that $H \supset D$ implies $H \supset C$, and thus $T = C$ as desired. □

A.2 Convex functions

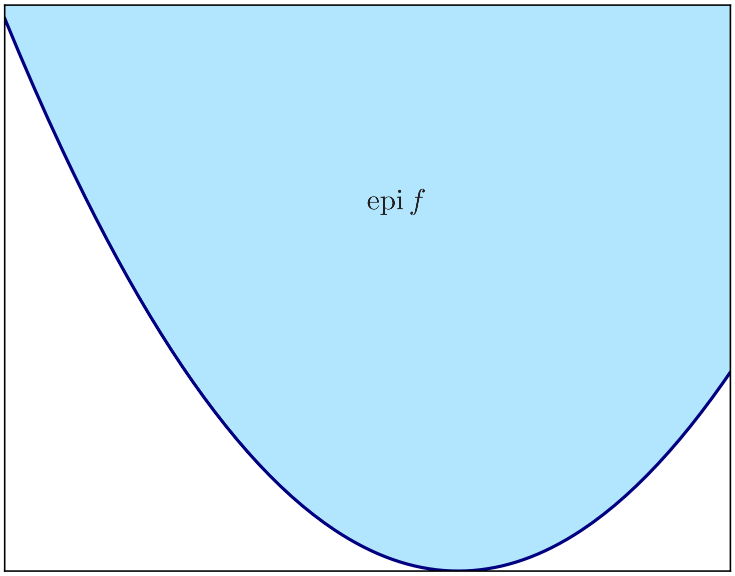

Figure A.1: The epigraph of a convex function.

We now build off of the definitions of convex sets to define convex functions. As we will see, convex functions have several nice properties that follow from the geometric (separation) properties of convex sets. First, we have

Definition A.6. *A function f is* convex *if for all $\lambda \in [0, 1]$ and $x, y \in \mathrm{dom}\, f$,*

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y). \tag{A.2.1}$$

We define the domain $\mathrm{dom}\, f$ of a convex function to be those points x such that $f(x) < +\infty$. Note that Definition A.6 implies that the domain of f must be convex.

An equivalent definition of convexity follows by considering a natural convex set attached to the function f, known as its epigraph.

Definition A.7. *The epigraph* epi f *of a function is the set*

$$\operatorname{epi} f := \{(x, t) : t \in \mathbb{R}, f(x) \le t\}.$$

That is, the epigraph of a function f is the set of points on or above the graph of the function itself, as depicted in Figure A.1. It is immediate from the definition of the epigraph that f is convex if and only if epi f is convex. Thus, we see that any convex set $C \subset \mathbb{R}^{d+1}$ that is unbounded "above," meaning that $C = C + \{0\} \times \mathbb{R}_+$, defines a convex function, and conversely, any convex function defines such a set C. This duality in the relationship between a convex function and its epigraph is central to many of the properties we exploit.

A.2.1 Equivalent definitions of convex functions

We begin our discussion of convex functions by enumerating a few standard properties that also characterize convexity. The simplest of these relate to properties of the derivatives and second derivatives of functions.

We begin with a first-order characterization. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is differentiable, and that for all $x, y \in \mathbb{R}$, we have

$$f(y) \ge f(x) + f'(x)(y - x). \tag{A.2.2}$$

We claim that inequality (A.2.2) implies that f is convex. Indeed, let $\lambda \in [0, 1]$ and $z = \lambda x + (1-\lambda)y$, so that $y - z = \lambda(y - x)$ and $x - z = (1 - \lambda)(x - y)$. Then

$$f(y) \ge f(z) + \lambda f'(z)(y - x) \ \text{ and } \ f(x) \ge f(z) + (1 - \lambda)f'(z)(x - y),$$

and multiplying the former by $(1 - \lambda)$ and the latter by λ and adding the two inequalities yields

$$\lambda f(x)+(1-\lambda)f(y) \ge \lambda f(z)+(1-\lambda)f(z)+\lambda(1-\lambda)f'(z)(y-x)+\lambda(1-\lambda)f'(z)(x-y) = f(\lambda x+(1-\lambda)y),$$

as desired. In Theorem A.14 to come, we see that the converse to inequality (A.2.2) holds as well, that is, differentiable convex functions satisfy inequality (A.2.2).

We may also give the standard second order characterization: if $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable and $f''(x) \ge 0$ for all x, then f is convex. To see this, note that

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2} f''(tx + (1 - t)y)(x - y)^2$$

for some $t \in [0, 1]$ by Taylor's theorem, so that $f(y) \ge f(x) + f'(x)(y - x)$ for all x, y because $f''(tx + (1 - t)y) \ge 0$. As a consquence, we obtain inequality (A.2.2), which implies that f is convex.

As convexity is a property that depends only on properties of functions on lines—one dimensional projections—we can straightforwardly extend the preceding results to functions $f : \mathbb{R}^d \to \mathbb{R}$. Indeed, noting that if $h(t) = f(x + ty)$ then $h'(0) = \langle \nabla f(x), y\rangle$ and $h''(0) = y^\top \nabla^2 f(x) y$, we have that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$f(y) \ge f(x) + \nabla f(x)^\top (y - x) \ \text{ for all } x, y,$$

while a twice differentabile function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$\nabla^2 f(x) \succeq 0 \ \text{ for all } x.$$

A.2.2 Continuity properties of convex functions

We now consider a few continuity properties of convex functions and a few basic relationships of the function f to its epigraph. First, we give a definition of the *subgradient* of a convex function.

Definition A.8. *A vector g is a* subgradient *of f at a point x_0 if for all x,*

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle . \tag{A.2.3}$$

See Figure A.2 for an illustration of the affine minorizing function given by the subgradient of a convex function at a particular point.

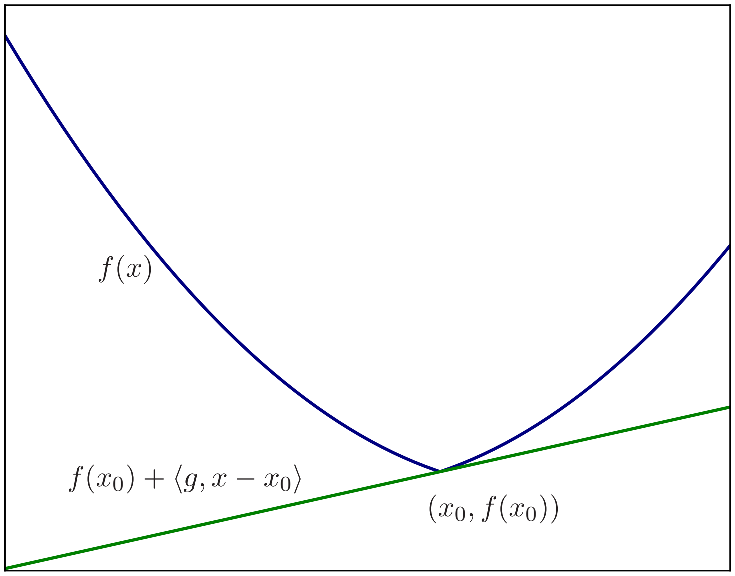

Figure A.2. The tangent (affine) function to the function f generated by a subgradient g at the point x_0.

Interestingly, convex functions have subgradients (at least, nearly everywhere). This is perhaps intuitively obvious by viewing a function in conjunction with its epigraph $\operatorname{epi} f$ and noting that $\operatorname{epi} f$ has supporting hyperplanes, but here we state a result that will have further use.

Theorem A.14. *Let f be convex. Then there is an affine function minorizing f. More precisely, for any $x_0 \in \operatorname{relint} \operatorname{dom} f$, there exists a vector g such that*

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle .$$

Proof If $\operatorname{relint} \operatorname{dom} f = \emptyset$, then it is clear that f is either identically $+\infty$ or its domain is a single point $\{x_0\}$, in which case the constant function $f(x_0)$ minorizes f. Now, we assume that $\operatorname{int} \operatorname{dom} f \neq \emptyset$, as we can simply always change basis to work in the affine hull of $\operatorname{dom} f$.

We use Theorem A.12 on the existence of supporting hyperplanes to construct a subgradient. Indeed, we note that $(x_0, f(x_0)) \in \operatorname{bd} \operatorname{epi} f$, as for any open set O we have that $(x_0, f(x_0)) + O$ contains points both inside and outside of $\operatorname{epi} f$. Thus, Theorem A.12 guarantees the existence of a vector v and $a \in \mathbb{R}$, not both simultaneously zero, such that

$$\langle v, x_0 \rangle + af(x_0) \leq \langle v, x \rangle + at \text{ for all } (x, t) \in \operatorname{epi} f. \tag{A.2.4}$$

Inequality (A.2.4) implies that $a \ge 0$, as for any x we may take $t \to +\infty$ while satisfying $(x, t) \in \operatorname{epi} f$. Now we argue that $a > 0$ strictly. To see this, note that for suitably small $\delta > 0$, we have $x = x_0 - \delta v \in \operatorname{dom} f$. Then we find by inequality (A.2.4) that

$$\langle v, x_0 \rangle + a f(x_0) \le \langle v, x_0 \rangle - \delta \left\|v\right\|^2 + a f(x_0 - \delta v), \quad \text{or} \quad a \left[f(x_0) - f(x_0 - \delta v)\right] \le -\delta \left\|v\right\|^2.$$

So if $v = 0$, then Theorem A.12 already guarantees $a \neq 0$, while if $v \neq 0$, then $\left\|v\right\|^2 > 0$ and we must have $a \neq 0$ and $f(x_0) \neq f(x_0 - \delta v)$. As we showed already that $a \ge 0$, we must have $a > 0$. Then by setting $t = f(x_0)$ and dividing both sides of inequality (A.2.4) by a, we obtain

$$\frac{1}{a} \langle v, x_0 - x \rangle + f(x_0) \le f(x) \quad \text{for all } x \in \operatorname{dom} f.$$

Setting $g = -v/a$ gives the result of the theorem, as we have $f(x) = +\infty$ for $x \notin \operatorname{dom} f$. □

Convex functions generally have quite nice behavior. Indeed, they enjoy some quite remarkable continuity properties just by virtue of the defining convexity inequality (A.2.1). In particular, the following theorem shows that convex functions are continuous on the relative interiors of their domains. Even more, convex functions are Lipschitz continuous on any compact subsets contained in the (relative) interior of their domains. (See Figure A.3 for an illustration of this fact.)

Theorem A.15. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and $C \subset \operatorname{relint} \operatorname{dom} f$ be compact. Then there exists an $L = L(C) \ge 0$ such that*

$$|f(x) - f(x')| \le L \left\|x - x'\right\|.$$

As an immediate consequence of Theorem A.15, we note that if $f : \mathbb{R}^d \to \mathbb{R}$ is convex and defined everywhere on $\mathbb{R}^d$, then it is continuous. Moreover, we also have that $f : \mathbb{R}^d \to \mathbb{R}$ is continuous everywhere on the (relative) interior of its domain: let any $x_0 \in \operatorname{relint} \operatorname{dom} f$. Then for small enough $\epsilon > 0$, the set $\operatorname{cl}(\{x_0 + \epsilon B\} \cap \operatorname{dom} f)$, where $B = \{x : \left\|x\right\|_2 \le 1\}$, is a closed and bounded—and hence compact—set contained in the (relative) interior of $\operatorname{dom} f$. Thus f is Lipschitz on this set, which is a neighborhood of x_0. In addition, if $f : \mathbb{R} \to \mathbb{R}$, then f is continuous everywhere except (possibly) at the endpoints of its domain.

Proof of Theorem A.15 To prove the theorem, we require a technical lemma.

Lemma A.16. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and suppose that there are x_0, $\delta > 0$, m, and M such that*

$$m \le f(x) \le M \quad \text{for } x \in B(x_0, 2\delta) := \{x : \|x - x_0\| < 2\delta\}.$$

Then f is Lipschitz on $B(x_0, \delta)$, and moreover,

$$|f(y) - f(y')| \le \frac{M - m}{\delta} \left\|y - y'\right\| \quad \text{for } y, y' \in B(x_0, \delta).$$

Proof Let $y, y' \in B(x_0, \delta)$, and define $y'' = y' + \delta(y' - y)/\left\|y' - y\right\| \in B(x_0, 2\delta)$. Then we can write y' as a convex combination of y and y'', specifically,

$$y' = \frac{\|y' - y\|}{\delta + \|y' - y\|} y'' + \frac{\delta}{\delta + \|y' - y\|} y.$$

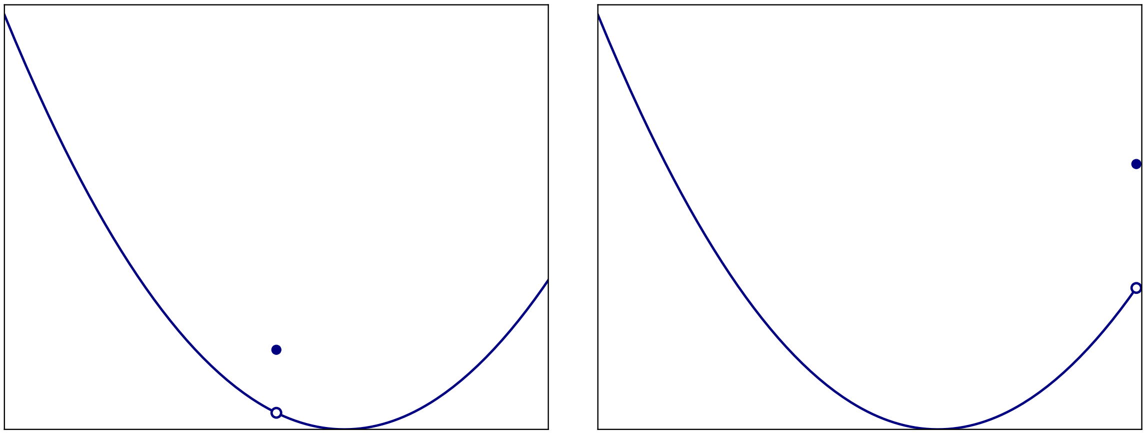

Figure A.3. Left: discontinuities in $\operatorname{int}\operatorname{dom} f$ are impossible while maintaining convexity (Theorem A.15). Right: At the edge of $\operatorname{dom} f$, there may be points of discontinuity.

Thus we obtain by convexity

$$\begin{aligned} f(y') - f(y) \le \frac{\|y' - y\|}{\delta + \|y' - y\|} f(y'') + \frac{\delta}{\delta + \|y' - y\|} f(y) - f(y) &= \frac{\|y - y'\|}{\delta + \|y - y'\|} [f(y'') - f(y)] \\ &\le \frac{M - m}{\delta + \|y - y'\|} \left\|y - y'\right\|. \end{aligned}$$

Here we have used the bounds on f assumed in the lemma. Swapping the assignments of y and y' gives the same lower bound, thus giving the desired Lipschitz continuity. □

With Lemma A.16 in place, we proceed to the proof proper. We assume without loss of generality that $\operatorname{dom} f$ has an interior; otherwise we prove the theorem restricting ourselves to the affine hull of $\operatorname{dom} f$. The proof follows a standard compactification argument. Suppose that for each $x \in C$, we could construct an open ball $B_x = B(x, \delta_x)$ with $\delta_x > 0$ such that

$$|f(y) - f(y')| \le L_x \left\|y - y'\right\| \quad \text{for } y, y' \in B_x. \tag{A.2.5}$$

As the B_x cover the compact set C, we can extract a finite number of them, call them $B_{x_1}, \ldots, B_{x_k}$, covering C, and then within each (overlapping) ball f is $\max_k L_{x_k}$ Lipschitz. As a consequence, we find that

$$|f(y) - f(y')| \le \max_k L_{x_k} \left\|y - y'\right\|$$

for any $y, y' \in C$.

We thus must derive inequality (A.2.5), for which we use the boundedness Lemma A.16. We must demonstrate that f is bounded in a neighborhood of each $x \in C$. To that end, fix $x \in \operatorname{int}\operatorname{dom} f$, and let the points $x_0, \ldots, x_d$ be affinely independent and such that

$$\Delta := \operatorname{Conv}\{x_0, \ldots, x_d\} \subset \operatorname{dom} f$$

and $x \in \operatorname{int} \Delta$; let $\delta > 0$ be such that $B(x, 2\delta) \subset \Delta$. Then by Carathéodory's theorem (Theorem A.3) we may write any point $y \in B(x, 2\delta)$ as $y = \sum_{i=0}^d \lambda_i x_i$ for $\sum_i \lambda_i = 1$ and $\lambda_i \ge 0$, and thus

$$f(y) \le \sum_{i=0}^{d} \lambda_i f(x_i) \le \max_{i \in \{0,\ldots,d\}} f(x_i) =: M.$$

Moreover, Theorem A.14 implies that there is some affine h function minorizing f; let $h(x) = a + \langle v, x \rangle$ denote this function. Then

$$m := \inf_{x \in C} f(x) \ge \inf_{x \in C} h(x) = a + \inf_{x \in C} \langle v, x \rangle > -\infty$$

exists and is finite, so that in the ball $B(x, 2\delta)$ constructed above, we have $f(y) \in [m, M]$ as required by Lemma A.16. This guarantees the existence of a ball B_x required by inequality (A.2.5). □

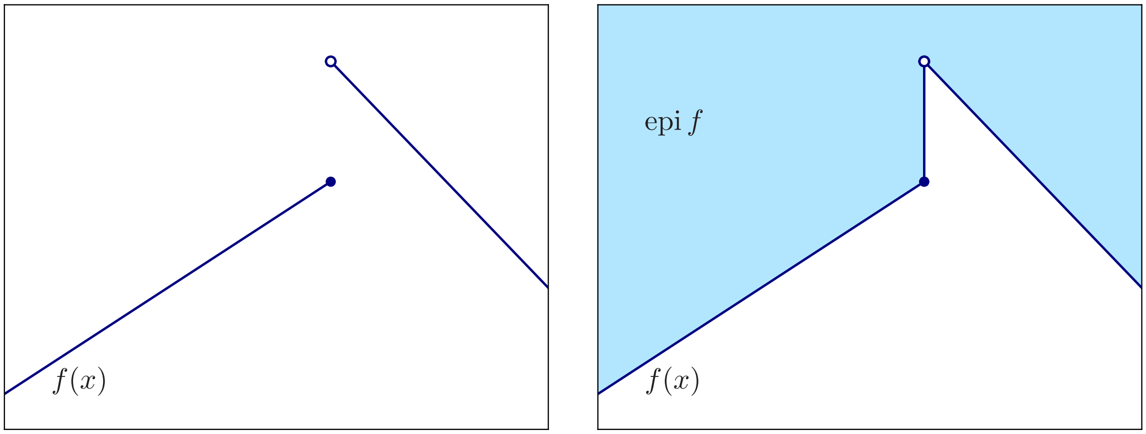

Figure A.4. A closed—equivalently, lower semi-continuous—function. On the right is shown the *closed* epigraph of the function.

Our final discussion of continuity properties of convex functions revolves around the most common and analytically convenient type of convex function, the so-called *closed-convex* functions.

Definition A.9. *A function f is* closed *if its epigraph,* epi f*, is a closed set.*

Equivalently, a function is closed if it is lower semi-continuous, meaning that

$$\liminf_{x \to x_0} f(x) \ge f(x_0) \tag{A.2.6}$$

for all x_0 and any sequence of points tending toward x_0. See Figure A.4 for an example such function and its associated epigraph.

Interestingly, in the one-dimensional case, closed convexity implies continuity. Indeed, we have the following observation (compare Figures A.4 and A.3 previously):

Observation A.17. *Let $f : \mathbb{R} \to \mathbb{R}$ be a closed convex function. Then f is continuous on its domain.*

Proof By Theorem A.15, we need only consider the endpoints of the domain of f (the result is obvious by Theorem A.15 if $\operatorname{dom} f = \mathbb{R}$); let $x_0 \in \operatorname{bd} \operatorname{dom} f$. Let $y \in \operatorname{dom} f$ be an otherwise arbitrary point, and define $x = \lambda y + (1 - \lambda)x_0$. Then taking $\lambda \to 0$, we have

$$f(x) \le \lambda f(y) + (1 - \lambda) f(x_0) \to f(x_0),$$

so that $\limsup_{x \to x_0} f(x) \le f(x_0)$. By the closedness assumption (A.2.6), we have $\liminf_{x \to x_0} f(x) \ge f(x_0)$, and continuity follows. □

A.2.3 Operations preserving convexity

We now turn to a description of a few simple operations on functions that preserve convexity. First, we extend the intersection properties of convex sets to operations on convex functions. (See Figure A.5 for an illustration of the proposition.)

Proposition A.18. *Let $\{f_\alpha\}_{\alpha \in \mathcal{A}}$ be an arbitrary collection of convex functions indexed by $\mathcal{A}$. Then*

$$f(x) := \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$$

is convex. Moreover, if for each $\alpha \in \mathcal{A}$, the function f_α is closed convex, f is closed convex.

Proof The proof is immediate once we consider the epigraph $\operatorname{epi} f$. We have that

$$\operatorname{epi} f = \bigcap_{\alpha \in \mathcal{A}} \operatorname{epi} f_\alpha,$$

which is convex whenever $\operatorname{epi} f_\alpha$ is convex for all α and closed whenever $\operatorname{epi} f_\alpha$ is closed for all α (recall Observation A.5). □

Another immediate result is that composition of a convex function with an affine transformation preserves convexity:

Proposition A.19. *Let $A \in \mathbb{R}^{d \times n}$ and $b \in \mathbb{R}^d$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then the function $g(y) = f(Ay + b)$ is convex.*

Lastly, we consider the functional analogue of the perspective transform. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, the *perspective transform* of f is defined as

$$\operatorname{pers}(f)(x, t) := \begin{cases} t f\left(\frac{x}{t}\right) & \text{if } t > 0 \text{ and } \frac{x}{t} \in \operatorname{dom} f \\ +\infty & \text{otherwise.} \end{cases} \tag{A.2.7}$$

In analogue with the perspective transform of a convex set, the perspective transform of a function is (jointly) convex.

Proposition A.20. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then $\operatorname{pers}(f) : \mathbb{R}^{d+1} \to \mathbb{R}$ is convex.*

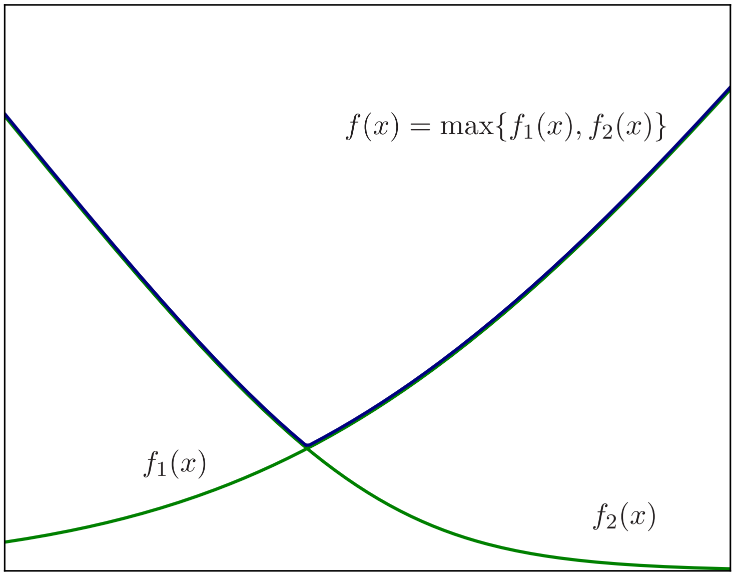

Figure A.5. The maximum of two convex functions is convex, as its epigraph is the intersection of the two epigraphs.

Proof The result follows if we can show that $\operatorname{epi}\operatorname{pers}(f)$ is a convex set. With that in mind, note that

$$\mathbb{R}^d \times \mathbb{R}_{++} \times \mathbb{R} \ni (x, t, r) \in \operatorname{epi}\operatorname{pers}(f) \text{ if and only if } f\left(\frac{x}{t}\right) \le \frac{r}{t}.$$

Rewriting this, we have

$$\begin{aligned}
\operatorname{epi}\operatorname{pers}(f) &= \left\{(x, t, r) \in \mathbb{R}^d \times \mathbb{R}_{++} \times \mathbb{R} : f\left(\frac{x}{t}\right) \le \frac{r}{t}\right\} \\
&= \left\{t(x', 1, r') : x' \in \mathbb{R}^d, t \in \mathbb{R}_{++}, r' \in \mathbb{R}, f(x') \le r'\right\} \\
&= \{t(x, 1, r) : t > 0, (x, r) \in \operatorname{epi} f\} = \mathbb{R}_{++} \times \{(x, 1, r) : (x, r) \in \operatorname{epi} f\}.
\end{aligned}$$

This is a convex cone. □

A.3 Conjugacy and duality properties

a. Closed convex function as a supremum of affine functions minorizing it

b. Fenchel Conjugate functions f^*

c. Fenchel biconjugate

A.4 Optimality conditions

Further reading

There are a variety of references on the topic, beginning with the foundational book by Rockafellar [7], which initiated the study of convex functions and optimization in earnest. Since then, a variety of authors have written (perhaps more easily approachable) books on convex functions, optimization, and their related calculus. Hiriart-Urruty and Lemaréchal [5] have written two volumes explaining in great detail finite-dimensional convex analysis, and provide a treatment of some first-order algorithms for solving convex problems. Borwein and Lewis [3] and Luenberger [6] give general treatments that include infinite-dimensional convex analysis, and Bertsekas [2] gives a variety of theoretical results on duality and optimization theory.

There are, of course, books that combine theoretical treatment with questions of convex modeling and procedures for solving convex optimization problems (problems for which the objective and constraint sets are all convex). Boyd and Vandenberghe [4] gives a very readable treatment for those who wish to use convex optimization techniques and modeling, as well as the basic results in convex analytic background and duality theory. Ben-Tal and Nemirovski [1], as well as Nemirovski's various lecture notes, give a theory of the tractability of computing solutions to convex optimization problems as well as methods for solving them.

Bibliography

[1] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization*. SIAM, 2001.

[2] D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[3] J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, 2006.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

[6] D. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

[7] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.