Lecture Notes in Mathematical Statistics

Jonathan Niles-Weed

September 12, 2022
Version 2.0.1

Contents

1 Introduction & Concentration inequlaities **5**
- 1.1 What is this class about? . . . 5
- 1.2 Basic inequalities . . . 7
- 1.3 Bounded random variables . . . 9
- 1.4 Sums of random variables . . . 10
- 1.5 Exercises . . . 12

2 Maximal inequalities and uniform convergence **15**
- 2.1 Maximal inequalities and the union bound . . . 15
- 2.2 Uniform convergence and empirical risk minimization . . . 17
  - 2.2.1 Empirical risk minimization . . . 17
- 2.3 Bracketing and Glivenko–Cantelli . . . 19
  - 2.3.1 The plug-in principle . . . 22
- 2.4 Exercises . . . 22

3 Asymptotics **25**
- 3.1 Modes of convergence . . . 25
- 3.2 Calculus of random variables . . . 26
- 3.3 Berry–Esseen Theorem . . . 28
- 3.4 Exercises . . . 29

4 Statistical modeling **31**
- 4.1 Models and statistical tasks . . . 31
- 4.2 Regression, Classification, and Clustering . . . 32
  - 4.2.1 Regression . . . 32
  - 4.2.2 Classification . . . 33
  - 4.2.3 Clustering . . . 34
- 4.3 Statistics and statistical tasks . . . 36
- 4.4 Exercises . . . 37

5 Point estimation **41**
- 5.1 What is a good estimator? . . . 41
  - 5.1.1 Unbiasedness . . . 41
  - 5.1.2 Asymptotic normality and efficiency . . . 42
  - 5.1.3 Decision theory framework . . . 42
  - 5.1.4 How to generalize Example 5.2? . . . 43
- 5.2 Method of Moments . . . 43
- 5.3 Maximum likelihood estimation (MLE) . . . 45
- 5.4 M-estimation . . . 48
  - 5.4.1 Consistency and Asymptotic Normality of M-estimators . . . 49
- 5.5 Exercises . . . 52

6 Testing **55**
6.1 Null and alternative hypotheses . . . 55
6.2 Test statistics and p-values . . . 56
6.2.1 p-values . . . 57
6.3 Likelihood ratio tests . . . 59
6.3.1 Neyman-Pearson Lemma . . . 60
6.4 Exercises . . . 61

7 Regularization **63**
7.1 The bias-variance tradeoff revisited . . . 63
7.2 Regularization as a variance-reduction strategy . . . 64
7.3 Smoothness in the Gaussian sequence model . . . 68
7.4 Exercises . . . 70

8 Monte-Carlo methods **73**
8.1 Monte Carlo p-values . . . 73
8.2 Permutation tests . . . 74
8.3 Rejection sampling . . . 75
8.4 Exercises . . . 76

9 Model selection and cross validation **77**
9.1 The basic problem . . . 77
9.2 Cross validation and AIC . . . 79
9.3 Model selection consistency and BIC . . . 81
9.4 Structural risk minimization and oracle inequalities . . . 82
9.5 Exercises . . . 85

10 Non-parametric estimation **87**
10.1 Kernel density estimation . . . 87
10.2 Kernel regression estimators . . . 89
10.3 Local linear regression . . . 91
10.4 Exercises . . . 93

11 High-dimensional linear regression **95**
11.1 Sparsity in the Gaussian sequence model . . . 95
11.2 ℓ_0 and ℓ_1 norms . . . 97
11.3 The Lasso . . . 99
11.4 Slow vs. fast rate . . . 100

Chapter 1

Introduction & Concentration inequlaities

1.1 What is this class about?

The goal of this class is to give mathematical tools for analyzing statistical procedures.

This raises the question: what do we mean by statistics? And what are valid statistical procedures? No matter the setting, the core question of statistics is the following: **given observations about the state of the world, make inferences about the processes that generated them**. The reason there is so much disagreement about the nature of statistics is that this is not yet a mathematical problem; it is a philosophical problem. We can turn it into a mathematical problem, but only by making axioms and formal assumptions. While this will give us the ability to ask mathematically precise questions, these assumptions and axioms are always up for debate.

The main assumption made in this class, and throughout much of statistics, is that the observations we make about the world are *random variables*, operating in accordance with the axioms of probability. We justify this for two reasons: first, there are some processes which, as far as we can tell, actually behave according to the axioms of probability as we understand them. These are exceedingly rare, and mostly involve quantum effects. The more common reason is that random variables seem to be accurate models for chaotic or unpredictable systems.

Using the language of random variables, we can restate the main question of statistics.

Statistics: given observations in the form of random variables, what can we infer about the probability distribution that generated them?

These questions are in some sense "opposite" to those typically asked in probability theory:

Probability: given a probability distribution, how do random variables with this distribution behave?

Of course, these two perspectives are complimentary. Statistics is a very broad field, with a lot of subareas—sometimes it's more an art than a science! The goal of mathematical statistics is to analyze this problem using mathematical tools. This course will try to cover parts of the classical theory, while giving a taste of some more modern challenges as well, such as the role of non-asymptotic results and high-dimensional phenomena.

Asymptotic vs. Non-asymptotic

Let $X_1, \dots$ be independent and identically distributed ("i.i.d.") random variables with finite expectation. The law of large numbers says

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\to} \mathbb{E}X_1 \,.$$

To make formal what we mean by the limiting notation, let us focus on the following version, called the *weak law of large numbers*.

Theorem 1.1. *Let $X_1, \dots$ be a sequence of i.i.d. random variables with finite expectation. Then*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X_1\right| > \varepsilon\right\} \underset{n\to\infty}{\to} 0 \quad \forall \varepsilon > 0\,.$$

This is an *asymptotic* statement. These are mathematically very clean but can be hard to put into use in practice, since without more information we "never see" the $n \to \infty$ limit. An example of a *non-asymptotic* statement is the following reformulation of the weak law of large numbers.

Theorem 1.2. *Let $X_1, \dots$ be a sequence of i.i.d. random variables with finite expectation and variance. Then*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X_1\right| \geq \varepsilon\right\} \leq \frac{\mathrm{Var}(X_1)}{n\varepsilon^2} \quad \forall \varepsilon > 0\,.$$

This tells us what happens for *any* value of n: if we know that the variance of X_1 is at most 1, then we can be 99% confident that our average is within .1 if we have $n = 10^4$ samples. We will show how to derive and strengthen such inequalities soon.

Classical statistics was mostly concerned with asymptotic results; non-asymptotic results are more common in parts of the modern machine learning literature. However, *both* types of results have important applications, and asymptotic results can still be extremely useful.

Low vs. High-dimensional

Consider the following setup. We observe (Y_i, X_i) for $i = 1, \dots, n$, where $Y_i \in \mathbb{R}$ is an outcome which we want to predict on the basis of a "covariate" $X_i \in \mathbb{R}^p$. We assume that these random variables are i.i.d. and satisfy

$$Y_i = \langle X_i, \beta^* \rangle + \varepsilon_i\,, \tag{1.1}$$

where ε_i models random noise, and where $\beta^* \in \mathbb{R}^p$ is some unknown vector which controls the effect that the covariates X_i have on the outcomes. This is an extremely famous model in statistics, known as a *linear model* because $\mathbb{E}[Y_i|X_i]$ is a *linear* function of X_i. We will see many examples of this type later.

It's reasonable to ask, given samples of the above form, can we estimate β^*? A classic idea known as *ordinary least squares* goes back to Gauss and Legendre: compute the *estimator*

$$\hat{\beta} := \underset{\beta\in\mathbb{R}^p}{\mathrm{argmin}} \sum_{i=1}^{n} (Y_i - \langle X_i, \beta\rangle)^2\,.$$

Minimizing the sum of squares encourages $\langle X_i, \beta\rangle$ to be close to Y_i, and, we hope, leads to a good estimate of β^*.

Under mild conditions, which we will explore later in the course, we can show that this estimator satisfies

$$\mathbb{E}\|\beta^* - \hat{\beta}\|^2 \approx \frac{p}{n}\,.$$

How to interpret this quantity? There is a big difference between what we shall call the *low-dimensional regime* and the *high-dimensional regime*.

Low-dimensional regime: $p \ll n$ This is the classical regime, and the error here is small. Think about predicting a subject's height by the height of their parents. Then $p = 2$, and it is easy to collect a large number of subjects, so n can be large.

High-dimensional regime: $n \approx p$ **or** $n \ll p$ Here, the error will be large; typically so large that $\hat{\beta}$ is not meaningful. For example, if one wants to predict the effectiveness of a drug based on genetic markers, then we might take $p \approx 10^6$ (the typical number of variations in a human genome). Of course, the number of subjects n will be much, much smaller.

1.2 Basic inequalities

If statistical inference is to be possible, we have to understand how random variables *typically* behave, and we have to be able to show that *atypical* behavior is rare. These are the job of *concentration* inequalities, which all are designed to say that a particular r.v. is near some value with high probability. The most famous example: averages of i.i.d. random variabes.

A very important tool is Markov's inequality, which allows us to transfer bounds on the expectation of a random variable to high probability bounds. This is useful because expectations are often possible to compute even for complicated random variables.

Proposition 1.3 (Markov's inequality)**.** *If X is a nonnegative scalar random variable, then*

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X}{t} \quad \forall t > 0\,.$$

Proof. Define the indicator $\mathbb{1}_{\{X \geq t\}}$:

$$\mathbb{1}_{\{X \geq t\}} = \begin{cases} 1 & \text{if } X \geq t, \\ 0 & \text{otherwise.} \end{cases}$$

Then $t \cdot \mathbb{1}_{\{X \geq t\}} \leq X$ (why?), so

$$\mathbb{E}X \geq \mathbb{E}[t \cdot \mathbb{1}_{\{X \geq t\}}] = t\mathbb{P}\{X \geq t\}\,.$$

□

Markov's inequality is quite weak. It shows its true power when it is applied to a *function* of a random variable.

Proposition 1.4. *Let X be a scalar random variable, and let $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a non-negative function which is non-decreasing on $[t_0, +\infty)$. Then*

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}\phi(X)}{\phi(t)} \qquad \forall t \geq t_0\,.$$

Proof. Since ϕ is increasing on $[t_0, +\infty)$ and $\phi(X)$ is nonnegative, we have

$$\mathbb{P}\{X \geq t\} \leq \mathbb{P}\{\phi(X) \geq \phi(t)\} \leq \frac{\mathbb{E}\phi(X)}{\phi(t)}\,.$$

□

Corollary 1.5. *Under the same assumptions as Proposition 1.4,*

$$\mathbb{P}\{X \leq -t\} \leq \frac{\mathbb{E}\phi(-X)}{\phi(t)} \quad \forall t \geq t_0\,.$$

Proof. Apply Proposition 1.4 to $-X$. □

To get the best possible tail bound, we should aim to choose a function ϕ such that $\phi(t)$ is as *large as possible* (this makes the denominator as large as possible), while $\mathbb{E}\phi(X)$ is *easy to bound* (and not too big).

Proposition 1.4, though simple, is at the heart of many extremely useful inequalities. The two most useful choices of ϕ in practice are $\phi(x) = x^2$ and $\phi(x) = e^{\lambda x}$ for $\lambda > 0$. The first choice leads to a simple deviation bound, often called Chebyshev's inequality. This is our first "concentration inequality," which says that a random variable has a small probability of being far from its mean.

Corollary 1.6 (Chebyshev's inequality)**.**

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \frac{\mathrm{Var}(X)}{t^2} \quad \forall t > 0\,.$$

Proof. Apply Proposition 1.4 with $\phi(x) = x^2$ to the random variable $|X - \mathbb{E}X|$. □

The choice $\phi(x) = e^{\lambda x}$ gives a more powerful concentration bound, particularly for random variables satisfying the following very useful definition.

Definition 1.7. A random variable X is σ^2-*subgaussian* if

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq e^{\frac{\lambda^2}{2}\cdot\sigma^2} \quad \forall \lambda \in \mathbb{R},\,. \tag{1.2}$$

When $\sigma^2 = 1$ or, more generally, when σ^2 can be understood from context, we simply say that X is *subgaussian*.

A motivation for the name "subgaussian" comes from the fact that, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then it satisfies

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} = e^{\frac{\lambda^2}{2}\cdot\sigma^2} \quad \forall \lambda \in \mathbb{R}\,.$$

In other words, for *gaussian* random variables, (1.2) holds with equality. In this sense, subgaussian random variables are "smaller than" or "below" gaussians.

Combined with Proposition 1.4, we obtain the following strong concentration bound for subgaussian random variables. This is often called the Chernoff bound, and the proof is called the Chernoff method.

Corollary 1.8 (Chernoff bound)**.** *If X is σ^2-subgaussian, then*

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq 2e^{-t^2/2\sigma^2} \quad \forall t \geq 0\,.$$

Proof. Let us first obtain a bound on $\mathbb{P}\{X - \mathbb{E}X \geq t\}$. We use Proposition 1.4 with $\phi(x) = e^{\lambda x}$ to $X - \mathbb{E}X$, for some $\lambda \geq 0$ to be specified. Applying Definition 1.7 yields

$$\begin{aligned}\mathbb{P}\{X - \mathbb{E}X \geq t\} &\leq \mathbb{E}e^{\lambda(X-\mathbb{E}X)} \cdot e^{-\lambda t} \\ &\leq e^{\frac{\lambda^2}{2}\cdot\sigma^2 - \lambda t} \quad \forall t \in \mathbb{R},\,.\end{aligned}$$

We are free to choose whichever $\lambda \in \mathbb{R}$ makes this inequality the strongest. In other words, we should seek to minimize the exponent $\frac{\lambda^2}{2}\cdot\sigma^2 - \lambda t$ over all $\lambda \in \mathbb{R}$. This is a quadratic function of λ, achieving its global minimum at $\lambda = t/\sigma^2$. Making this choice of λ, we obtain

$$\begin{aligned}\mathbb{P}\{X - \mathbb{E}X \geq t\} &\leq e^{\frac{\lambda^2}{2}\cdot\sigma^2 - \lambda t} \\ &= e^{\frac{t^2}{2\sigma^2} - \frac{t^2}{\sigma^2}} && (\lambda = t/\sigma^2) \\ &= e^{-t^2/2\sigma^2}\,.\end{aligned}$$

We have shown $\mathbb{P}\{X - \mathbb{E}X \geq t\} \leq e^{-t^2/2\sigma^2}$ for all $t \in \mathbb{R}$. We now note that if X is σ^2-subgaussian, then so is $-X$. (Check this!) It follows that we also obtain, by the exact same argument, the bound

$$\mathbb{P}\{X - \mathbb{E}X \leq -t\} \leq e^{-t^2/2\sigma^2}\,.$$

For $t \geq 0$, the event $|X - \mathbb{E}X| \geq t$ holds if and only if $X - \mathbb{E}X \geq t$ or $X - \mathbb{E}X \leq -t$. We therefore obtain for $t \geq 0$ that

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \mathbb{P}\{X - \mathbb{E}X \geq t\} + \mathbb{P}\{X - \mathbb{E}X \leq -t\} \leq 2e^{-t^2/2\sigma^2}\,.$$

□

1.3 Bounded random variables

Often, it is possible to compute $\mathrm{Var}(X_i)$ for each i without much trouble. But it is also good to know how to obtain good bounds on $\mathrm{Var}(X_i)$ when it cannot be calculated exactly. The following bound is extremely useful.

Proposition 1.9. *If $X \in [a, b]$, then*

$$\mathrm{Var}(X) \leq (b-a)^2/4\,.$$

We will prove the weaker bound

$$\mathrm{Var}(X) \leq (b-a)^2$$

and leave the stronger bound for homework.

Proof of Proposition 1.9, weak form. Since $X \geq a$, taking expectations yields $\mathbb{E}X \geq a$. Similarly, $\mathbb{E}X \leq b$. Therefore $\mathbb{E}X \in [a, b]$.

Since X and $\mathbb{E}X$ both lie in $[a, b]$, it follows that $|X - \mathbb{E}X| \leq (b-a)$. Therefore

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 \leq (a-b)^2\,,$$

as claimed. □

It turns out much more is true: a bounded random variable is also subgaussian. This is the content of the famous lemma due to Hoeffding.

Proposition 1.10 (Hoeffding's Lemma)**.** *If $X \in [a, b]$, then X is $(b-a)^2/4$-subgaussian.*

Once again, we will prove a looser bound, and leave a tighter bound to homework.

Proof of Proposition 1.10, weak form. We employ the inequality $e^t \leq t + e^{t^2}$, valid for all $t \in \mathbb{R}$, which follows from the inequality

$$1 - t + e^t \leq e^t + e^{-t} \leq 1 + e^{t^2}\,.$$

Therefore

$$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} \leq \mathbb{E}\lambda(X - \mathbb{E}X) + \mathbb{E}e^{\lambda^2 (X - \mathbb{E}X)^2}\,.$$

The first term vanishes, and as in the proof of Proposition 1.9, we have that $(X - \mathbb{E}X)^2 \leq (a-b)^2$ with probability 1. Therefore

$$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} \leq e^{\lambda^2 (a-b)^2}\,.$$

□

An important note is that Proposition 1.9 and Proposition 1.10 are often far from tight. For instance, if $X_1, \ldots, X_n$ are independent random variables taking values in $[-1, 1]$, then their average $\bar{X}$ clearly also takes values in $[-1, 1]$. Applying Proposition 1.10 to $\bar{X}$ directly, we could conclude that $\bar{X}$ is 1-subgaussian. But this is way off: as we will see soon, the fluctuations of $\bar{X}$ are of size $1/\sqrt{n}$! Indeed, we will show (Proposition 1.15) that $\bar{X}$ is actually $\frac{1}{n}$-subgaussian. The lesson is that simple bounds like Proposition 1.9 and Proposition 1.10 are often good enough when applied to "simple" random variables, but they can be very loose when the random variable is "complicated," particularly if the complicated random variable is a function of a large number of independent random variables. Such random variables typically have much better behavior than the individual random variables that comprise them.

1.4 Sums of random variables

As previewed last time, the most important concentration fact in all of statistics is the following statement:

> An average of n i.i.d. random variables is within $\mathcal{O}(1/\sqrt{n})$ of its expectation with high probability.

Informally, we can approximate the expectation of a random variable by taking independent samples. If this were not true, it would be nearly impossible to learn from data.

Today, we will try to understand this phenomenon in more detail via concentration bounds. Recall Proposition 1.4. We obtained the Chebyhsev and Chernoff bounds by choosing $\phi(x) = x^2$ and $\phi(x) = e^{\lambda x}$, respectively. These choices are useful because they interact very well with sums of independent random variables. We demonstrate this when $\phi(x) = x^2$.

Lemma 1.11. *If $X_1, \dots, X_n$ are independent, then*

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}(X_i)\,.$$

Proof. First, we can assume without loss of generality that the X_i are centered. (This is an important and common trick.) Indeed, if we replace X_i by the centered random variables $X_i - \mathbb{E}X_i$ for $i = 1, \dots, n$, then both sides are unchanged. (Check this!)

We can then compute:

$$\begin{aligned}
\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \mathbb{E}\frac{1}{n^2}\sum_{i,j=1}^{n} X_i X_j \\
&= \frac{1}{n^2}\sum_{i,j=1}^{n} \mathbb{E}X_i X_j \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}X_i^2 && (X_i \text{ independent, centered}) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}(X_i)\,.
\end{aligned}$$

□

Note that this proof only uses the weaker fact that X_i are uncorrelated (i.e., $\mathbb{E}X_iX_j = \mathbb{E}X_i\mathbb{E}X_j$ for $i \neq j$).

This lemma immediately yields a version Chebyshev's inequality for sums.

Proposition 1.12 (Chebyshev's inequality for sums)**.** *Let $X_1, \dots, X_n$ be independent random variables, with $\operatorname{Var}(X_i) = \sigma^2$ for $i = 1, \dots n$. Let $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ be their average. Then*

$$\mathbb{P}\left\{\left|\bar{X} - \mathbb{E}\bar{X}\right| \geq s\frac{\sigma}{\sqrt{n}}\right\} \leq s^{-2} \quad \forall s \geq 0\,.$$

Before giving the proof, let's interpret this inequality. The term $\sigma/\sqrt{n}$ gives us a sense of the *scale of the fluctuations* of $\bar{X}$. When $s \leq 1$, the right side is at least 1, so Chebyshev's inequality gives no information. Therefore, the average can vary as much as it wants within $\sigma n^{-1/2}$ of the mean. However, once $s > 1$, the probability on the right side begins to decrease—so the average cannot have a too large probability of straying far outside a $\sigma n^{-1/2}$ radius.

Proof. Corollary 1.6 implies

$$\mathbb{P}\left\{\left|\bar{X} - \mathbb{E}\bar{X}\right| \geq t\right\} \leq \frac{\operatorname{Var}\bar{X}}{t^2} \quad \forall t \geq 0\,.$$

Plugging in the bound of Lemma 1.11 and choosing $t = s\sigma/\sqrt{n}$ yields the claim. □

Note that the claim still holds with the same proof if the X_i have unequal variances, as long as $\operatorname{Var}(X_i) \leq \sigma^2$.

The story with subgaussian random variables and the Chernoff bound is similar. As in Lemma 1.11, sums of independent subaussian random variables are themselves subgaussian, with an easy-to-control subgaussianity parameter.

Lemma 1.13. *If $X_1, \ldots, X_n$ are independent and X_i is σ_i^2-subgaussian for $i = 1, \ldots, n$, then $\frac{1}{n}\sum_{i=1}^n X_i$ is*

$$\left(\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2\right)\text{-subgaussian}\,.$$

Proof. We assume as above that each X_i are centered. For any $\lambda \in \mathbb{R}$, we compute

$$\begin{aligned}
\mathbb{E}e^{\lambda\left(\frac{1}{n}\sum_{i=1}^n X_i\right)} &= \mathbb{E}\prod_{i=1}^n e^{\frac{\lambda}{n}X_i} \\
&= \prod_{i=1}^n \mathbb{E}e^{\frac{\lambda}{n}X_i} && \text{(independence)} \\
&\leq \prod_{i=1}^n e^{\frac{(\lambda/n)^2}{2}\cdot\sigma_i^2} && \text{(Definition 1.7)} \\
&= e^{\frac{\lambda^2}{2}\cdot\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2}\,.
\end{aligned}$$

The claim now follows from Definition 1.7. □

We can apply Lemma 1.13 to get a good concentration inequality for sums of subgaussian random variables.

Proposition 1.14 (Chernoff bound for sums)**.** *Let $X_1, \ldots, X_n$ be independent random variables, each σ^2-subgaussian. Let $\bar{X}$ be their average. Then*

$$\mathbb{P}\left\{|\bar{X} - \mathbb{E}\bar{X}| \geq s\frac{\sigma}{\sqrt{n}}\right\} \leq 2e^{-s^2/2} \quad \forall s \geq 0\,.$$

As with Proposition 1.12, Proposition 1.14 implies that $\sigma/\sqrt{n}$ is the natural scale of the fluctuations of $\bar{X}$. As before, when $s \approx 1$, this inequality is uninformative, and, as s grows, the the probability on the right side begins to decrease. Here, though, the stronger assumption (that the X_i are subgaussian) gives a much stronger result, since the probability on the right side decreases to 0 exponentially fast.

Proof. By Lemma 1.13, the average $\bar{X}$ is $\left(\frac{1}{n}\sigma^2\right)$-subgaussian. Corollary 1.8 then implies that

$$\mathbb{P}\left\{|\bar{X} - \mathbb{E}\bar{X}| \geq t\right\} \leq 2e^{-nt^2/2\sigma^2} \quad \forall t \geq 0\,.$$

Choosing $t = s\frac{\sigma}{\sqrt{n}}$ yields the claim. □

Combining this with Proposition 1.10, we obtain Hoeffding's inequality.

Proposition 1.15 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables, with $X_i \in [a, b]$ for $i = 1, \ldots, n$. Let $\bar{X}$ be their average. Then*

$$\mathbb{P}\left\{|\bar{X} - \mathbb{E}\bar{X}| \geq s\frac{(b-a)}{2\sqrt{n}}\right\} \leq 2e^{-s^2/2} \quad \forall s \geq 0\,.$$

This shows the average of bounded random variables is very sharply concentrated.

1.5 Exercises

1. Show that several of our inequalities cannot be improved.

   (a) Markov's inequality is tight: for all $t > 0$, there exists a nonnegative random variable variable X such that $\mathbb{P}\{X \geq t\} = \mathbb{E}X/t$. (Hint: make every inequality in the proof of Markov's inequality an equality.)

   (b) Chebyshev's inequality is tight: for all $t > 0$, there exists a random variable X such that $\mathbb{P}\{|X - \mathbb{E}X| \geq t\} = \mathrm{Var}(X)/t^2$. (Hint: adapt your example from part (a), above.)

2. The goal of this exercise is to give full proofs of Proposition 1.9 and Proposition 1.10. Assume throughout that $X \in [a, b]$.

   (a) Prove that $\mathrm{Var}(X) \leq \mathbb{E}(X - c)^2$ for all random variables X and $c \in \mathbb{R}$. (Hint: prove and use that $\mathbb{E}(X - c)^2 = \mathrm{Var}(X) + (\mathbb{E}X - c)^2$.)

   (b) Prove Proposition 1.9 by choosing $c = (a + b)/2$.

   (c) Assume that X is a centered random variable with pdf p. Let

   $$q_\lambda(x) = \frac{e^{\lambda x}}{\mathbb{E}e^{\lambda X}} p(x)\,.$$

   Show that q_λ is a probability density.

   (d) Define the function $K(\lambda) \coloneqq \log \mathbb{E}e^{\lambda X}$. Show that

   $$K'(\lambda) = \int x q_\lambda(x)\,\mathrm{d}x \tag{1.3}$$

   $$K''(\lambda) = \int x^2 q_\lambda(x)\,\mathrm{d}x - \left(\int x q_\lambda(x)\,\mathrm{d}x\right)^2\,. \tag{1.4}$$

   (e) Show that $K(0) = K'(0) = 0$. By using Proposition 1.9, show that $K''(\lambda) \leq (b - a)^2/4$. (Hint: interpret (1.4) as the variance of a bounded random variable.)

   (f) Prove Proposition 1.10 by showing $K(\lambda) \leq \frac{\lambda^2}{2} \cdot \frac{(b-a)^2}{4}$. (Hint: integrate.)

3. Above, we claimed that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then

   $$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} = e^{\frac{\lambda^2}{2} \cdot \sigma^2} \quad \forall \lambda \in \mathbb{R}\,.$$

   Prove this fact.

4. Comparing Corollary 1.6 to Corollary 1.8 suggests that assuming that a random variable is subgaussian is stronger than assuming that its variance is small. Indeed, this is true.

   (a) Prove that if X is σ^2-subgaussian, then $\mathrm{Var}(X) \leq \sigma^2$. (Hint: consider $\lim_{\lambda \to 0} \frac{\mathbb{E}e^{\lambda(X - \mathbb{E}X)} - 1}{\lambda^2}$.)

   (b) Prove that if $X \sim \mathrm{Exp}(1)$, then $\mathrm{Var}(X) = 1$ but X is not σ^2-subgaussian for any finite σ^2.

5. Though Chebyshev's inequality (Proposition 1.12) is typically weaker than Hoeffding's inequality (Proposition 1.15), there are situations where it can be stronger. This is particularly true when the variance is small. There is an improvement of Hoeffding's inequality called Bernstein's inequality which gives better bounds in the case, which is outside the scope of this course. This exercise will examine the failure of Hoeffding's inequality for rare events.

   Let $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$ be independent, and let $\bar{X}$ be their average.

   (a) Compute $\mathrm{Var}(\bar{X})$ as a function of n and p.

(b) Using Proposition 1.12 and Proposition 1.15, give two different upper bounds on the probability that $|\bar{X} - p| \geq 1/(2\sqrt{n})$.

(c) Show that if $p = \lambda/n$ for some $\lambda > 0$, then Proposition 1.12 gives a nontrivial concentration inequality but Proposition 1.15 does not.

Chapter 2

Maximal inequalities and uniform convergence

2.1 Maximal inequalities and the union bound

One central aspect of high-dimensional statistics is the necessity of controlling the fluctuations of many random variables at once.

For instance, if $X_1, \dots, X_n$ are independent scalar random variables with values in $[-1, 1]$, then Proposition 1.15 guarantees that their average $\bar{X}$ satisfies

$$\bar{X} \in (\mathbb{E}\bar{X} - 3.3n^{-1/2}, \mathbb{E}\bar{X} + 3.3n^{-1/2}) \quad \text{with probability at least .99.} \tag{2.1}$$

(Check this!) This shows that even after a modest number of samples, we can be confident that $\bar{X}$ is near $\mathbb{E}\bar{X}$.

However, suppose instead that $X_1, \dots, X_n$ are independent *vector* random variables in $\mathbb{R}^p$, where again each coordinate of each X_i takes values in $[-1, 1]$. (Equivalently, we may say that X_i takes values in $[-1, 1]^p$.) For example, these may represent p different measurements taken of n different subjects. Now, we can still form the empirical average $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \in \mathbb{R}^p$. Does a claim like (2.1) still hold?

Let us specify what we mean. If we ignore all but the jth coordinate of $\bar{X}$, then (2.1) guarantees that

$$\bar{X}_j \in (\mathbb{E}\bar{X}_j - 3.3n^{-1/2}, \mathbb{E}\bar{X}_j + 3.3n^{-1/2}) \quad \text{with probability at least .99.}$$

That is, there is only a small probability that $\bar{X}_j$ deviates by a large amount from its expectation. The problem is that, as p becomes large, there are more and more opportunities for this "bad event" to occur. Indeed, you will show on the homework that if the entries of X_i are independent nondegenerate random variables and n is sufficiently large, we will always have

$$\mathbb{P}\left\{\bar{X}_j \in (\mathbb{E}\bar{X}_j - 3.3n^{-1/2}, \mathbb{E}\bar{X}_j + 3.3n^{-1/2}) \quad \forall j = 1, \dots, p\right\} \underset{p\to\infty}{\to} 0\,. \tag{2.2}$$

The message is striking: even though it is very likely that that any *particular* coordinate is near its mean, it is very unlikely that *all* of them are near their means!

Clearly we want to be able to control many random variables at once. We seek what is called a *maximal inequality*. Recall that if X is σ^2-subgaussian (Definition 1.7), then

$$\mathbb{P}\left\{|X - \mathbb{E}X| \geq t\right\} \leq 2e^{-t^2/2\sigma^2}\,.$$

Equivalently, by taking $t = \sqrt{2\sigma^2 \log(2/\delta)}$, we can say that if X is σ^2-subgaussian, then

$$|X - \mathbb{E}X| < \sqrt{2\sigma^2 \log(2/\delta)} \quad \text{with probability at least } 1 - \delta. \tag{2.3}$$

We now show prove a maximal inequality for subgaussian random variables—that is, an inequality like (2.3) that holds for many random variables at once. The key idea is deceptively simple, and goes by the name "union bound": for any events $A_1, \dots, A_p$,

$$\mathbb{P}\left\{\bigcup_{i=1}^{p} A_i\right\} \le \sum_{i=1}^{p} \mathbb{P}\{A_i\} .$$

We can therefore bound the probability that a set of events occurs by the sum of the probabilities of each event.

Proposition 2.1. *Suppose $X_1, \dots, X_p$ are σ^2-subgaussian random variables,* not necessarily independent. *Then*

$$\max_{i=1,\dots,p} |X_i - \mathbb{E}X_i| \le \sqrt{2\sigma^2 \log(2p/\delta)} \quad \textit{with probability at least } 1-\delta.$$

In other words, up to the appearance of a p in the logarithm, the *maximum* (i.e. worst) deviation scales almost like the deviation of a single random variable.

Proof. We will show that

$$\mathbb{P}\left\{\max_{i=1,\dots,p} |X_i - \mathbb{E}X_i| \ge \sqrt{t^2 + 2\sigma^2 \log p}\right\} \le 2e^{-t^2/2\sigma^2} , \tag{2.4}$$

from which the claim follows by choosing $t = \sqrt{2\sigma^2 \log(2/\delta)}$. (Note: since $a + b \ge \sqrt{a^2 + b^2}$ for $a, b \ge 0$, a slightly weaker but easier-to-use version of (2.4) is

$$\mathbb{P}\left\{\max_{i=1,\dots,p} |X_i - \mathbb{E}X_i| \ge t + \sqrt{2\sigma^2 \log p}\right\} \le 2e^{-t^2/2\sigma^2} .$$

We will often use the second one.)

Let us write $\mathcal{B}$ for the bad event that $\max_{i=1,\dots,p} |X_i - \mathbb{E}X_i| \ge \sqrt{t^2 + 2\sigma^2 \log p}$. The bad event $\mathcal{B}$ occurs if and only if $|X_i - \mathbb{E}X_i| \ge \sqrt{t^2 + 2\sigma^2 \log p}$ for some $i = 1, \dots, p$. In other words, if we write $\mathcal{B}_i$ for the event that $|X_i - \mathbb{E}X_i| \ge \sqrt{t^2 + 2\sigma^2 \log p}$, then $\mathcal{B} = \bigcup_{i=1}^{p} \mathcal{B}_i$.

We obtain

$$\begin{aligned}
\mathbb{P}\{\mathcal{B}\} &= \mathbb{P}\left\{\bigcup_{i=1}^{p} \mathcal{B}_i\right\} \\
&\le \sum_{i=1}^{p} \mathbb{P}\{\mathcal{B}_i\} && \text{(union bound)} \\
&= \sum_{i=1}^{p} \mathbb{P}\left\{|X_i - \mathbb{E}X_i| > \sqrt{t^2 + 2\sigma^2 \log p}\right\} \\
&\le \sum_{i=1}^{p} 2e^{-(t^2 + 2\sigma^2 \log p)/2\sigma^2} && \text{(subgaussianity of } X_i\text{)} \\
&= \sum_{i=1}^{p} 2p^{-1} e^{-t^2/2\sigma^2} \\
&= 2e^{-t^2/2\sigma^2} .
\end{aligned}$$

□

Returning to the setting at the beginning of the section, we recall from the proof of Proposition 1.15 that the average of n independent random variables taking values in $[-1,1]$ is n^{-1}-subgaussian. Therefore, for *any* p,

$$\mathbb{P}\left\{\bar{X}_j \in (\mathbb{E}\bar{X}_j - (3.3+\sqrt{2\log p})n^{-1/2}, \mathbb{E}\bar{X}_j + (3.3+\sqrt{2\log p})n^{-1/2}) \quad \forall j = 1,\dots,p\right\} \geq .99$$

Comparing this with (2.2), we see that we only need to expand our interval by a little bit to guarantee that it contains all p coordinates simultaneously. (And it really is just a little bit: we can handle $p = 10^6$ at a price of $\sqrt{2\log p} \approx 5.25$.)

2.2 Uniform convergence and empirical risk minimization

A very useful application of Proposition 2.1 is to bounding the convergence of functions of random variables to their expectations. To illustrate, given i.i.d. samples $X_1,\dots,X_n$, the law of large numbers ensures that for any bounded function f,

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \underset{n\to\infty}{\to} \mathbb{E}f(X_1)\,.$$

(For now, we don't emphasize what type of convergence we mean here.) However, it is often important to show that this convergence happens *simultaneously* for all functions in some class $\mathcal{F}$. The mathematical jargon for this is to ask that the convergence happens "uniformly" over the set $\mathcal{F}$.

In fact, if the set $\mathcal{F}$ is finite, we can obtain such a guarantee directly from Proposition 2.1

Proposition 2.2. *Let $\mathcal{F}$ be a finite collection of $[-1,1]$-valued functions and let $X_1,\dots,X_n$ be i.i.d. Then for any $\delta > 0$,*

$$\max_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1)\right| \leq \sqrt{\frac{2\log(2|\mathcal{F}|/\delta)}{n}} \quad \textit{with probability at least } 1-\delta.$$

Proof. For each f, the quantity $\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1)$ is n^{-1}-subgaussian. (Why?) The claim follows from Proposition 2.1. □

As long as $|\mathcal{F}|$ is not too large, we therefore obtain uniform convergence at the price of a logarithmic factor.

2.2.1 Empirical risk minimization

Proposition 2.2 is exceedingly useful for analyzing an idea known as empirical risk minimization. This is perhaps the most common procedure in modern statistics/machine learning.

Let's consider a problem called *binary classification*. We observe a *training set* $\{(y_1,X_1),\dots(y_n,X_n)\}$ of i.i.d. random variables, where $X_i \in \mathbb{R}^p$ are some covariates (or, in ML-speak, "features"), based on which we are trying to predict $y_i \in \{\pm 1\}$ which is called the *response* (i.e., yes or no). Our goal is to find a good classifier: a function $f : \mathbb{R}^d \to \{\pm 1\}$ which does a good job of predicting y on the basis of f.

Formally, we would like to find a function f which minimizes the *risk*

$$\mathcal{R}(f) \coloneqq \mathbb{P}\left\{f(X) \neq y\right\}\,.$$

We have no way of evaluating the actual risk, but for any f we can compute the *empirical risk*:

$$\hat{\mathcal{R}}(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{f(X_i)\neq y_i}$$

For any fixed f, it is easy to use Proposition 1.15 to show that $\hat{\mathcal{R}}(f) \approx \mathcal{R}(f)$.

Lemma 2.3. *For any fixed f,*

$$\mathbb{P}\left\{|\hat{\mathcal{R}}(f) - \mathcal{R}(f)| \geq \frac{s}{\sqrt{n}}\right\} \leq 2e^{-s^2/2}\,.$$

Proof. The indicators $\{\mathbb{1}_{f(X_i)\neq y_i}\}$ are independent, bounded random variables. Proposition 1.15 therefore implies

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{f(X_i)\neq y_i} - \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{f(X_i)\neq y_i}\right| \geq \frac{s}{\sqrt{n}}\right\} \leq 2e^{-s^2/2}\,.$$

We recognize $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{f(X_i)\neq y_i} = \hat{\mathcal{R}}(f)$. Similarly,

$$\begin{aligned}\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{f(X_i)\neq y_i} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\mathbb{1}_{f(X_i)\neq y_i}\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\left\{f(X_i)\neq y_i\right\}\\ &= \mathcal{R}(f)\,.\end{aligned}$$

This proves the claim. □

To summarize:

- We would like to minimize $\mathcal{R}$.
- We have no access to $\mathcal{R}$, but only to $\hat{\mathcal{R}}$.
- For a fixed classifier f, $\mathcal{R}(f)$ and $\hat{\mathcal{R}}(f)$ are close.

This suggests the following idea, called *empirical risk minimization*: to choose a good classifier from a set $\mathcal{F}$ of candidate classifiers, we should choose the one which minimizes $\hat{\mathcal{R}}$.

Formally, let

$$\hat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \hat{\mathcal{R}}(f)\,.$$

We want to argue that the *excess risk* $\mathcal{R}(\hat{f}) - \mathcal{R}(f^*)$ is small, where $f^* = \operatorname{argmin}_{f\in\mathcal{F}}\mathcal{R}(f)$ is the optimal choice.

We can analyze this error as follows:

$$\begin{aligned}\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) &= (\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})) + (\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f^*)) + (\hat{\mathcal{R}}(f^*) - \mathcal{R}(f^*))\\ &= \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3\,,\end{aligned}$$

where

$$\begin{aligned}\mathcal{E}_1 &\coloneqq \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})\\ \mathcal{E}_2 &\coloneqq \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f^*)\\ \mathcal{E}_3 &\coloneqq \hat{\mathcal{R}}(f^*) - \mathcal{R}(f^*)\,.\end{aligned}$$

We can easily bound $\mathcal{E}_2$ and $\mathcal{E}_3$. Since $\hat{f}$ minimizes $\hat{\mathcal{R}}$, we always have $\mathcal{E}_2 \leq 0$. By Lemma 2.3, $\mathcal{E}_3$ is small with high probability. However, we come to an extremely important point: **we cannot apply Lemma 2.3 to $\mathcal{E}_1$.** This is because $\hat{f}$ is not a fixed classifier—it depends on the data $\{(y_1, X_1), \ldots, (y_n, X_n)\}$—and you can check that the proof of Lemma 2.3 fails.

To obtain an upper bound on $\mathcal{E}_1$, we replace it by a much stronger quantity that we *can* control:

$$\mathcal{E}_1 = \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f}) \leq \max_{f\in\mathcal{F}} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)|\,.$$

The important point is that $\max_{f \in \mathcal{F}} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)|$ is something we can control using Proposition 2.2 or Proposition 2.5.

For example, if $|\mathcal{F}|$ is finite, then the above reasoning implies

$$\begin{aligned}\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) &\leq |\hat{\mathcal{R}}(f^*) - \mathcal{R}(f^*)| + \max_{f \in \mathcal{F}} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)| \\ &\leq 2 \max_{f \in \mathcal{F}} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)| \\ &\leq 2\sqrt{\frac{2\log(2|\mathcal{F}|/\delta)}{n}}\,.\end{aligned}$$

with probability at least $1 - \delta$. This says that, if $|\mathcal{F}|$ is finite, then empirical risk minimization works—it produces a classifier which is almost as good as the optimal classifier with high probability.

2.3 Bracketing and Glivenko–Cantelli

Proposition 2.2 provides a good bound if $\mathcal{F}$ is finite, but if $\mathcal{F}$ is infinite then it is totally uninformative. It turns out that this isn't a failure of our proof idea: there exist infinite classes for which our uniform bound fails, as you will show on the homework. This raises the question of whether anything like Proposition 2.2 can be obtained for sets $\mathcal{F}$ which contain an infinite number of functions. And of course, most interesting classes in statistics and machine learning are infinite!

Finding good uniform convergence results for infinite classes is a very important topic in statistics and probability. We will only scratch the surface of this topic in this class, but the example we present will represent a key idea in this area: it is possible to bound an infinite class if it can be approximated well by a finite class.

To approximate an infinite class, we will use the idea of a *bracket*.

Definition 2.4. Given a pair of functions (g, h), the *bracket* $[g, h]$ is the set of all functions satisfying

$$g(x) \leq f(x) \leq h(x) \quad \forall x\,.$$

The idea of the next result is that we can obtain a uniform convergence result over an infinite class as long as it can be covered by a small number of small brackets.

Proposition 2.5. *Let $X_1, \ldots, X_n$ be i.i.d. Suppose that $\mathcal{F}$ is a class of $[-1, 1]$-valued functions and suppose that there exist m brackets $[g_1, h_1], \ldots, [g_m, h_m]$ and an $\varepsilon > 0$ such that:*

- *The functions defining each bracket are ε close in expectation, i.e.,*

$$\mathbb{E}h_j(X_i) - \mathbb{E}g_j(X_i) \leq \varepsilon \quad \forall j = 1, \ldots, m$$

- *The brackets $[g_1, h_1], \ldots, [g_m, h_m]$ cover $\mathcal{F}$, i.e., for each $f \in \mathcal{F}$ there exists a $j \leq m$ such that $f \in [g_j, h_j]$.*

Then for any $\delta > 0$,

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i) \right| \leq \sqrt{\frac{2\log(4m/\delta)}{n}} + \varepsilon$$

with probability at least $1 - \delta$.

Note that typically m and ε will have an inverse relationship: the smaller ε is, the closer g_j and h_j are for each j, so the smaller the bracket $[g_j, h_j]$ is, and the larger m must be to cover all of $\mathcal{F}$. There is therefore a trade-off between the two terms in the bound.

Proof. Suppose that $f \in \mathcal{F}$ lies in the bracket $[g, h]$. Then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i) &\le \frac{1}{n}\sum_{i=1}^{n} h(X_i) - \mathbb{E}g(X_i) \\ &\le \frac{1}{n}\sum_{i=1}^{n} h(X_i) - \mathbb{E}h(X_i) + \varepsilon \\ &\le \max_j \left| \frac{1}{n}\sum_{i=1}^{n} h_j(X_i) - \mathbb{E}h_j(X_i) \right| + \varepsilon\,.\end{aligned}$$

Likewise, we have

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i) \ge -\max_j \left| \frac{1}{n}\sum_{i=1}^{n} g_j(X_i) - \mathbb{E}g_j(X_i) \right| - \varepsilon\,.$$

Combining these bounds yields

$$\left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i) \right| \le \max_{\bar{f} \in \{g_1,\ldots,g_m,h_1,\ldots,h_m\}} \left| \frac{1}{n}\sum_{i=1}^{n} \bar{f}(X_i) - \mathbb{E}\bar{f}(X_i) \right| + \varepsilon\,.$$

Since the right side does not depend on f, this inequality holds uniformly over $\mathcal{F}$. Therefore, since the $\bar{\mathcal{F}} \coloneqq \{g_1, \ldots, g_m, h_1, \ldots, h_m\}$ has $2m$ elements, we can apply Proposition 2.2 to obtain

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i) \right| \le \max_{\bar{f} \in \bar{\mathcal{F}}} \left| \frac{1}{n}\sum_{i=1}^{n} \bar{f}(X_i) - \mathbb{E}\bar{f}(X_i) \right| + \varepsilon \le \sqrt{\frac{2\log(4m/\delta)}{n}} + \varepsilon$$

with probability at least $1 - \delta$. □

We now present our first result deserving the name "Theorem." This result is extremely important in statistical practice. The idea of this theorem is that the so-called empirical distribution is close to the true distribution with high probability. We first recall the definition of the distribution function of a random variable.

Definition 2.6. The *(cumulative) distribution function* or *CDF* of a real-valued random variable X is

$$F(t) \coloneqq \mathbb{P}\{X \le t\}\,.$$

The CDF contains all the information about the distribution of X. (In particular, if X is a continuous random variable, then F' is its pdf.)

We now define what we mean by the empirical distribution: this is a *random* distribution induced by i.i.d. copies of a random variable.

Definition 2.7. Given i.i.d. random variables $X_1, \ldots, X_n$, the *empirical distribution* is the probability distribution that assigns probability $1/n$ to each of the n points $X_1, \ldots, X_n$.

In words, we generate the empirical distribution by sampling n i.i.d. copies and then constructing the uniform distribution on the n numbers we sampled.

The empirical distribution is a probability distribution, so we can compute its CDF:

$$\hat{F}_n(t) = \frac{|\{X_i : X_i \le t\}|}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le t\}}\,. \tag{2.5}$$

Note that $\hat{F}_n$ *is a random variable* taking values in the set of functions from $\mathbb{R}$ to $[0, 1]$. Make sure you understand this point!

The Glivenko-Cantelli says that the empirical distribution approaches the underlying distribution, in the sense that the empirical CDF $\hat{F}_n$ approaches F.

Theorem 2.8 (Glivenko-Cantelli)**.** *Let $X_1, \dots, X_n$ be i.i.d. copies of a random variable X having distribution function F. For any $\delta > 0$,*

$$\sup_{t \in \mathbb{R}} |F(t) - \hat{F}_n(t)| \leq 2\sqrt{\frac{2\log(4n/\delta)}{n}}$$

with probability at least $1 - \delta$.

The bound we present is not the best possible: one can in fact show that the $O(\sqrt{(\log n)/n})$ dependence can be improved to $O(\sqrt{1/n})$, but we won't pursue this here. We assume that $n \geq 3$, since if $n \leq 2$ the bound is vacuous.

Proof. We will define a suitable set of brackets and apply Proposition 2.5.

Step 1 : For any point $t \in \mathbb{R}$, write

$$F(t-) \coloneqq \lim_{s \to t^-} F(s) = \mathbb{P}\{X < t\} .$$

If F is continuous at t, then $F(t-) = F(t)$, but it is possible (e.g., if X is discrete), for $F(t-) < F(t)$. For any ε, we claim that we can find an ordered set $T = \{t_1, \dots, t_{m-1}\} \subset \mathbb{R}$ of points with $m \leq 1/\varepsilon + 1$ such that

$$\begin{aligned} F(t_1-) &\leq \varepsilon \\ F(t_{m-1}) &\geq 1 - \varepsilon \\ F(t_{j+1}-) - F(t_j) &\leq \varepsilon \quad \forall j = 0, \dots, m-1 . \end{aligned}$$

Indeed, taking $t_0 = -\infty$ and $t_m = +\infty$, we need to find $t_1 < \cdots < t_{m-1}$ such that

$$\mathbb{P}\{X \in (t_j, t_{j+1})\} \leq \varepsilon \quad \forall j = 0, \dots, m-1 ,$$

and it is easy to check that for any random variable and any $\varepsilon > 0$, one can partition the real line into at most $1/\varepsilon + 1$ intervals such that X lies in each open interval with probability at most ε.

Step 2 : Let us consider the class $\mathcal{F} \coloneqq \{\mathbb{1}_{\{x \leq t\}} : t \in \mathbb{R}\}$ and the brackets $[\mathbb{1}_{\{x \leq t_j\}}, \mathbb{1}_{\{x < t_{j+1}\}}]$ for $j = 0, \dots, m-1$. First, note that

$$\mathbb{E}\mathbb{1}_{\{X < t_{j+1}\}} - \mathbb{E}\mathbb{1}_{\{X \leq t_j\}} = \mathbb{P}\{X < t_{j+1}\} - \mathbb{P}\{X \leq t_j\} = F(t_{j+1}-) - F(t_j) \leq \varepsilon \quad \forall j = 0, \dots, m-1$$

by construction. Second, for any $t \in \mathbb{R}$, there exists a unique $j \in \{0, \dots, m-1\}$ such that $t \in [t_j, t_{j+1})$, and for this j we have

$$\mathbb{1}_{\{x \leq t_j\}} \leq \mathbb{1}_{\{x \leq t\}} \leq \mathbb{1}_{\{x < t_{j+1}\}} \quad \forall x \in \mathbb{R} .$$

Therefore, the m brackets $[\mathbb{1}_{\{x \leq t_j\}}, \mathbb{1}_{\{x < t_{j+1}\}}]$ for $j = 0, \dots, m-1$ satisfy the requirements of Proposition 2.5.

Step 3 : By Proposition 2.5, we obtain for any $\varepsilon > 0$,

$$\begin{aligned} \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| &= \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq t\}} - \mathbb{E}\mathbb{1}_{\{X \leq t\}} \right| \\ &= \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X) \right| \\ &\leq \sqrt{\frac{2\log(4m/\delta)}{n}} + \varepsilon \\ &\leq \sqrt{\frac{2\log(4(\varepsilon^{-1}+1)/\delta)}{n}} + \varepsilon . \end{aligned}$$

If we choose $\varepsilon = 1/\sqrt{n}$, we have

$$\sqrt{\frac{2\log(4(\varepsilon^{-1}+1)/\delta)}{n}} + \varepsilon = \sqrt{\frac{2\log(4(\sqrt{n}+1)/\delta)}{n}} + \frac{1}{\sqrt{n}} \le 2\sqrt{\frac{2\log(4n/\delta)}{n}}\,,$$

where we use the fact that if $n \ge 3$, then $\sqrt{n}+1 \le n$ and $2\log(4n/\delta) \ge 1$. This proves the claim. □

2.3.1 The plug-in principle

The empirical distribution provides a simple and often good tool for estimating properties of a distribution based on data. This is called the *plug-in principle*, and it works like this: given i.i.d. copies of a random variable X (with unknonwn distribution F), form the empirical distribution (with distribution function $\hat{F}_n$). Then, given any property of interest of the distribution F, we can estimate this property by the corresponding property of $\hat{F}_n$. To give a simple example, if we are interested in the mean of X, then the plug-in principle suggests to use the mean of the empirical distribution instead. Since the empirical distribution puts probability mass $1/n$ on each of the n samples $X_1, \dots, X_n$, this "plug-in estimator" is simply

$$\frac{1}{n}\sum_{i=1}^{n} X_n\,,$$

i.e., the sample average.

To understand this slightly more formally, recall that since the CDF of a random variable contains all the information about its distribution, we can always write properties of a random variable's distribution as functions of its distribution function. In other words, if we define Φ to be the set of all valid CDFs, then we can realize any fact about the distribution of X (e.g., its mean or variance) as a mapping $T : \Phi \to \mathbb{R}$. As a simple example, if F is the cdf of X, then

$$\mathbb{P}\{X > t\} = 1 - F(t)\,,$$

and we can think of this as a mapping

$$\begin{aligned} T : \Phi &\to \mathbb{R} \\ F &\mapsto 1 - F(t)\,. \end{aligned}$$

Theorem 2.8 says that $\hat{F}_n$ approaches F uniformly. Therefore, as long as the mapping T is continuous in an appropriate sense, we can argue that $T(\hat{F}_n) \to T(F)$, so that when n is large $T(\hat{F}_n)$ is a good estimate of $T(F)$. This idea forms the basis an important idea in applied statistics called the bootstrap, which we may talk about later in the course.

2.4 Exercises

1. This exercise will investigate (2.2).

   (a) Let X be any nondegenerate random variable on $[-1, 1]$. (Nondegenerate means that X takes more than one value.) Show that there exists a positive integer n_0 and a constant $c > 0$ depending on the distribution of X but not on n such that, if $X_1, \dots, X_n$ are independent copies of X, then

   $$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X\right| > 3.3n^{-1/2}\right\} > c \quad \forall n \ge n_0\,.$$

   (Hint: the central limit theorem guarantees that

   $$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}X)\right| > 3.3\right\} = \mathbb{P}\left\{|Z| > 3.3 \cdot \mathrm{Var}(X)^{-1/2}\right\}\,,$$

   where Z is a standard gaussian random variable.)

(b) Conclude that if $\bar{X}_j$ is an independent copy of $\bar{X} \coloneqq \frac{1}{n}\sum_{i=1}^n X_i$ for $j = 1, \ldots, p$ and $n \geq n_0$, then

$$\mathbb{P}\left\{|\bar{X}_j - \mathbb{E}X| \leq 3.3n^{-1/2} \quad \forall j = 1, \ldots, p\right\} < (1-c)^p \underset{p\to\infty}{\to} 0\,.$$

2. This exercise will show that uniform convergence and empirical risk minimization can easily fail for infinite classes. Let $X_1, \ldots X_n \sim \text{Unif}([0,1])$ be i.i.d. Let $\mathcal{F}$ be the set of all indicator functions, i.e., functions of the form

$$f(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases}$$

for some $S \subset [0,1]$.

(a) Show that, for any $f \in \mathcal{F}$ and any $\delta > 0$,

$$\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \int_0^1 f(x)\,\mathrm{d}x\right| < \sqrt{\frac{\log(2/\delta)}{2n}} \quad \text{with probability } 1-\delta\,.$$

In other words, when n is large, any particular $f \in \mathcal{F}$ is close to its expectation with high probability.

(b) Nevertheless, show that, no matter the value of $X_1, \ldots, X_n$, there exists an $f \in \mathcal{F}$ such that

$$\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \int_0^1 f(x)\,\mathrm{d}x\right| = 1\,.$$

In other words, no matter how large n is, there is always a function in $\mathcal{F}$ which is very far from its expectation.

(c) Let $\{(y_1, X), \ldots, (y_n, X_n)\}$ be i.i.d. copies of (y, X), where $X \sim \text{Unif}([0,1])$ and $y = 1$ always. Show that for any n, there exists a minimizer $\hat{f}$ of the empirical risk $\hat{R}$ such that

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) = 1\,.$$

In other words, no matter how large n is, empirical risk minimization fails.

3. Theorem 2.8 showed that the empirical CDF $\hat{F}_n$ converges to F, in the sense that $\max_{t\in\mathbb{R}} |F(t) - \hat{F}_n(t)| \leq \mathcal{O}(\sqrt{\log n/n})$ with high probability. In this exercise you will show convergence in a different sense.

(a) For a fixed $t \in \mathbb{R}$, show that $\mathbb{E}\hat{F}_n(t) = F(t)$.

(b) For a fixed $t \in \mathbb{R}$, show that $\text{Var}(\hat{F}_n(t)) = \frac{1}{n}F(t)(1-F(t))$.

(c) Using the above two claims, show that

$$\mathbb{E}\int_{-\infty}^{\infty} (F(t) - \hat{F}_n(t))^2\,\mathrm{d}t \leq \frac{1}{n}\int_0^{\infty} \mathbb{P}\left\{|X| \geq t\right\}\,\mathrm{d}t\,.$$

(You may assume that you can interchange expectation and integration.)

(d) Conclude that

$$\mathbb{E}\int_{-\infty}^{\infty} (F(t) - \hat{F}_n(t))^2\,\mathrm{d}t \leq \frac{\mathbb{E}|X|}{n}\,.$$

(Hint: Write $\int_0^\infty \mathbb{P}\left\{|X| \geq t\right\}\,\mathrm{d}t = \int_0^\infty \mathbb{E}\mathbb{1}_{|X|\geq t}\,\mathrm{d}t$ and interchange expectation and integration.)

Mathematically, bounds on $\max_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)|$ are known as L_∞ bounds and bounds on $\int_{-\infty}^{\infty} (F(t) - \hat{F}_n(t))^2 \, \mathrm{d}t$ are known as L_2 bounds. Neither bound is stronger than the other, but they give different information: the first says that the error at each point is small, and the second says that the total squared error over the whole real line is small.

4. In this exercise, we will use the following strengthened form of Theorem 2.8, which goes by the name "Dvoretzky–Kiefer–Wolfowitz inequality": if $\hat{F}_n$ is the empirical distribution corresponding to n i.i.d. samples from F, then

$$\mathbb{P}\left\{\sup_{t\in\mathbb{R}} |F(t) - \hat{F}_n(t)| \geq s\right\} \leq 2e^{-2ns^2} \quad \forall s \geq 0\,. \tag{2.6}$$

   (a) Show that (2.6) is indeed stronger than Theorem 2.8.

   (b) Let $X_1, \ldots, X_n$ be i.i.d. copies of a random variable X with unknown distribution. Suppose we wish to decide whether the distribution function of X is equal to some known CDF F. In light of Theorem 2.8, one reasonable approach is to compare the empirical distribution function $\hat{F}_n$ corresponding to $X_1, \ldots, X_n$ to F. Suppose we decide that we will declare that X does *not* have distribution F if $\sup_{t\in\mathbb{R}} |F(t) - \hat{F}_n(t)| \geq \sqrt{3/2n}$. Show using (2.6) that if X did actually come from F, then we are wrong with probability at most 10%.

   (c) Let $X_1, \ldots, X_n$ be i.i.d. copies of X and $Y_1, \ldots, Y_n$ be i.i.d. copies of Y, where the distributions of both X and Y are unknown. Suppose we wish to decide whether X and Y have the same distribution. Let $\hat{F}_n$ and $\hat{G}_n$ be the empirical CDF's corresponding to $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively, and suppose that we declare that the distributions of X and Y are different if $\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - \hat{G}_n(t)| \geq \sqrt{8/n}$. Show using (2.6) that if X and Y do actually have the same distribution, then we are wrong with probability at most 10%.

   (d) In parts (b) and (c) above, our test depended on being able to decide whether $\sup_{t\in\mathbb{R}} |F(t) - \hat{F}_n(t)|$ or $\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - \hat{G}_n(t)|$ is larger than some threshold. Assuming that the samples are sorted, show that this question can be answered in $\emptyset(n)$ time.

   The procedures described in this question are called Kolmogorov–Smirnov tests. We will learn much more about testing soon.

Chapter 3

Asymptotics

3.1 Modes of convergence

In the asymptotic setting, we usually consider an infinite sequence $X_1, \ldots, \in \mathcal{X}$ of independent, identically distributed random variables, and a sequence of random variables T_n obtained as functions of the first n data points: $T_n = T_n(X_1, \ldots, X_n)$. For instance, for each $n \in \mathbb{N}$, we might consider the average $T_n = \frac{1}{n}\sum_{i=1}^n X_i$. We will often want to know how this family of statistics behaves as n grows. To that end, we briefly recall two notions of convergence which you will have seen in your probability class.

Definition 3.1. A sequence $(T_n)_{n=1}^\infty$ of random variables *converges in probability* to a random variable T (written $T_n \xrightarrow{p} T$) if

$$\mathbb{P}\left\{|T_n - T| \geq \varepsilon\right\} \underset{n\to\infty}{\to} 0 \quad \forall \varepsilon > 0\,.$$

We note that in this definition T could be deterministic—for example, the weak law of large numbers says that if $X_1, \ldots, X_n$ are i.i.d., then $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{p} \mathbb{E}X_1$.

Definition 3.2. A sequence $(T_n)_{n=1}^\infty$ of random variables *converges in distribution* to T (written $T_n \xrightarrow{d} T$ if

$$\mathbb{P}\left\{T_n \leq t\right\} = F_{T_n}(t) \underset{n\to\infty}{\to} F_T(t) = \mathbb{P}\left\{T \leq t\right\}$$

for all $t \in \mathbb{R}$ at which F_T is continuous.

The definition of convergence in probability has a natural extension to random variables T_n taking values in a space other than $\mathbb{R}$, but the generalization of convergence in distribution is less clear. The following statement, which we will not prove, gives an alternative definition of convergence in distribution which generalizes more naturally to other settings.

Lemma 3.3 (Portmanteau Lemma)**.** *The following are equivalent:*

1. $F_{T_n}(t) \underset{n\to\infty}{\to} F_T(t)$ *for all* $t \in \mathbb{R}$ *at which* F_T *is continuous.*
2. $\mathbb{E}f(T_n) \underset{n\to\infty}{\to} \mathbb{E}f(T)$ *for all bounded, continuous* $f : \mathbb{R} \to \mathbb{R}$*.*

Thanks to Lemma 3.3, we can generalize the definition of convergence in distribution to random variables in more general spaces (such as $\mathbb{R}^d$, or to any metric space).

The most important example of convergence in distribution is the central limit theorem: if $X_1, \ldots, X_n$ are independent with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$ for $i = 1, \ldots, n$, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(0,1)\,.$$

An analogous statement holds for independent random variables in $\mathbb{R}^d$. It is possible to formulate similar statements the context of other spaces, but several mathematically subtle issues arise that we will not treat here.

Convergence in probability is stronger than convergence in distribution, though they are equivalent in some special cases. We record these facts without proof below.

Proposition 3.4. *1. If $T_n \xrightarrow{p} T$, then $T_n \xrightarrow{d} T$.*

2. If $T_n \xrightarrow{d} c$ for a constant c, then $T_n \xrightarrow{p} c$.

3. If $T_n \xrightarrow{d} T$ and $Y_n \xrightarrow{p} c$ for a constant c, then $(T_n, Y_n) \xrightarrow{d} (T, c)$.

A key difference between convergence in probability and convergence in distribution is that, for convergence in probability, the random variables $(T_n)_{n=1}^{\infty}$ and T must all be defined on the same probability space; for convergence in distribution, only the distributions of each random variable matter. In particular, it is enough to just specify a distribution (say $\mathcal{N}(0,1)$) to which the variables converge, without specifying a random variable.

3.2 Calculus of random variables

The key benefit of using asymptotic statements over non-asymptotic ones is that asymptotic statements are much easier to manipulate: they give rise to what might be called a "calculus of random variables," which can be used to derive sophisticated limit theorems for functions of random variables.

We will start with some simple examples.

Proposition 3.5 (Continuous mapping)**.** *For any continuous function g, if $T_n \xrightarrow{*} T$, then $g(T_n) \xrightarrow{*} g(T)$, where $\xrightarrow{*}$ indicates either $\xrightarrow{p}$ or $\xrightarrow{d}$.*

Proof. We will prove the statement for convergence in distribution and leave the proof for convergence in probability for homework. By Lemma 3.3, it suffices to show that

$$\mathbb{E}f(g(T_n)) \to \mathbb{E}f(g(T)) \tag{3.1}$$

for all bounded, continuous functions f. But for any such f, the function $f \circ g$ is bounded and continuous, so that (3.1) holds by the assumption that $T_n \xrightarrow{d} T$. □

Though we do not prove this, the conclusion of Proposition 3.5 also holds if g is only continuous on a set C such that $\mathbb{P}\{T \in C\} = 1$.

An important application of this principle is the following result:

Proposition 3.6 (Slutzky's Theorem)**.** *If $T_n \xrightarrow{d} T$ and $Y_n \xrightarrow{p} c$ for a* constant *c, then*

$$Y_n T_n \xrightarrow{d} cT\,.$$

Proof. By Proposition 3.4, $(T_n, Y_n) \xrightarrow{d} (T, c)$. Apply the continuous mapping theorem with the continuous function $g(t, y) = ty$. □

Together, Proposition 3.5 and Proposition 3.6 tell a powerful story: if we understand the limits of a sequence of random variables, we can also understand the limits of *functions* of random variables. This is extremely important in practice, because the central limit theorem guarantees that many natural objects have Normal limits—so understanding functions of these objects requires only understanding functions of normal random variables. To formalize this idea, we give a basic definition.

Definition 3.7. A sequence T_n of random variables satisfies a *central limit theorem* if there exist constants μ and σ^2 such that

$$\sqrt{n}(T_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)\,.$$

The sequence $\sqrt{n}(T_n - \mu)$ is called *asymptotically normal*.

As we have already seen, if T_n is the average of i.i.d. random variables $X_1, \ldots, X_n$, then the classical central limit theorem implies that $\sqrt{n}(T_n - \mathbb{E}X_1)$ is asymptotically normal with variance $\mathrm{Var}(X_1)$. An important fact is that many *other* objects also enjoy this property, and almost every piece of empirical analysis done in statistics relies on this fact!

The following theorem is the main result in the "calculus of random variables," and it shows the power of the asymptotic approach.

Theorem 3.8 (Delta method)**.** *Let g be a function that is differentiable at μ. If*

$$\sqrt{n}(T_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)\,,$$

then

$$\sqrt{n}(g(T_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, g'(\mu)^2\sigma^2)\,.$$

In words, if $\sqrt{n}(T_n - \mu)$ is asymptotically normal, then so is $\sqrt{n}(g(T_n) - g(\mu))$ for any differentiable function g.

Proof. The proof is based on Taylor expansion. The basic idea is that

$$g(T_n) \approx g(\mu) + g'(\mu)(T_n - \mu)\,,$$

so

$$\sqrt{n}(g(T_n) - g(\mu)) \approx \sqrt{n}(T_n - \mu)g'(\mu)\,,$$

which suggests the claim.

To be more formal, Taylor's theorem gives the existence of a function R such that

$$g(T_n) = g(\mu) + g'(\mu)(T_n - \mu) + R(T_n - \mu)\,,$$

where $R(h)/|h| \to 0$ as $h \to 0$. Therefore, the function $\delta(h)$ given by

$$\delta(h) \coloneqq \begin{cases} R(h)/h & \text{if } h \neq 0, \\ 0 & \text{if } h = 0 \end{cases}$$

is continuous at 0.

The fact that $\sqrt{n}(T_n - \mu)$ is asymptotically normal implies that $T_n - \mu \xrightarrow{p} 0$ (see Exercise 3), so Proposition 3.5 implies

$$\delta(T_n - \mu) \xrightarrow{p} 0\,.$$

Slutzky's theorem then implies

$$\sqrt{n}R(T_n - \mu) = \sqrt{n}(T_n - \mu)\delta(T_n - \mu) \xrightarrow{d} 0\,.$$

which implies $\sqrt{n}R(T_n - \mu) \xrightarrow{p} 0$ by Proposition 3.4.

We obtain

$$\sqrt{n}(g(T_n) - g(\mu)) = \sqrt{n}(T_n - \mu)g'(\mu) + \sqrt{n}R(T_n - \mu)\,,$$

where

$$\sqrt{n}(T_n - \mu)g'(\mu) \xrightarrow{d} \mathcal{N}(0, g'(\mu)^2\sigma^2)$$

and

$$\sqrt{n}R(T_n - \mu) \xrightarrow{p} 0\,.$$

Applying Proposition 3.4 and Proposition 3.5 to the pair $(\sqrt{n}(T_n - \mu)g'(\mu), \sqrt{n}R(T_n - \mu))$ yields the claim. □

Examining the proof of the delta method, it is clear that a similar statement will hold in more general contexts as long as the remainder term $R(T_n - \mu)$ can be controlled. This observation is the basis of several far-reaching generalizations of the delta method to infinite-dimensional settings.

3.3 Berry–Esseen Theorem

It is natural to ask whether the key statements of asymptotic statistics, such as the central limit theorem, can be made non-asymptotic. The *Berry–Esseeen theorem* provides one way of doing so, and shows that it is possible to give finite-sample results for distributional limit theorems.

We will present one version of this theorem that highlights the main ideas.

Theorem 3.9 (Berry–Esseen)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables with finite mean and variance σ^2, and let $\bar{X}_n$ be their average. If f is a continuous function such that f''' exists and is bounded, then*

$$|\mathbb{E}f(\sqrt{n}(\bar{X}_n - \mathbb{E}X_1)) - \mathbb{E}f(Z)| \leq \frac{3}{2\sqrt{n}}\|f'''\|_\infty \mathbb{E}|X_1 - \mathbb{E}X_1|^3\,,$$

where $Z \sim \mathcal{N}(0, \sigma^2)$ and where $\|f'''\|_\infty = \sup_{x\in\mathbb{R}} |f'''(x)|$.

Examining this statement in light of Lemma 3.3, we see that this theorem gives a version of convergence in distribution with explicit error bounds, so long as f is sufficiently smooth.

The proof uses a technique known as the *Lindeberg exchange method*.

Proof. We assume without loss of generality that the X_i are centered. Let $Z_1, \dots, Z_n \sim \mathcal{N}(0, \mathrm{Var}(X_1))$ be i.i.d. Note that these Z_i are chosen so that they match the first two moments of the X_i. Moreover, note that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i = \sqrt{n}(\bar{X}_n - \mathbb{E}X_1)$$
$$\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i \overset{d}{=} Z\,.$$

In other words, we can get from $\sqrt{n}(\bar{X}_n - \mathbb{E}X_1)$ to Z by replacing each X_i with the corresponding Z_i. This suggests a natural idea: let's replace these random variables one at a time!

Specifically, for $k = 0, \dots, n$, define

$$\bar{X}_n^{(k)} \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^k Z_i + \frac{1}{\sqrt{n}}\sum_{i=k+1}^n X_i\,.$$

In this notation, our goal is to bound

$$|\mathbb{E}f(\bar{X}_n^{(0)}) - \mathbb{E}f(\bar{X}_n^{(n)})| \leq \sum_{i=1}^n |\mathbb{E}f(\bar{X}_n^{(i-1)}) - \mathbb{E}f(\bar{X}_n^{(i)})|\,.$$

For any x and x', we have by Taylor's theorem that

$$f(x') = f(x) + (x' - x)f'(x) + \frac{1}{2}(x' - x)^2 f''(x) + R\,,$$

where R is a quantity satisfying

$$|R| \leq \frac{1}{6}\|f'''\|_\infty |x - x'|^3\,.$$

Letting

$$U_k \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^{k-1} Z_i + \frac{1}{\sqrt{n}}\sum_{i=k+1}^n X_i\,,$$

we have

$$\bar{X}_n^{(i-1)} = U_i + \frac{1}{\sqrt{n}} X_i$$
$$\bar{X}_n^{(i)} = U_i + \frac{1}{\sqrt{n}} Z_i \,.$$

We obtain that

$$\mathbb{E} f(\bar{X}_n^{(i-1)}) = \mathbb{E} f(U_i) + \frac{1}{\sqrt{n}} \mathbb{E} f'(U_i) X_i + \frac{1}{2n} \mathbb{E} f''(U_i) X_i^2 + R$$
$$\mathbb{E} f(\bar{X}_n^{(i)}) = \mathbb{E} f(U_i) + \frac{1}{\sqrt{n}} \mathbb{E} f'(U_i) Z_i + \frac{1}{2n} \mathbb{E} f''(U_i) Z_i^2 + R'$$

where the error terms satisfy

$$|R| + |R'| \leq \frac{1}{6n^{3/2}} \|f'''\|_\infty (\mathbb{E}|Z_i|^3 + \mathbb{E}|X_i|^3) \,.$$

We now make the crucial observation that X_i and Z_i are both independent of U_i, and that $\mathbb{E} X_i = \mathbb{E} Z_i$ and $\mathbb{E} X_i^2 = \mathbb{E} Z_i^2$. Therefore the first three terms in each expansion match exactly, and

$$|\mathbb{E} f(\bar{X}_n^{(i-1)}) - \mathbb{E} f(\bar{X}_n^{(i)})| \leq \frac{1}{6n^{3/2}} \|f'''\|_\infty (\mathbb{E}|Z_i|^3 + \mathbb{E}|X_i|^3) \,.$$

Summing these errors and using that $\mathbb{E}|Z_i|^3 \leq 8\sigma^3 \leq 8\mathbb{E}|X_i|^3$ (by Exercise 4 and Jensen's inequality) yields the claim. □

3.4 Exercises

1. This question investigates different modes of convergence.

   (a) Show that if $\mathbb{E}|X_n - X| \to 0$ as $n \to \infty$, then $X_n \xrightarrow{p} X$. Show that the conclusion is unchanged if $\mathbb{E}|X_n - X| \to 0$ is replaced by $\mathbb{E}|X_n - X|^r \to 0$ for any $r > 0$.

   (b) Let $X_n \sim \text{Bern}(\lambda_n)$, for some sequence λ_n of numbers in $(0, 1)$. Show that if $\lambda_n \to 0$ as $n \to \infty$, then $X_n \xrightarrow{p} 0$.

   (c) Let $Y_n = \lambda_n^{-1} X_n$, with X_n as in part (b). Show that if $\lambda_n \to 0$ as $n \to \infty$, then $Y_n \xrightarrow{p} 0$. What do you conclude about part (a)?

   (d) Let $X_n \sim \text{Bern}(1/2 + 1/n)$ be independent, and let $X \sim \text{Bern}(1/2)$. Does $X_n \xrightarrow{p} X$? Does $X_n \xrightarrow{d} X$?

2. This question proves Proposition 3.5 for convergence in probability, under the additional assumption that $(T_n)_{n \geq 1}$ and T all lie in a compact set. (This assumption is not necessary, but simplifies the proof.)

   (a) Assume that g is *Lipschitz*, that is, that there exists an $L \in \mathbb{R}$ such that $|g(x) - g(y)| \leq L|x - y|$ for all $x, y \in \mathbb{R}$. Show that for any $\varepsilon > 0$,

   $$\mathbb{P}\{|g(T_n) - g(T)| \geq \varepsilon\} \leq \mathbb{P}\{|T_n - T| \geq \varepsilon / L\} \,.$$

   Conclude that $g(T_n) \xrightarrow{p} g(T)$.

(b) Let g be a continuous function on a compact set K. It is a fact from real analysis that for any such function, there exists a continuous, non-decreasing function $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ called a *modulus of continuity* such that $\omega(0) = 0$ and

$$|g(x) - g(y)| \leq \omega(|x - y|) \qquad \forall x, y \in K\,.$$

Mimicking the proof in the first part, show that $g(T_n) \xrightarrow{p} g(T)$.

3. Show that if $\sqrt{n}(T_n - \mu)$ is asymptotically normal, then $T_n - \mu \xrightarrow{p} 0$. (Hint: For any random variable X we have

$$\mathbb{P}\{|X| \geq \varepsilon\} \leq \mathbb{P}\{X \leq -\varepsilon\} + (1 - \mathbb{P}\{X \leq \varepsilon/2\})\,,$$

for any $\varepsilon > 0$.)

4. In this exercise, we will prove the bound $\mathbb{E}|Z|^3 \leq 3\sqrt{2\pi}\sigma^3 \leq 8\sigma^3$ if $Z \sim \mathcal{N}(0, \sigma^2)$. In fact, we will prove something stronger: that the claim holds as long as Z is centered and σ^2-subgaussian.

   (a) Show that it suffices to prove the claim for $\sigma = 1$.

   (b) Show that

$$\mathbb{E}|Z|^3 = \int_0^\infty 3t^2 \mathbb{P}\{|Z| \geq t\}\, dt\,.$$

   (Hint: write $\int_0^\infty 3t^2 \mathbb{P}\{|Z| \geq t\}\, dt = \int_0^\infty \mathbb{E} 3t^2 \mathbb{1}_{|Z| \geq t} dt$ and exchange expectation and integration.)

   (c) Conclude via Corollary 1.8.

5. In statistical practice, it is often the case that a sequence of random variables T_n satisfies a central limit theorem with unknown μ. Moreover, in some situations, the limiting variance σ^2 can depend on μ, which poses a challenge when attempting to use asymptotic normality for inference. The delta method provides a trick for avoiding this problem.

   (a) Let μ be an unknown constant and suppose that a T_n satisfies a central limit theorem:

$$\sqrt{n}(T_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\mu))\,,$$

   where $\sigma : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a known function. Suppose that g is a function such that $g'(\mu) = \frac{1}{\sigma(\mu)}$. Show that $\sqrt{n}(g(T_n) - g(\mu))$ is asymptotically normal with variance 1, no matter what μ is. Such a g is called a *variance-stabilizing transformation*.

   (b) If $Z \sim \mathcal{N}(0, \sigma^2)$, then $\mathbb{E}Z^2 = \sigma^2$ and $\mathbb{E}Z^4 = 3\sigma^4$. If $Z_1, \ldots, Z_n \sim \mathcal{N}(0, \sigma^2)$ are i.i.d., show that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Z_i^2 - \sigma^2\right) \xrightarrow{d} \mathcal{N}(0, 2\sigma^4)$$

   (c) In the same setting as the previous item, show that

$$\sqrt{n}\left(\log\left(\frac{1}{n}\sum_{i=1}^n Z_i^2\right) - \log(\sigma^2)\right) \xrightarrow{d} \mathcal{N}(0, 2)\,.$$

Chapter 4

Statistical modeling

4.1 Models and statistical tasks

Recall that the fundamental statistical question is: given random observations, what probability distribution did they arise from? To specify the parameters of this question, we define the notion of a statistical model (also sometimes known as a *statistical experiment*).

Recall that a probability space is a tuple $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space, $\mathcal{F}$ is a collection of events, and $\mathbb{P}$ is a probability measure. The definition of a statistical model is a slight tweak of this definition:

Definition 4.1. A *statistical model* is a tuple $(\Omega, \mathcal{F}, \mathcal{P})$, where Ω is the sample space, $\mathcal{F}$ is a collection of events, and $\mathcal{P}$ is a *family* of probability measures.

When we fix a statistical model, we are assuming that our observation $\omega \in \Omega$ was generated according to some $\mathbb{P} \in \mathcal{P}$, and the fundamental goal is to decide which distribution in $\mathbb{P}$ best explains our observations.

Definition 4.1 is so general that it is almost meaningless; nevertheless, specifying the model—the set of probability distributions under consideration—is an extremely important part of statistical analysis. Here are some common models:

- $(\mathbb{R}, \{\text{all probability distributions on } \mathbb{R}\})$. This is the largest possible model on $\mathbb{R}$, since it doesn't exclude any distributions.
- $(\{0,1\}, \{\text{Bern}(p) : p \in [0,1]\})$. This is the set of all probability distributions on $\{0,1\}$.
- $(\mathbb{R}, \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \geq 0\})$, the set of all univariate Gaussian distributions.
- $(\mathbb{R} \times \mathbb{R}^p, \mathcal{P})$, where $\mathcal{P} = \{\text{Law}(Y, X) : Y = \langle X, \beta^* \rangle + \varepsilon, \beta^* \in \mathbb{R}^p, X \sim \mathcal{N}(0, I), \varepsilon \sim \mathcal{N}(0,1)\}$. Here, we are specifying a set of possible *joint* distributions for two random variables: X is a standard Gaussian random variable on $\mathbb{R}^p$, and the relationship between Y and X is given by the equation $Y = \langle X, \beta^* \rangle + \varepsilon$, where β^* is an unknown vector in $\mathbb{R}^p$. This "functional form" (where we indicate equations satisfied by our data) is a common way to specify a statistical model.

Since we will usually imagine that we obtain n i.i.d. observations, we will often abuse notation and use the same symbol to refer to the distribution of a single random variable as well as to the distribution of n i.i.d. copies of that random variable. For instance, if we observe $X_1, \ldots, X_n \sim \text{Bern}(p)$ for $p \in [0,1]$ unknown, we are actually considering the model $(\{0,1\}^n, \{\text{Bern}(p)^{\otimes n} : p \in [0,1]\})$, where $\text{Bern}(p)^{\otimes n}$ denotes the distribution of $\omega = (X_1, \ldots, X_n)$ when $X_1, \ldots, X_n$ are i.i.d.

The models listed above are all simple, but some models used in practice are enormously complex.[1] *Modeling* is the task of choosing an appropriate statistical model for a particular inference task. The challenge for the modeler is to balance two extremes:

[1] For instance, think about the forecasting models used by statisticians like Nate Silver to predict US presidential elections.

- If $\mathcal{P}$ is large, the model is richer. In other words, we can be more confident that our observations really do correspond to a distribution $P \in \mathcal{P}$. Unfortunately, doing inference in larger models is typically harder (both practically and theoretically).
- If $\mathcal{P}$ is small, then the model is simpler, and inference is typically easier. However, choosing a small model runs the risk of excluding the distribution P from which the observations originally arose. In other words, small models can fail to correctly capture phenomena present in the data.

The question of how to balance these considerations belongs to a topic called model selection, which we will investigate later in the course.

It is customary to label (or *parametrize*) the elements of $\mathcal{P}$ by writing $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ for some set Θ. This is without loss of generality, as we can always take $\Theta = \mathcal{P}$ (that is, label the elements of $\mathcal{P}$ by themselves). We will always assume that the model is *identifiable*, so that if $\theta_1 \neq \theta_2$, then $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$. In the situation when the model is correct, we will write θ_* for the true value of the parameter.

We also employ the following traditional names.

Definition 4.2. If $\Theta \subseteq \mathbb{R}^k$ for some k, then the model is called *parametric*. Otherwise, the model is called *non-parametric*.

The second, third, and fourth examples above are all parametric models. By contrast, the first example is nonparametric—there is no obvious way to label a distribution on $\mathbb{R}$ by an element of $\mathbb{R}^k$.[2] However, we can still label the elements of $\mathcal{P}$ in the first example: we can associate with each probability distribution its CDF.

4.2 Regression, Classification, and Clustering

The most natural question when faced with data from an unknown distribution $\mathbb{P}_\theta \in \mathcal{P}$ is "which distribution was it?" This is indeed the usual question that we will turn to, and it makes sense in any model. However, these are some special classes of model which appear often in statistics and machine learning, which give rise to special versions of this problem. It is worth keeping these examples in mind throughout the course.

4.2.1 Regression

Consider i.i.d. samples (X_i, Y_i) from an unknown joint distribution on $\mathbb{R}^p \times \mathbb{R}$. Here, X_i is a vector of *features* or *covariates*, and $Y_i \in \mathbb{R}$ is an *outcome* or *response*. The goal is to predict the response on the basis of the covariates, i.e., we seek a function f such that $f(X) \approx Y$, in some sense. We may imagine that Y has some natural variation, part of which is explained by X, and part of which is "unexplained" variation, coming from additional random noise. For this reason, one is therefore typically interested in the *regression function*

$$r(x) \coloneqq \mathbb{E}[Y \mid X = x]$$

which provides the "best guess" for the value of Y on the basis of having observed the covariate x.

If r is the optimal regression function of Y onto X, then we can define

$$\varepsilon \coloneqq Y - r(X)\,.$$

Note that these random variables satisfy

$$\mathbb{E}[\varepsilon \mid X] = \mathbb{E}[Y \mid X] - r(X) = r(X) - r(X) = 0\,.$$

Therefore, we can rewrite the relationship between X and Y as

$$Y = r(X) + \varepsilon\,, \qquad \mathbb{E}[\varepsilon \mid X] = 0\,. \tag{4.1}$$

[2] This distinction between parametric and nonparametric is not mathematically precise. For instance, since the number of probability distribution on $\mathbb{R}$ has the same cardinality as $\mathbb{R}$, there *does* exist a labeling of this set by elements of $\mathbb{R}$, but it is not natural. However, the traditional terminology is still widely used and is often appropriate.

The relationship in (4.1) is known as a *regression model*. Note that even though we call it a "model", we have actually not made any modeling restrictions so far—(4.1) is a completely general description of the joint relationship between X and Y that holds without any assumptions.[3] However, so far, this model is too general to be useful: we typically cannot say much about r without making further restrictions.

The main way of restricting this model is to restrict the types of functions that r can be. If r is assumed to be a linear function, so that $r(x) = \langle \beta, x \rangle$ for some $\beta \in \mathbb{R}^p$, then we recover the *linear model*. If we fix the distribution of ε and X, then this model is parametric (since it can be specified entirely by the finite-dimensional parameter β.

On the other hand, we can adopt other models for r as well. *Polynomial regression* is the model where $r(x)$ is assumed to be a polynomial rather than a linear function. This is also a parametric model. Nonparametric models include those where r is assumed to be smooth, for instance, to lie in a "Sobolev space":

$$\int (r'(x))^2 \, \mathrm{d}x < \infty,$$

or to possess certain geometric properties, for instance, to be strictly increasing, which is usually called *isotonic regression*.

In specifying a regression model, we typically have the choice of how to model the covariates X. In one, we specify a statistical model for X as well, by specifying a distribution (or family of distributions) which X could follows. In another, we fix a deterministic set of X's, and let ε alone be random. The first option is called "random design," and the second is called "fixed design." The first is the more common (and more realistic) setting; the second reflects the situation of the experimenter who is able to choose herself which X's to measure. (For example, the biologist who is able to create cell lines with particular features in the lab, and then measures them for certain properties.)

4.2.2 Classification

There is a special name for regression problems where Y takes only a discrete set of values: *classification*. When Y takes only the values 0 and 1, this is *binary classification*. The classification problem is mathematically identical to the regression problem, but the classical models are different. Indeed, except in special circumstances, the linear model where $r(x) = \langle \beta, x \rangle$ is usually not sensible for binary classification.

The classical solution to this problem is to choose an increasing function $g : [0, 1] \to \mathbb{R}$ and consider the model

$$g(r(x)) = \beta_0 + \langle \beta, x \rangle .$$

The function g is called a *link* function, and we can equivalently write

$$r(x) = g^{-1}(\beta_0 + \langle \beta, x \rangle) . \tag{4.2}$$

In other words, we consider the class of functions expressible as the *composition* of g^{-1} with a linear function. The benefit of this formulation is that we are guaranteed that $r(x) \in [0, 1]$, so that this class of functions always makes sense as a model for $\mathbb{E}[Y \mid X = x]$.

The classic choice for the link function is the logit function.

Definition 4.3. The *logit function* is the function $\operatorname{logit} : (0, 1) \to \mathbb{R}$ given by

$$\operatorname{logit}(p) \coloneqq \log \frac{p}{1 - p} .$$

The inverse of the logit is the logistic function.

Proposition 4.4. *The function* logit *is invertible, with inverse given by the* logistic function σ*, where*

$$\sigma(x) \coloneqq \operatorname{logit}^{-1}(x) = \frac{1}{1 + e^{-x}} .$$

[3] Technically, we need that $\mathbb{E}[Y \mid X]$ is finite, so that r is well defined.

Proof. A direct computation yields

$$\sigma(\text{logit}(p)) = \frac{1}{1 + (1-p)/p} = p \quad \forall p \in (0,1)$$

and

$$\text{logit}(\sigma(x)) = \log \frac{\frac{1}{1+e^{-x}}}{\frac{e^{-x}}{1+e^{-x}}} = x \quad \forall x \in \mathbb{R}\,.$$

□

Instantiating (4.2) with the choice $g = \text{logit}$, we obtain the *logistic regression model*:

$$r(x) = \sigma(\beta_0 + \beta^\top x)\,,$$

or, equivalently,

$$Y \mid X = x \sim \text{Bern}(\sigma(\beta_0 + \langle \beta, x \rangle))\,. \tag{4.3}$$

To gain intuition, note that σ is a strictly increasing function of x, so that larger values of $\beta_0 + \beta^\top x$ yield larger probabilities. The role of σ is to squeeze the whole real line $\mathbb{R}$ into the interval $(0,1)$. Also, σ has a nice symmetry property around 0:

$$\sigma(-x) = 1 - \sigma(x) \quad \forall x \in \mathbb{R}\,.$$

Of course, we can obtain other more complicated models by considering $\sigma \circ f$ for $f \in \mathcal{F}$ for some interesting class of functions $\mathcal{F}$. If you let $\mathcal{F}$ be the set of neural networks and you would like to perform binary classification, then this corresponds to using a "softmax activation" in the final layer.

4.2.3 Clustering

Both regression and classification are what are known in machine learning as "supervised" learning problems, since we observe both X (the covariate) and Y (the outcome we intend to predict). Clustering is what is called in machine learning "unsupervised" learning. Here, only covariates are observed, and the goal is to separate samples into different subgroups, with different statistical properties.

For instance, in 1894, Karl Pearson discovered that the width of the foreheads of crabs near Naples did not appear to follow a Gaussian distribution. (See Fig. 4.1.) Since the measurements of many parts of animals' bodies tend to have approximate Normal distributions, this suggested that there was an unexplained phenomenon at play. Pearson hypothesized that there were actually *two* species of crab. For crabs from each species, forehead widths did follow a Normal law. Formally, if a measurement X_i arose from a crab of species 1, then it might be modeled as a random variable with distribution $\mathcal{N}(\mu_1, \sigma_1^2)$; if it arose from species 2, then it might have law $\mathcal{N}(\mu_2, \sigma_2^2)$. If a fraction $\alpha \in [0,1]$ of crabs were of species 1, then the overall law of a forehead from a *random* crab is

$$X_i \sim \alpha \mathcal{N}(\mu_1, \sigma_1^2) + (1-\alpha)\mathcal{N}(\mu_2, \sigma_2^2)\,,$$

where the distribution on the right side is the continuous distribution with density

$$\frac{\alpha}{\sqrt{2\pi\sigma_1^2}} e^{-(x-\mu_1)^2/2\sigma_1^2} + \frac{(1-\alpha)}{\sqrt{2\pi\sigma_2^2}} e^{-(x-\mu_2)^2/2\sigma_2^2}\,.$$

This is a *Gaussian mixture model*, a fundamental model for clustering problems. Via very extensive calculations—without the aid of computers!—Pearson showed that this model was an excellent fit for the observations.

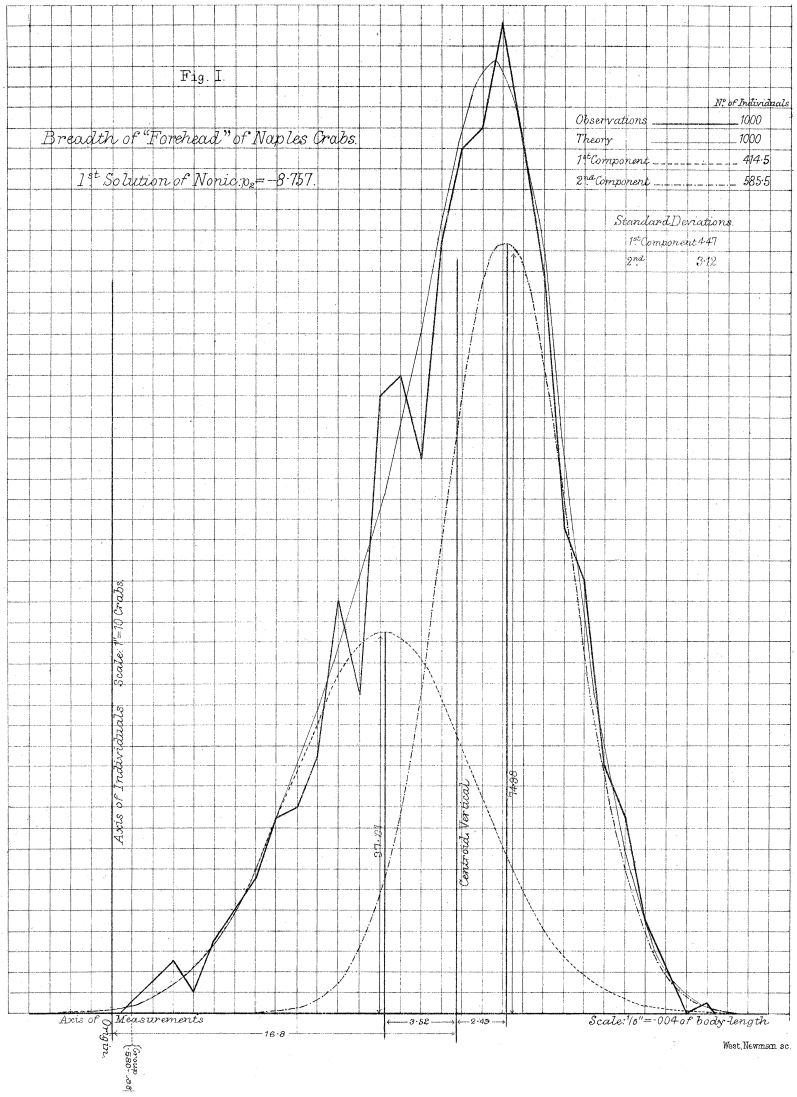

Figure 4.1: Pearson's crab data

4.3 Statistics and statistical tasks

In addition to being the name of the field, "statistic" has a particular technical meaning.

Definition 4.5. A *statistic* is a random variable on Ω. (That is, a function of the data.)

Usually these functions take values in $\mathbb{R}$ or $\mathbb{R}^p$ for some p. The goal of *statistics* (the field) is to choose good *statistics* (random variables) to compute.

To formalize this, we will give names to some canonical statistical problems. We observe data $X_1, \dots, X_n \sim \mathbb{P}_\theta \in \mathcal{P}$. The following tasks or variants thereof form the backbone of most work in statistics.

Point Estimation Construct $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ (i.e., a function of $X_1, \dots, X_n$) such that $\hat{\theta}$ is close to θ with high probability.

Confidence sets Construct $\hat{C} = \hat{C}(X_1, \dots, X_n) \subseteq \Theta$ so that $\hat{C} \ni \theta$ with high probability.[4]

Testing Given disjoint $\Theta_0, \Theta_1 \subseteq \Theta$, construct $\psi(X_1, \dots, X_n) \in \{0, 1\}$ such that if $\theta \in \Theta_i$, then $\psi = i$ with high probability.

Estimation, constructing confidence sets, and testing, as described above, all involve the judicious choice of some statistic (the estimators $\hat{\theta}$, the set $\hat{C}$, or the test ψ).

We will be interested in performing these tasks well, but it is not clear what "well" means. In 1939, Abraham Wald developed an influential framework for making questions of this type precise. This framework, called statistical decision theory, is quite abstract but has proven extremely important in the development of the mathematical theory of statistics. His main idea was to formalize statistics as the business of making decisions on the basis of data: for each decision, there is a cost. This cost depends both on what I choose to do and on the state of the world. If I decide to carry an umbrella because I think it will rain, I have "paid" a cost in my time or energy. On the other hand, if I do not carry an umbrella, but it does rain, then I have "paid" a different cost in being wet. Statistics then reduces to the study of which decision procedures incur the smallest cost. Since the cost we incur is random (it depends on the random state of the world), the best way to make sense of this question is to ask which procedures incur the smallest cost *in expectation*.

We first review the key objects in this formalism.

- **Model**: A family $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ of probability distributions on a sample space Ω.
- **Data**: We observe $\omega \in \Omega$.
- **Functional of interest**: A function $F : \Theta \to \mathcal{Y}$ reflecting the quantity we care about.
- **Decision rule**: A function $\hat{F} : \mathcal{X}^n \to \hat{\mathcal{Y}}$ reflecting the procedure we perform on the data.
- **Loss**: a function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$ which represents the penalty we obtain based on our functional and our decision. The quantity $\ell(F(\theta), \hat{F}(\omega))$, which we will sometimes abbreviate $\ell(F, \hat{F})$, is a random variable, because ω is random.
- **Risk:** to obtain a deterministic evaluation of our decision rule, we consider the expected loss, or risk:

$$\mathcal{R}_\theta(\hat{F}) = \mathbb{E}_\theta \ell(F(\theta), \hat{F}(\omega)) \,.$$

This framework is extremely general, and encompasses all the tasks defined above.

[4]The following point is often confusing: in all of the above examples, the parameter θ is **fixed**, while the observations $X_1, \dots, X_n$ (and therefore the estimators $\hat{\theta}$, the confidence set $\hat{C}$, and the test ψ) are random. For example, in the statement "$\hat{C} \ni \theta$ with high probability," we mean that we want the *random* set $\hat{C}$ to contain the *fixed* parameter θ with high probability.

Estimation Here, the application of Wald's ideas are relatively clear. Suppose that we consider the model $X_1, \dots, X_n \sim \mathcal{N}(\theta, I)$, where $\theta \in \mathbb{R}^p$.

Our functional F could be $F(\theta) = \theta$, if we want to estimate the parameter itself, but we could also consider other choices if we care about something more specific:

- $F(\theta) = \|\theta\|_2$
- $F(\theta) = \max_i \theta_i$
- $F(\theta) = |\{i : \theta_i \neq 0\}|$

As for the loss, the most typical choice is $\ell(F, \hat{F}) = (F - \hat{F})^2$ or $\ell(F, \hat{F}) = \|F - \hat{F}\|^2$ (the square loss), but we could also consider other options, depending on our goals:

- $\ell(F, \hat{F}) = |F - \hat{F}|$
- $\ell(F, \hat{F}) = \mathbb{1}_{|F - \hat{F}| \geq \tau}$

In each case, the loss function enforces the goal that $\hat{F}$ should be close to F.

Confidence sets Recall that that we would like to compute a random set C—say, a closed interval in $\mathbb{R}$—depending on our observations $X_1, \dots, X_n$ such that $C \ni \theta$ with high probability when $X_1, \dots, X_n \sim P_\theta$. To formalize this, we can consider $F(\theta) = \theta$ and let $\hat{F} : \mathcal{X}^n \to \{[a, b] \subseteq \mathbb{R} : a \leq b\}$ be a function which takes values in the set of closed intervals in $\mathbb{R}$. The function $\hat{F}$ represents our procedure for computing C on the basis of our observations. If we choose the loss function $\ell(F, \hat{F}) = \mathbb{1}_{F \notin \hat{F}}$, then we have exactly recovered the goal of constructing a confidence set.

Hypothesis testing Recall that in this task, we have two disjoint subsets Θ_0 and Θ_1 in Θ, and our goal is to decide whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$. In this case, we consider the functional $F : \Theta_0 \cup \Theta_1 \to \{0, 1\}$ given by

$$F(\theta) \coloneqq \begin{cases} 0 & \text{if } \theta \in \Theta_0, \\ 1 & \text{if } \theta \in \Theta_1. \end{cases}$$

Our decision rule $\hat{F}$—which in this case is called a *test*—is a function from $\mathcal{X}^n \to \{0, 1\}$. The simplest choice is to consider the loss $\ell(F, \hat{F}) = \mathbb{1}_{F \neq \hat{F}}$; however, it's often more useful in practice to allow for some asymmetry between Θ_0 and Θ_1, where we assign different costs to errors in one direction than in another. In this case, we could consider losses of the form $\ell(F, \hat{F}) = c_1 \mathbb{1}_{F=0, \hat{F}=1} + c_2 \mathbb{1}_{F=1, \hat{F}=0}$.

4.4 Exercises

1. This problem shows how concentration bounds can be used to obtain estimators, confidence sets, and tests. Suppose we observe n i.i.d. samples from a parametric model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$, and assume the existence of an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ such that

$$\mathbb{P}_\theta \left\{ |\hat{\theta} - \theta| \geq t \right\} \leq \rho(\sqrt{n} t) \qquad \forall t \geq 0, \theta \in \Theta \,, \tag{4.4}$$

   where $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a strictly decreasing, continuous function satisfying $\lim_{s \to \infty} \rho(s) = 0$. Note that the meaning of (4.4) is as follows: our data $\omega = (X_1, \dots, X_n)$ lives in the sample space Ω. No matter which probability measure $\mathbb{P}_\theta \in \mathcal{P}$ the space Ω is equipped with, the random variable $\hat{\theta}(\omega)$ under that measure satisfies (4.4).

   (a) Show under the Bernoulli model, with $X_1, \dots, X_n \sim \text{Bern}(\theta)$, and the Gaussian model, with $X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$, the estimator $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$ satisfies (4.4) with $\rho(t) = 2e^{-t^2/2}$.

(b) Show that (4.4) implies $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$. This property is known as *consistency*.

(c) Fix $\alpha \in (0,1)$. Show that if we define

$$\hat{C} \coloneqq [\hat{\theta} - \rho^{-1}(\alpha)/\sqrt{n}, \hat{\theta} + \rho^{-1}(\alpha)/\sqrt{n}]\,,$$

then $\mathbb{P}_\theta\left\{\theta \in \hat{C}\right\} \geq 1-\alpha$. State carefully the interpretation of this statement. Such a set is known as a $1-\alpha$ *confidence interval*.

(d) Let $\Theta_0, \Theta_1 \subseteq \Theta$ be *separated*, in the sense that

$$|\theta_0 - \theta_1| \geq 2\delta > 0 \quad \forall \theta_0 \in \Theta_0, \theta_1 \in \Theta_1\,.$$

Consider the test

$$\psi = \mathbb{1}_{\exists \theta_1 \in \Theta_1 \text{ s.t. } |\hat{\theta} - \theta_1| \leq \delta}\,.$$

Show that if $\delta > \rho^{-1}(\alpha)/\sqrt{n}$, then

$$\mathbb{P}_\theta\{\psi = i\} \geq 1 - \alpha \qquad \forall \theta \in \Theta_i, i \in \{0,1\}\,.$$

(e) Would the superficially similar test

$$\tilde{\psi} = \mathbb{1}_{\hat{\theta} \in \Theta_1}$$

yield the same guarantee? Why or why not?

This exercise justifies the attention we paid to concentration inequalities in the first lecture: with good concentration inequalities, we can estimate, create confidence sets, and test.

2. Suppose that we observe n i.i.d. samples from a parametric model $(\mathbb{R}, \mathcal{P})$, and suppose that under any $\mathbb{P}_\theta \in \mathcal{P}$, a sequence of statistics T_n satisfies a CLT centered at the parameter θ, i.e.,

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta)) \qquad \forall \theta \in \Theta\,,$$

for some function $\sigma : \mathbb{R} \to \mathbb{R}_{>0}$. Fix $\alpha \in (0,1)$. A sequence of sets $\hat{C}_n$ is called an *asymptotic* $(1-\alpha)$ *confidence set* if

$$\mathbb{P}_\theta\left\{\theta \in \hat{C}_n\right\} \underset{n\to\infty}{\to} 1 - \alpha \qquad \forall \theta \in \Theta\,.$$

Define $z_{\alpha/2}$ to be the unique positive real number satisfying

$$\mathbb{P}\left\{|Z| \geq z_{\alpha/2}\right\} = \alpha \quad Z \sim \mathcal{N}(0,1)\,.$$

Assuming that the function σ is continuous, show that the set

$$\hat{C}_n = [T_n - \sigma(T_n) z_{\alpha/2}/\sqrt{n}, T_n + \sigma(T_n) z_{\alpha/2}/\sqrt{n}]$$

is an asymptotic $(1-\alpha)$ confidence interval. (Hint: use Slutzky's theorem and the continuous mapping theorem.)

3. Consider n i.i.d. samples from the fully nonparametric model $(\mathbb{R}, \{\text{all probability distributions on } \mathbb{R}\})$. A pair of functions $\underline{F}, \overline{F} : \mathbb{R} \to \mathbb{R}$ constructed from the data is a $1-\alpha$ *confidence band* for the CDF F if

$$\mathbb{P}_F\left\{\underline{F}(t) \leq F(t) \leq \overline{F}(t)\ \forall t \in \mathbb{R}\right\} \geq 1 - \alpha \qquad \forall \mathbb{P}_F \in \mathcal{P}\,,$$

where $\mathbb{P}_F$ represents the probability measure with CDF F.

(a) Use (2.6) to construct a $1-\alpha$ confidence band for F.

(b) Show that if $(\underline{F}, \overline{F})$ is a $1-\alpha$ confidence band for F, then $(\max\{\underline{F}, 0\}, \min\{\overline{F}, 1\})$ is as well. Therefore, the confidence band constructed in part (a) can always be truncated (if necessary) so that both $0 \le \underline{F}(t) \le \overline{F}(t) \le 1$ for all $t \in \mathbb{R}$.

4. Given a statistical model $(\Omega, \mathcal{P})$, a function $g : \Omega \times \Theta \to \mathbb{R}$ is called *pivotal* if

$$\mathbb{P}_\theta \{g(\omega, \theta) \le t\} = \mathbb{P}_{\theta'} \{g(\omega, \theta') \le t\} \qquad \forall t \in \mathbb{R}, \theta, \theta' \in \Theta\,,$$

that is, if the distribution of the random variable $g(\omega, \theta)$ under $\mathbb{P}_\theta$ does not depend on θ. Note that $g(\omega, \theta)$ is *not* a statistic, because it is not a function of the data alone.

(a) Consider the Gaussian model, where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Show that $(\frac{1}{n}\sum_{i=1}^n X_i - \mu)/\sigma$ is pivotal.

(b) Suppose that $\underline{c}, \overline{c} \in \mathbb{R}$ satisfy

$$\mathbb{P}_\theta \{\underline{c} \le g(\omega, \theta) \le \overline{c}\} \ge 1 - \alpha\,. \tag{4.5}$$

Show that the set

$$C \coloneqq \{\theta \in \Theta : g(\omega, \theta) \in [\underline{c}, \overline{c}]\}$$

is a $1-\alpha$ confidence set.

(c) Why can $\underline{c}, \overline{c}$ satisfying (4.5) be computed without knowledge of θ?

Chapter 5

Point estimation

5.1 What is a good estimator?

Recall the following fundamental statistical task: given a model $\mathcal{P} = \{P_\theta : \Theta \in \Theta\}$ and i.i.d. samples from some $P_\theta \in \mathcal{P}$, construct a statistic $\hat{\theta}$ which is close to θ.[1] As usual, we will actually be interested in a *sequence* of statistics $\hat{\theta}_n$ indexed by sample size, and we will be interested in guaranteeing that $\hat{\theta}_n$ is close to θ when n is sufficiently large.

We have already seen one basic desirable property.

Definition 5.1. A sequence of statistics $T_n = T_n(X_1, \ldots, X_n)$ is a *consistent estimator* of θ if

$$T_n \xrightarrow{p} \theta$$

under $\mathbb{P}_\theta$, for all $\theta \in \Theta$.

We have already seen a very fundamental example of a consistent estimator: the sample average.

Example 5.2. Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$ be i.i.d. By the weak law of large numbers,

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}_\theta X_1 = \theta \quad \forall \theta \in \mathbb{R}\,.$$

Therefore, the sample average is a consistent estimator of θ.

In a sense, Example 5.2 is almost the *only* situation in which everyone can agree that the sample average is "the" canonical estimator of θ. As a result, the properties of this estimator have sometimes been used as a guide to say what properties other estimators should have.

5.1.1 Unbiasedness

We begin with a definition.

Definition 5.3. The *bias* of an estimator $\hat{\theta}$ of θ is

$$\mathbb{E}_\theta[\hat{\theta}] - \theta\,.$$

An estimator is *unbiased* if its bias is zero.

[1] More generally, we can try to construct a statistic $\hat{F}$ which is close to a functional of interest F.

Example 5.4 (Example 5.2, continued)**.** The sample average $\bar{X}_n$ satisfies

$$\mathbb{E}_\theta \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\theta X_i = \theta\,.$$

Therefore, the sample average is unbiased.

Being unbiased sounds really great (who wants bias?), and a great deal of early work in statistics put a lot of emphasis on finding estimators that are unbiased. However, this quality is now considered much less important—essentially all the interesting estimators in modern statistics are biased. Moreover, there are many estimation problems for which unbiased estimators do not exist. (Note that this is quite different from saying that our estimators are "biased" in the colloquial sense that they produce unfair or unjust outcomes. That can *also* be true, but it is distinct from the definition above.)

5.1.2 Asymptotic normality and efficiency

As has already been emphasized, asymptotic normality is a very useful property for a statistic to have, since it allows for the construction of asymptotic confidence sets. Of course, the average of Gaussian samples is asymptotically normal.

Example 5.5 (Example 5.2, continued)**.** The sample average $\bar{X}_n$ satisfies

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{d} \mathcal{N}(0,1) \qquad \text{under } \mathbb{P}_\theta.$$

In fact, we do not even need to take the limit: the distribution of $\sqrt{n}(\bar{X}_n - \theta)$ is exactly $\mathcal{N}(0,1)$ for all $n \geq 1$.

It will turn out that many estimators are asymptotically normal, which can be obtained as a consequence of the delta method. For instance, one can show that in the setting of Example 5.2, the median $\hat{X}_n^{\text{med}}$ of the samples is also asymptotically normal, and

$$\sqrt{n}(\hat{X}_n^{\text{med}} - \theta) \xrightarrow{d} \mathcal{N}(0, \pi/2)\,.$$

Since $\pi/2 > 1$, the classical theorists said we should prefer the sample average to the sample median: they have the same limiting distribution, but the *asymptotic variance* of $\bar{X}_n$ is smaller. This leads to a notion called *asymptotic efficiency*: an estimator is asymptotically efficient if it is asymptotically normal and its variance is as small as possible (in the sense that it matches a lower bound known as the *Cramér–Rao bound*). It turns out that $\bar{X}_n$ is efficient.

5.1.3 Decision theory framework

We can also adopt the decision theory framework. Let us employ the square loss: $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, so that the risk of an estimator is

$$\mathcal{R}_\theta(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta} - \theta)^2\,,$$

which is called the *mean squared error (MSE)*. We record an important fact about this risk.

Proposition 5.6 (Bias-Variance decomposition)**.** *When ℓ is the square loss, the risk satisfies*

$$\mathcal{R}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) + (\mathbb{E}_\theta\hat{\theta} - \theta)^2\,.$$

Proposition 5.6 says that the risk under the square loss can be high for two reasons: first, if $\hat{\theta}$ is highly variable (i.e., if the variance $\mathrm{Var}_\theta(\hat{\theta})$ is large) and second, if in expectation $\hat{\theta}$ does not agree with θ (i.e., if the bias is large in magnitude).

Proof of Proposition 5.6. Recall that for any random variable X and any $c \in \mathbb{R}$,

$$\mathbb{E}(X-c)^2 = \mathrm{Var}(X) + (\mathbb{E}X - c)^2 \,. \tag{5.1}$$

Indeed,

$$\begin{aligned}\mathbb{E}(X-c)^2 &= \mathbb{E}(X - \mathbb{E}X + \mathbb{E}X - c)^2 \\ &= \mathbb{E}(X-\mathbb{E}X)^2 + 2\mathbb{E}(X-\mathbb{E}X)(\mathbb{E}X - c) + (\mathbb{E}X-c)^2 \\ &= \mathrm{Var}(X) + (\mathbb{E}X - c)^2 \,.\end{aligned}$$

Applying (5.1) with $X = \hat{\theta}$ and $c = \theta$ yields the claim. □

Example 5.7 (Example 5.2, continued)**.** Under $\mathbb{P}_\theta$ the sample average $\bar{X}_n$ has distribution $\mathcal{N}(\theta, n^{-1})$; therefore,

$$\mathcal{R}_\theta(\bar{X}_n) = \mathbb{E}_\theta(\bar{X}_n - \theta)^2 = n^{-1} \,.$$

Since $\bar{X}_n$ is unbiased, the mean squared error arises entirely from the variance term in Proposition 5.6.

If we wish to ask whether the estimator $\bar{X}_n$ achieves the smallest possible risk, we of course need to ask, "at which θ?", since the risk depends on this quantity. Now, it is *not* true that $\bar{X}_n$ has lower risk than every other estimator.

Example 5.8 (Example 5.2, continued)**.** Consider the estimator $\hat{T} = 0$ which is identically zero. We have

$$\begin{aligned}\mathcal{R}_\theta(\bar{X}_n) &= \frac{1}{n} \\ \mathcal{R}_\theta(\hat{T}) &= \mathbb{E}_\theta(0-\theta)^2 = \theta^2 \,.\end{aligned}$$

Note that neither estimator dominates the other: for some θ, $\bar{X}_n$ achieves lower risk, and for others, $\hat{T}$ does. Though $\hat{T}$ is not a sensible estimator, it has lower risk than $\bar{X}_n$ when θ is very near the origin. In fact, when $\theta = 0$, we clearly have $\mathcal{R}_\theta(0) < \mathcal{R}_\theta(\bar{X}_n)$ for all $n \geq 1$. Note also that, unlike $\bar{X}_n$, the mean squared error of $\hat{T}$ is driven entirely by bias, since $\mathrm{Var}(\hat{T}) = 0$.

However, it is true that $\bar{X}_n$ achieves the best worst case (*minimax*) risk, in the sense that for any estimator $\hat{\theta}$, it holds

$$\sup_{\theta \in \mathbb{R}} \mathcal{R}_\theta(\hat{\theta}) \geq n^{-1} \,.$$

Therefore, if we decide to judge an estimator by its worst-case performance, then $\bar{X}_n$ is optimal. We will show how to establish such results later in this course.

5.1.4 How to generalize Example 5.2?

If everyone agrees that $\bar{X}_n$ is the only reasonable estimator for θ (and if it enjoys other properties like unbiasedness, asymptotic efficiency, and minimax optimality), we should ask what the analogy of $\bar{X}_n$ is for other models. There are at least *three* different ways to do this in general: the method of moments, maximum likelihood estimation (MLE), and M-estimation, which, as we will see, generalizes the previous two.

5.2 Method of Moments

Let us fix a model $\mathcal{P}$ and let $X_1, \ldots, X_n \sim P_\theta \in \mathcal{P}$ be i.i.d. Given $f : \mathcal{X} \to \mathbb{R}$, we can easily estimate the expectation of $f(X)$ when $X \sim P_\theta$, which we will write as $\mathbb{E}_\theta f(X)$. Indeed, by the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n f(X_i) \xrightarrow{p} \mathbb{E}_\theta f(X) \,.$$

Typically, the value of $\mathbb{E}_\theta f(X)$ will be different for different values of θ. So it's possible that, if we choose f carefully, knowing $\mathbb{E}_\theta f(X)$ will uniquely determine θ. More generally, if we know $\mathbb{E}_\theta f_1(X), \ldots, \mathbb{E}_\theta f_K(X)$ for suitable $f_1, \ldots, f_K$, we may be able to solve for θ.

One easy choice is to use low-degree polynomials.

Definition 5.9. For $k \geq 1$, the *kth moment* of a random variable X is

$$\mathbb{E}X^k \,.$$

Given samples $X_1, \ldots, X_n$, the *kth sample moment* is

$$\frac{1}{n}\sum_{i=1}^n X_i^k \,.$$

The method of moments estimator uses the moments to estimate θ.

Definition 5.10. Fix a model $\mathcal{P}$ and assume $X_1, \ldots, X_n \sim P \in \mathcal{P}$ are i.i.d. For any $k \geq 1$, write $m_k(\theta) = \mathbb{E}_\theta X^k$. Fix $K \geq 1$. The *method of moments estimator* $\hat{\theta}_{MOM}$ is any solution to

$$\frac{1}{n}\sum_{i=1}^n X_i^k = m_k(\theta) \quad \forall 1 \leq k \leq K \,. \tag{5.2}$$

In other words, if we write $\hat{m}_k$ for the kth sample moment, the method of moments says to find a $\theta \in \Theta$ such that $\hat{m}_k = m_k(\theta)$. In parametric models, K is typically chosen to be equal to the dimension of the parameter space, so that the system (5.2) has a unique solution.

Let us see how this method works on a few examples.

Example 5.11 (Gaussian mean)**.** Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$. What is the method of moments estimator for θ? We can compute that

$$m_1(\theta) = \mathbb{E}_\theta X = \theta \,,$$

so if we choose $K = 1$ then the method of moments estimator is any solution to

$$\frac{1}{n}\sum_{i=1}^n X_i = m_1(\theta) = \theta \,,$$

in other words, $\hat{\theta}_{MOM} = \bar{X}_n$. The method of moments therefore recovers the estimator considered in Example 5.2.

The method of moments was Pearson's approach for estimating the parameters of a Gaussian mixture model. For notational simplicity, we will show how this procedure works in a more constrained model.

Example 5.12 (Mixture of Gaussians)**.** Let $X_1, \ldots, X_n \sim \frac{1}{2}\mathcal{N}(\theta_1, 1) + \frac{1}{2}\mathcal{N}(\theta_2, 1)$. We will assume that $\theta_2 \geq \theta_1$; otherwise, the parameters are not identifiable. (Why?)

What does the method of moments yield? First, we compute

$$\begin{aligned} m_1(\theta_1, \theta_2) &= \frac{1}{2}(\theta_1 + \theta_2) \\ m_2(\theta_1, \theta_2) &= \frac{1}{2}(\theta_1^2 + \theta_2^2) + 1 \,. \end{aligned}$$

We therefore seek to solve the system of equations

$$\begin{aligned} \hat{m}_1 &= \frac{1}{2}(\theta_1 + \theta_2) \\ \hat{m}_2 &= \frac{1}{2}(\theta_1^2 + \theta_2^2) + 1 \,. \end{aligned}$$

for θ_1 and θ_2. If we write $s = \theta_1 + \theta_2$ and $d = \theta_2 - \theta_1$, then this system reads

$$\hat{m}_1 = \frac{1}{2}s$$
$$\hat{m}_2 = \frac{1}{4}(s^2 + d^2) + 1\,,$$

which, if we assume that $\hat{m}_2 - 1 \geq \hat{m}_1^2$, has solutions

$$s = 2\hat{m}_1$$
$$d = 2\sqrt{\hat{m}_2 - \hat{m}_1^2 - 1}\,.$$

Therefore the method of moments estimator is

$$\hat{\theta}_{MOM} = (\hat{m}_1 - \sqrt{\hat{m}_2 - \hat{m}_1^2 - 1}, \hat{m}_1 + \sqrt{\hat{m}_2 - \hat{m}_1^2 - 1})\,.$$

An important caveat is that the method of moments estimator may be undefined if the system (5.2) is inconsistent (i.e., fails to have a solution). For example, in Example 5.12, if $\hat{m}_2 - 1 < \hat{m}_1^2$, then the quadratic equation we obtained has no real solutions. A fix for this is to consider a slightly modified procedure known as the generalized method of moments. Intuitively, in this method, we seek an estimator which is only required to solve (5.2) approximately.

Definition 5.13. Fix $K \geq 1$. The *generalized method of moments estimator* $\hat{\theta}_{GMOM}$ is any solution to

$$\operatorname*{argmin}_{\theta \in \Theta} \sum_{k=1}^{K} (\hat{m}_k - m_k(\theta))^2\,. \tag{5.3}$$

This and similar procedures have proven very useful for analyzing models like the simple mixture of Gaussians in Example 5.12.

5.3 Maximum likelihood estimation (MLE)

The method of moments seems to be a little ad-hoc. Why use moments, rather than some other collection of test functions? Maximum likelihood estimation was advocated by Fisher as a more principled general technique. It is still the first technique one should try on a statistical model.

Maximum likelihood estimation is based on the likelihood function, which we now define: for simplicity, we assume that the distributions are either all continuous (with pdfs) or discrete (with pmfs).[2]

Definition 5.14. Fix a model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, and assume that $\mathbb{P}_\theta$ has pdf (or pmf) p_θ for all $\theta \in \Theta$. Given and observation ω, the *likelihood function* is

$$\mathcal{L}(\theta) \coloneqq p_\theta(\omega)\,.$$

The *log-likelihood* is $\ell(\theta) \coloneqq \log \mathcal{L}(\theta)$. When we wish to emphasize the dependence on the data, we write $\mathcal{L}(\theta) = \mathcal{L}(\theta \mid \omega)$.

Note that if $\omega = (X_1, \ldots, X_n)$ where X_i are assumed to be i.i.d., then the likelihood reduces to $\prod_{i=1}^n p_\theta(X_i)$. Using the log-likelihood instead of the likelihood is useful in this case because if $X_1, \ldots, X_n$ are i.i.d, then $\ell(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$ is a sum of i.i.d. random variables. Note that $\mathcal{L}(\theta)$ is exactly the joint

[2] If you know some measure theory, the correct condition is that there exists a σ-finite measure λ such that $\mathbb{P}_\theta \ll \lambda$ for all $\theta \in \Theta$. In the continuous case, we can take λ to be the Lebesgue measure; in the discrete case, we can take the counting measure on the sample space.

density of the data, except **we view it as a function of** θ. The likelihood can be viewed as a "function-valued statistic": like all statistics, the likelihood is defined based on observations, but the resulting object is not a number but a function from Θ to $\mathbb{R}$.

Since we typically are interested in comparing $\mathcal{L}(\theta)$ and $\mathcal{L}(\theta')$, it is enough to specify $\mathcal{L}$ up to a constant of proportionality. That is, we would draw the same inference if we replaced $\mathcal{L}(\theta)$ by $c \cdot \mathcal{L}(\theta)$, for any $c > 0$ independent of θ (but possibly dependent on ω). More formally, we can say that the likelihood in fact is a member of an equivalence class of functions, where two functions are equivalent if there exists a function $c : \Omega \to \mathbb{R}$ such that

$$f(\theta \mid \omega) = c(\omega) g(\theta \mid \omega)$$

where $c(\omega) > 0$ with probability 1 under all $\mathbb{P}_\theta \in \mathcal{P}$. When two functions are equivalent in this sense, it is common to write $f \propto g$.

The importance of the likelihood stems from the intuitive principle that if $\mathcal{L}(\theta) > \mathcal{L}(\theta')$, then θ "better explains" the observations than θ'. This is the basis of maximum likelihood estimation.

Definition 5.15. Fix a model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, and assume we observe $\omega = (X_1, \ldots, X_n)$ which are i.i.d. samples from an element of $\mathcal{P}$. The *maximum likelihood estimator (MLE)* is

$$\hat{\theta} \coloneqq \operatorname*{argmax}_{\theta \in \Theta} \mathcal{L}(\theta \mid X_1, \ldots, X_n) = \operatorname*{argmax}_{\theta \in \Theta} \ell(\theta \mid X_1, \ldots, X_n) = \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} \ell(\theta \mid X_i) \,.$$

In general, the maximizer of $\mathcal{L}(\theta \mid X_1, \ldots, X_n)$ may not be unique, in which case we can define $\hat{\theta}$ to be any maximizer of the likelihood function. It is also possible that no maximizers of $\mathcal{L}$ exist (for instance, if $\mathcal{L}$ grows without bound as $\theta \to \pm\infty$), in which case the MLE is undefined.

Note that if ℓ is differentiable, finding the MLE can often be reduced (e.g., when ℓ is concave and $\Theta = \mathbb{R}^k$) to solving the equation

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{i=1}^{n} \ell(\theta \mid X_i) = 0 \,.$$

Example 5.16 (Gaussian mean)**.** Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$ for some unknown $\theta \in \mathbb{R}$. What is the maximum likelihood estimator? By definition of the Gaussian distribution,

$$\mathcal{L}(\theta | x_i) = p_\theta(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \theta)^2} \propto e^{-\frac{1}{2}(x_i - \theta)^2} \,.$$

Therefore

$$\hat{\theta} = \operatorname*{argmax}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \ell(\theta | X_i) = \operatorname*{argmax}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} -\frac{1}{2}(X_i - \theta)^2 = \operatorname*{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (X_i - \theta)^2 \,.$$

By taking derivatives, we see that the unique solution of this equation is at $\hat{\theta} = \bar{X}_n$. We have again recovered

Example 5.17 (Linear model)**.** Consider the linear model

$$Y_i = \beta_0 + \beta^\top X_i + \varepsilon_i \,,$$

where $(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p$ is the unknown parameter of interest and $\varepsilon_i \sim \mathcal{N}(0, 1)$ are i.i.d. Note that we have not specified the distribution of the X_i—but we assume that this law does not depend on β, so it is not necessary to specify any model on X_i to perform maximum likelihood estimation.

Reasoning as before, we have

$$\mathcal{L}(\beta_0, \beta | (x_i, y_i)) \propto e^{-\frac{1}{2}(\beta^\top x_i + \beta_0 - y_i)^2} \,,$$

and since $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent, the log-likelihood reads

$$\sum_{i=1}^{n} -\frac{1}{2}(\beta^\top X_i + \beta_0 - Y_i)^2 \,.$$

Therefore the maximum-likelihood estimator for (β_0, β) is

$$(\hat{\beta}_0, \hat{\beta}) = \operatorname*{argmax}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \sum_{i=1}^n -\frac{1}{2}(\beta^\top X_i + \beta_0 - Y_i)^2 = \operatorname*{argmin}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \sum_{i=1}^n (\beta^\top X_i + \beta_0 - Y_i)^2 \,.$$

We therefore recognize that the maximum likelihood estimator is given by ordinary least squares, which we mentioned earlier in the course in connection with (1.1). This is a convex function of β_0, β and is easy to optimize. In fact, in this case, it is possible to solve it explicitly by computing derivatives. (This is why statisticians in the pre-computer era were especially fond of linear models!) We have

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^n (\beta^\top X_i + \beta_0 - Y_i)^2 = \sum_{i=1}^n 2(\beta_0 + \beta^\top X_i - Y_i) = 2\left(\beta_0 + \beta^\top \left(\sum_{i=1}^n X_i\right) - \left(\sum_{i=1}^n Y_i\right)\right)$$

$$\nabla_\beta \sum_{i=1}^n (\beta^\top X_i + \beta_0 - Y_i)^2 = \mathbb{E} 2X_i(\beta_0 + \beta^\top X_i - Y_i) = 2\left(\beta_0 \left(\sum_{i=1}^n X_i\right) + \beta^\top \left(\sum_{i=1}^n X_i X_i^\top\right) - \left(\sum_{i=1}^n X_i Y_i\right)\right) .$$

Writing $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and defining

$$\hat{\Sigma}_X = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top - \bar{X}\bar{X}^\top \in \mathbb{R}^{p \times p}$$

$$\hat{\Sigma}_{XY} = \frac{1}{n}\sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y} \in \mathbb{R}^p \,,$$

we obtain that the maximum likelihood estimator for β_0 satisfies

$$\hat{\beta}_0 = \bar{Y} - \beta^\top \bar{X}$$

and therefore

$$\hat{\Sigma}_X \beta = \hat{\Sigma}_{XY} \,.$$

As long as $\hat{\Sigma}_X$ is invertible, we obtain that $\hat{\beta} = \hat{\Sigma}_X^{-1}\hat{\Sigma}_{XY}$. If $\hat{\Sigma}_X$ is not invertible, then the MLE is not unique, but a minimizer can be found by using the *Moore–Penrose pseudoinverse* of $\hat{\Sigma}_X$ in place of $\hat{\Sigma}_X^{-1}$.

As Examples 5.16 and 5.17 suggest, Gaussian models often give rise to maximum likelihood estimators which are obtained by minimizing sums of squares—this whole class of methods is called "least squares."

Example 5.18 (Classification)**.** Consider a model parametrized by Θ, under which $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ are i.i.d, with $Y_i \mid X_i = x \sim \mathrm{Bern}(r_\theta(x))$, where $r_\theta : \mathbb{R}^p \to [0,1]$ for $\theta \in \Theta$. What is the MLE for θ?

We write the likelihood for a single sample:

$$\mathcal{L}(\theta \mid x_i, y_i) = r_\theta(x_i)^{y_i}(1 - r_\theta(x_i))^{1-y_i} \,.$$

Hence

$$\ell(\theta) = \sum_{i=1}^n Y_i \log r_\theta(X_i) + (1 - Y_i)\log(1 - r_\theta(X_i)) \,,$$

The maximum likelihood estimator is therefore

$$\operatorname*{argmax}_{\theta \in [0,1]} \sum_{i=1}^n Y_i \log r_\theta(X_i) + (1 - Y_i)\log(1 - r_\theta(X_i)) \,.$$

In machine learning, this function is called the "cross entropy."

If we focus on logistic regression, then we consider the model

$$r_\theta(x) = \sigma(\beta_0 + \beta^\top x) \qquad \theta = (\beta_0, \beta) \,.$$

In this case, maximum likelihood estimation corresponds to

$$\operatorname*{argmax}_{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta^\top X_i)Y_i - \log(1+e^{\beta_0+\beta^\top X_i})\,. \tag{5.4}$$

Indeed, The log-likelihood function satisfies

$$\begin{aligned}\ell(\theta \mid x_i, y_i) &= y_i \log(\sigma(\beta_0+\beta^\top x_i)) + (1-y_i)\log(1-\sigma(\beta_0+\beta^\top x_i))\\ &= \log(1-\sigma(\beta_0+\beta^\top x_i)) + y_i \operatorname{logit}(\sigma(\beta_0+\beta^\top x_i))\\ &= \log(1-\sigma(\beta_0+\beta^\top x_i)) + y_i(\beta_0+\beta^\top x_i)\\ &= \log(\sigma(-\beta_0-\beta^\top x_i)) + y_i(\beta_0+\beta^\top x_i)\,.\end{aligned}$$

Unlike ordinary least squares, the estimator in (5.4) does not have a closed-form expression. However, the objective is a concave function of β_0 and β and can be optimized efficiently.

5.4 M-estimation

Maximum likelihood estimation and the method of moments are two examples of a more general family of estimators. These estimators are very natural from the perspective of machine learning, since their definition and analysis resembles those of empirical risk minimization. (See Section 2.2.1.)

We first give the abstract definition.

Definition 5.19. Fix a family of probability distributions $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ on a space $\mathcal{X}$. We will be considering the i.i.d. setting, where $\Omega = \mathcal{X}^n$. Let $\rho(x,\theta) : \mathcal{X}\times\Theta \to \mathbb{R}$ be a function satisfying

$$\theta^* = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}_{\theta^*}\rho(X,\theta) \quad \forall \theta^* \in \Theta\,.$$

The *M-estimator* based on i.i.d. samples $X_1,\dots,X_n$ is any solution to

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\rho(X_i,\theta)\,.$$

The function ρ is often called a *contrast function*. This function is chosen so that, for any fixed $\theta^* \in \Theta$, minimizing the function $\theta \mapsto \mathbb{E}_{\theta^*}\rho(X,\theta)$ gives θ^*. M-estimation corresponds to the natural idea of minimizing $\theta \mapsto \frac{1}{n}\sum_{i=1}^{n}\rho(X_i,\theta)$ in place of $\theta \mapsto \mathbb{E}_{\theta^*}\rho(X,\theta)$. If $X_1,\dots,X_n \sim \mathbb{P}_{\theta^*}$ are i.i.d., the law of large numbers implies $\frac{1}{n}\sum_{i=1}^{n}\rho(X_i,\theta) \xrightarrow{p} \mathbb{E}_{\theta^*}\rho(X,\theta)$, which gives a justification for this idea.

Example 5.20 (Gaussian mean)**.** Let $X_1,\dots,X_n \sim \mathcal{N}(\theta,1)$. Consider the contrast function $\rho(x,\theta) = (\theta-x)^2$. If $X \sim \mathcal{N}(\theta^*,1)$, then

$$\mathbb{E}_{\theta^*}(X-\theta)^2 = (\theta^*-\theta)^2 + 1\,,$$

which implies that $\theta^* = \operatorname{argmin}_{\theta\in\mathbb{R}} \mathbb{E}_{\theta^*}\rho(X,\theta)$ for all $\theta^* \in \mathbb{R}$.

The corresponding M-estimator is

$$\operatorname*{argmin}_{\theta\in\mathbb{R}} \sum_{i=1}^{n}(\theta-X_i)^2\,.$$

We again recover the least squares estimator, our old friend the sample average.

Example 5.21 (Maximum likelihood estimation)**.** The most fundamental example of M-estimation is maximum likelihood estimation. (Indeed, the "M" in "M-estimation" stands for maximum likelihood.) This corresponds to the choice $\rho(x,\theta) = -\ell(\theta|x)$. The fact that θ^* maximizes $\mathbb{E}_{\theta^*}\ell(\theta|X)$ can be shown by defining a quantity known as the Kullback–Leibler divergence, which is of great importance in information theory.

Example 5.22 (Method of moments)**.** Let $m_k(\theta)$ be defined as in Section 5.2, and set

$$\rho(x,\theta) \coloneqq \sum_{k=1}^{K} (x^k - m_k(\theta))^2 .$$

Note that for any $\theta^* \in \Theta$,

$$\mathbb{E}_{\theta^*} \rho(X,\theta) = \sum_{k=1}^{K} \mathbb{E}_{\theta^*} (X^k - m_k(\theta))^2 = \sum_{k=1}^{K} \mathrm{Var}_{\theta^*}(X^k) + (\mathbb{E}_{\theta^*} X^k - m_k(\theta))^2 .$$

Hence, if K is chosen such that the map $\theta \mapsto (m_1(\theta), \dots, m_K(\theta))$ is injective, then θ^* is the unique minimizer of $\theta \mapsto \mathbb{E}_{\theta^*} \rho(X,\theta)$.

The corresponding M-estimator coincides with the generalized method of moments (5.3).

Originally, M-estimators were defined by Peter Huber in order to design estimators that are *robust* to corruptions. This can be viewed as a question of model misspecification: how should you design estimators when you are worried that part of your model is incorrect?

Example 5.23 (Gaussian mean with corruptions)**.** Suppose that $X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$ are i.i.d., and that $X_{n+1}, \dots, X_m$ are arbitrary *outliers*, which the statistician would like to ignore. However, the statistician does not know which of the n data points (out of the m total points) are trustworthy. What is a good way to estimate θ?

If she uses the sample average $\frac{1}{m}\sum_{i=1}^{m} X_i$, the result can fail to be consistent. We can decompose this statistic into two terms, one which gives the contribution coming from the good data, and the other giving the contribution from the outliers:

$$\frac{1}{m}\sum_{i=1}^{m} X_i = \frac{n}{m} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) + \frac{m-n}{m} \cdot \left(\frac{1}{m-n} \sum_{i=n+1}^{m} X_i \right) .$$

The average $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to θ by the law of large numbers, but we have no control over the second term, which can be very far from θ. The overall sample average can therefore fail to be close to θ, even if the number of outliers is small. We say that the sample average is *non-robust*.

Huber accounted for the non-robustness of the sample average by noting that it corresponds to the M-estimator with contrast function $\rho(x,\theta) = (x-\theta)^2$. This function has the good property that it grows as θ gets farther from x (which encourages the estimator to be near the mean); however, the problem is that it grows too fast, which makes it very sensitive to the values of outliers. Huber proposed to replace this function by a different function which grows more slowly. For fixed $t \geq 0$, he defined

$$\rho_t(x,\theta) \coloneqq \begin{cases} (x-\theta)^2 & \text{if } |x-\theta| \leq t \\ 2t|x-\theta| - t^2 & \text{otherwise.} \end{cases}$$

This is a differentiable function which behaves quadratically when $x \approx \theta$, and grows linearly when $|x-\theta|$ is large. The value of t defines where the change between the quadratic and linear regimes occurs. As $t \to 0$, the function $\frac{1}{2t}\rho_t$ approaches the simple absolute value function $|x-\theta|$, whose corresponding M-estimator is the sample median.

5.4.1 Consistency and Asymptotic Normality of M-estimators

It turns out that M-esitmators are consistent under mild assumptions The proof strategy is familiar: as in Section 2.2.1, it suffices to understand the uniform convergence of the contrast function over $\theta \in \Theta$.

Theorem 5.24. *Fix $\theta^* \in \Theta$, and assume $X_1, \dots, X_n \sim P_{\theta^*}$ are i.i.d. observations. Write $M(\theta) = \mathbb{E}_{\theta^*} \rho(X,\theta)$ and $M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \rho(X_i,\theta)$ for its empirical counterpart.*

Assume:

1. *For all* $\varepsilon > 0$,

$$\inf_{\theta : |\theta - \theta^*| \geq \varepsilon} M(\theta) > M(\theta^*) \,.$$

2.

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0 \,.$$

Then the M-estimator $\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} M_n(\theta)$ *is consistent:*

$$\hat{\theta} \xrightarrow{p} \theta^* \,.$$

Before giving the proof, let us understand the assumptions. The first assumption is essentially equivalent to requiring that θ^* is the unique minimizer of M. In practice, this assumption is easy to verify as long as ρ is properly chosen. The second assumption is stronger, and is related to the topics of Chapter 2: we will be able to prove such a result whenever we can prove a suitable maximal inequality, for instance by bracketing (Proposition 2.5).

The proof is very similar to the proof of the continuous mapping theorem.

Proof of Theorem 5.24. As in Section 2.2.1, we write

$$M(\hat{\theta}) - M(\theta^*) = M(\hat{\theta}) - M_n(\hat{\theta}) + M_n(\hat{\theta}) - M_n(\theta^*) + M_n(\theta^*) - M(\theta^*) \leq 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \,,$$

where we have used the fact that $M_n(\hat{\theta}) - M_n(\theta^*) \leq 0$ by the definition of $\hat{\theta}$. Since $M(\hat{\theta}) - M(\theta^*) \geq 0$, we have

$$|M(\hat{\theta}) - M(\theta^*)| \leq 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \,.$$

Hence, by the second assumption, $|M(\hat{\theta}) - M(\theta^*)| \xrightarrow{p} 0$.

By the first assumption, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$M(\theta) - M(\theta^*) \geq \delta \quad \text{if } |\theta - \theta^*| \geq \varepsilon.$$

Therefore, for any $\varepsilon > 0$,

$$\mathbb{P}\left\{ |\hat{\theta} - \theta^*| \geq \varepsilon \right\} \leq \mathbb{P}\left\{ |M(\hat{\theta}) - M(\theta^*)| \geq \delta \right\} \to 0 \,.$$

We obtain that $\hat{\theta} \xrightarrow{p} \theta^*$, as claimed. □

Note that the proof strategy of Theorem 5.24 also allows us to extract a *rate* of convergence, if desired: as long as we have quantitative versions of the two assumptions, we will be able to obtain a proof that $\hat{\theta}$ approaches θ^* at a specified rate. However, this rate is usually not sharp, and a more sophisticated technique is needed to get the right rate.

Under stronger assumptions, we are also able to obtain asymptotic normality statements for M-estimators. These assumptions are very common in the analysis of M-estimators: in fact, in statistics textbooks, they typically go by the name "the usual assumptions"!

Proposition 5.25. *Let* ρ *be a twice-differentiable contrast function on* Θ*, and* $\hat{\theta}$ *the associated M-estimator. If* $\hat{\theta}$ *is a consistent estimator of a univariate parameter* θ*, then under "the usual assumptions" (marked in the proof),*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\rho, \theta)) \,,$$

where

$$\sigma^2(\rho, \theta) = \frac{\mathbb{E}\left(\frac{\partial}{\partial \theta} \rho(X, \theta)\right)^2}{\left(\mathbb{E} \frac{\partial^2}{\partial \theta^2} \rho(X, \theta)\right)^2} \,. \tag{5.5}$$

Note that this proposition implies that, in parametric situations, $\hat{\theta} - \theta$ will typically have fluctuations of size $1/\sqrt{n}$, just like the sample average. The fact that this scaling is shared by a large number of estimators is the reason $1/\sqrt{n}$ is often called the "parametric rate."

Before giving the proof, it's useful to think about what the asymptotic variance $\sigma^2(\rho, \theta)$ is saying. This quantity can be large for two reasons: either $\mathbb{E}\left(\frac{\partial}{\partial\theta}\rho(X,\theta)\right)^2$ can be large—indicating that ρ has larger fluctuations—or $\mathbb{E}\frac{\partial^2}{\partial\theta^2}\rho(X,\theta)$ can be small—indicating that ρ has small curvature. If the curvature is small, then even a small fluctuation in $\frac{\partial}{\partial\theta}\rho(X,\theta)$ can lead to a large error in terms of how close $\hat{\theta}$ is to θ.

Proof. Let us write $\psi(x,\theta) = \frac{\partial}{\partial\theta}\rho(x,\theta)$. First, let us **assume** that $\hat{\theta}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\hat{\theta}) = 0\,,$$

which will hold as long as $\hat{\theta}$ is achieved in the interior of Θ. We also **assume** that $\mathbb{E}\psi(X,\theta) = 0$, which holds as long as θ is in the interior of Θ and θ minimizes $\theta' \mapsto \mathbb{E}\rho(X,\theta')$.

Under the assumption that $\hat{\theta}$ lies in the interior of Θ, we will have that ψ is differentiable at $\hat{\theta}$. As in the proof of Theorem 3.8, we have

$$0 = \frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\theta) + (\hat{\theta}-\theta)\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\psi(X_i,\theta) + \delta(X_i,\theta)\right\},$$

where $\delta(X_i,\theta)$ is a function such that $\delta(X_i,\hat{\theta}) \to 0$ as $\hat{\theta}\to\theta$. We **assume** this convergence is uniform, so that in fact $\sup_{x\in\mathcal{X}}\delta(X_i,\hat{\theta}) \to 0$ as $\hat{\theta}\to\theta$. Let us rewrite this as

$$\sqrt{n}(\hat{\theta}-\theta) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i,\theta)}{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\psi(X_i,\theta) + \delta(X_i,\theta)} \tag{5.6}$$

Note that $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\psi(X_i,\theta)$ is an average of i.i.d. terms, so by the Weak Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\psi(X_i,\theta) \xrightarrow{p} \mathbb{E}\frac{\partial}{\partial\theta}\psi(X,\theta)\,.$$

We **assume** that $\tau \coloneqq \mathbb{E}\frac{\partial}{\partial\theta}\psi(X,\theta) \neq 0$, which holds as long as θ is not a saddle point of $\mathbb{E}\rho(X,\theta)$. If $\hat{\theta}$ is consistent, then $\delta(X_i,\theta)\xrightarrow{p}0$. We obtain that the denominator of (5.6) converges in probability to τ, a nonzero constant. On the other hand, by the regular central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i,\theta) \xrightarrow{d} \mathcal{N}(0,\mathbb{E}\psi^2(X,\theta))$$

So, using Slutzky's theorem, we obtain that

$$\sqrt{n}(\hat{\theta}-\theta) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i,\theta)}{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\psi(X_i,\theta) + \delta(X_i,\theta)} \xrightarrow{d} \mathcal{N}(0,\mathbb{E}\psi^2(X,\theta)/\tau^2)\,,$$

and $\mathbb{E}\psi^2(X,\theta)/\tau^2 = \frac{\mathbb{E}\psi^2(X,\theta)}{(\mathbb{E}\frac{\partial}{\partial\theta}\psi(X,\theta))^2} = \sigma^2(\rho,\theta)$, as claimed. □

We conclude with a non-example, showing that if the assumed conditions fail to hold, then asymptotic normality can fail. The classic example is maximum likelihood estimation in the uniform model. Given i.i.d. samples $X_1,\ldots,X_n$ from $\mathrm{Unif}([0,\theta])$, for some unknown $\theta > 0$, the maximum likelihood estimator $\hat{\theta}$ is the largest sample. Note that the log-likelihood function $\ell(\theta \mid X_1,\ldots,X_n) = -n\log\theta + \log\mathbb{1}_{\max_i X_i\le\theta}$ is *not* differentiable, and so the Taylor expansion used in the proof cannot hold everywhere.

Note that, for $t \in [0, \theta]$,

$$\mathbb{P}\left\{\hat{\theta} < t\right\} = (t/\theta)^n \,,$$

which implies for $s > 0$ and n sufficiently large,

$$\mathbb{P}\left\{n(\theta - \hat{\theta}) \leq s\right\} = 1 - \mathbb{P}\left\{\theta - \hat{\theta} > s/n\right\} = 1 - \left(1 - \frac{s}{n\theta}\right)^n \underset{n\to\infty}{\to} 1 - e^{s/\theta} \,,$$

so

$$n(\theta - \hat{\theta}) \xrightarrow{d} \mathrm{Exp}(1/\theta) \,.$$

The limiting behavior of the MLE for this model is therefore totally different from what is predicted by Proposition 5.25.

5.5 Exercises

1. A parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of continuous distributions with corresponding densities p_θ is known as an *exponential family* if there exist functions $h : \mathbb{R} \to \mathbb{R}$, $T : \mathbb{R} \to \mathbb{R}$, and $A : \Theta \to \mathbb{R}$ such that

$$p_\theta(x) = h(x)e^{\theta T(x) - A(\theta)} \quad \forall \theta \in \Theta \,.$$

   (a) Show that h and A in the above definition are not unique (so that the model is not identifiable as written).

   (b) Show that the set of Gaussian distributions with variance 1 forms an exponential family. Identify T, h, and A in this example.

   (c) Show that the set of all distributions on $[1, +\infty)$ with densities of the form $p_\theta(x) = (\theta - 1)x^{-\theta}$ for $\theta > 1$ forms an exponential family. Identify T, h, and A.

   (d) Show that for any exponential family, $A(\theta) = \log \int h(x)e^{\theta T(x)} \, \mathrm{d}x$. Conclude that for all $\theta \in \Theta$, $A'(\theta) = \mathbb{E}_\theta T(X)$, where $\mathbb{E}_\theta$ is the expectation when $X \sim P_\theta$. (You may assume that it is valid to interchange differentiation and integration, and that Θ is open.)

   (e) Show that in an exponential family, if $X_1, \ldots, X_n \sim P_\theta$ are i.i.d., then any solution to

$$\frac{1}{n}\sum_{i=1}^n T(X_i) = \mathbb{E}_\theta T(X) \tag{5.7}$$

   is a maximum likelihood estimator. (Hint: you may use the fact the following two facts: A is a differentiable convex function of θ, and if f is a differentiable convex function and $f'(\theta) = 0$, then θ is a global minimum of f.)

   Note the similarity between (5.7) and (5.2). For exponential families, maximum likelihood estimation is equivalent to a method-of-moments-like procedure, with the function $T(X)$ used in place of X^k.

2. Let $X_1, \ldots, X_n \sim \mathrm{Unif}([0, \theta])$ be i.i.d. for some $\theta > 0$.

   (a) Compute the MLE of θ.

   (b) Use the method of moments to estimate θ.

   (c) Show that if $n \geq 3$, then with positive probability the method of moments applied to this example gives an estimator $\hat{\theta}$ which is unrealistic, in the sense that the statistician can be sure that $\mathrm{Unif}([0, \hat{\theta}])$ did *not* produce the data.

(d) Show that both the MLE and the method of moments estimator are consistent. (Hint: in the case of the MLE, it may be useful to use the fact that for independent random variables $X_1, \ldots, X_n$,

$$\mathbb{P}\left\{\max_{i=1,\ldots,n} X_i \leq t\right\} = \prod_{i=1}^{n} \mathbb{P}\left\{X_i \leq t\right\} ,$$

which follows from the fact that $\max_{i=1,\ldots,n} X_i \leq t$ if and only if $X_i \leq t$ for all i.)

3. This exercise will explore the notation of "asymptotic efficiency," and give another justification for the use of the maximum likelihood estimator. We assume in this problem that each $\mathbb{P}_\theta \in \mathcal{P}$ is continuous with a smooth, strictly positive density p_θ. We write $s(\theta \mid x)$ for the *score function*:

$$s(\theta \mid x) = \frac{\partial}{\partial\theta}\ell(\theta \mid x) = \frac{\partial}{\partial\theta}\log p_\theta(x) .$$

(a) Show that

$$s(\theta \mid \omega) := \sum_{i=1}^{n} s(\theta \mid x_i) = \frac{\frac{\partial}{\partial\theta}\prod_{i=1}^{n} p_\theta(x_i)}{\prod_{i=1}^{n} p_\theta(x_i)}$$

(b) Using part (a), show that for a suitably smooth function f of $\omega = (X_1, \ldots, X_n)$ for $X_1, \ldots, X_n$ i.i.d., we have

$$\mathbb{E}_\theta f(\omega)s(\theta \mid \omega) = \frac{\mathrm{d}}{\mathrm{d}\theta}\left\{\mathbb{E}_\theta f(\omega)\right\} .$$

(Hint: write

$$\mathbb{E}_\theta f(\omega)s(\theta \mid \omega) = \int f(x_1, \ldots, x_n)\frac{\frac{\partial}{\partial\theta}\prod_{i=1}^{n} p_\theta(x_i)}{\prod_{i=1}^{n} p_\theta(x_i)}\prod_{i=1}^{n} p_\theta(x_i) ,$$

and then assume that you can interchange differentiation and integration.)

(c) Let $\hat{\theta}$ be any unbiased estimator of θ. Using the previous part, show that

$$\mathbb{E}_\theta s(\theta \mid \omega) = 0 \qquad \forall\theta \in \Theta$$

and

$$\mathbb{E}_\theta\left\{\hat{\theta}(\omega)s(\theta \mid \omega)\right\} = 1 \qquad \forall\theta \in \Theta .$$

(d) Conclude that for any $\lambda \in \mathbb{R}$ and any unbiased estimator $\hat{\theta}$,

$$0 \leq \mathbb{E}_\theta(\lambda(\hat{\theta}(\omega) - \theta) - s(\theta \mid \omega))^2 = \lambda^2 \operatorname{Var}_\theta(\hat{\theta}) + \operatorname{Var}_\theta(s(\theta \mid \omega)) - 2\lambda .$$

In particular, by choosing $\lambda = \operatorname{Var}_\theta(s(\theta \mid \omega))$ and rearranging, show that any unbiased estimator $\hat{\theta}$ of θ satisfies

$$\operatorname{Var}_\theta(\hat{\theta}) \geq \frac{1}{\operatorname{Var}_\theta(s(\theta \mid \omega))} = \frac{1}{n\operatorname{Var}_\theta(s(\theta \mid X_1))} .$$

This is known as the *Cramér–Rao bound*, and $\operatorname{Var}_\theta(s(\theta \mid X_1))$ is known as the *Fisher information*, and denoted $I(\theta)$. It gives a lower bound on the variance of *any* unbiased estimator. An unbiased estimator matching this bound is called *efficient*.

(e) **[Optional, ungraded]** Show that when $\rho(x, \theta) = -\ell(\theta \mid x)$, then the asymptotic variance defined in (5.5) equals $\frac{1}{I(\theta)}$. This implies that the maximum likelihood estimator is asymptotically efficient.

4. Suppose that $Y_1, \ldots, Y_n$ satisfy

$$Y_i = \beta x_i + \varepsilon_i ,$$

where $x_1, \ldots x_n$ are fixed and known, $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ are i.i.d., and β and σ^2 are unknown.

(a) Compute the MLE $\hat{\beta}$ of β.

(b) Compute $\mathbb{E}\hat{\beta}$ and $\mathrm{Var}(\hat{\beta})$.

(c) Conclude that if $\sum_{i=1}^n x_i^2 \underset{n\to\infty}{\to} \infty$, then $\hat{\beta}$ is consistent.

(d) Give an example showing that if $\sum_{i=1}^n x_i^2$ does not approach infinity, then $\hat{\beta}$ can fail to be consistent.

5. Unlike linear regression, the maximum likelihood estimator in logistic regression does not always exist. This exercise examines this phenomenon.

(a) Consider a set of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $Y_i \in \{0, 1\}$ for all i. We say that a vector v *perfectly separates* the data if for all $i = 1, \ldots, n$,

$$\begin{aligned} Y_i = 1 &\implies X_i^\top v > 0 \\ Y_i = 0 &\implies X_i^\top v < 0\,. \end{aligned}$$

Explain in geometric terms why such a definition is sensible.

(b) Let us consider the simplified logistic regression model where $\beta_0 = 0$. Recall that the MLE, if it exists, solves

$$\operatorname*{argmax}_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \beta^\top X_i Y_i - \log(1 + e^{\beta^\top X_i})\,.$$

Show that

$$\beta^\top X_i Y_i - \log(1 + e^{\beta^\top X_i}) < 0$$

for all i.

(c) Suppose that v perfectly separates the data. Show that

$$v^\top X_i Y_i - \log(1 + e^{v^\top X_i}) = -\log(1 + e^{-|v^\top X_i|})\,.$$

Conclude that

$$\sup_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \beta^\top X_i Y_i - \log(1 + e^{\beta^\top X_i}) = 0\,.$$

(Hint: take $\beta = \lambda v$ for $\lambda > 0$, and let $\lambda \to \infty$.)

(d) Combining parts (b) and (c), show that if there exists a vector perfectly separating the data, then the MLE fails to exist.

Chapter 6

Testing

6.1 Null and alternative hypotheses

In hypothesis testing, we consider a model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ and two disjoint subsets $\Theta_0, \Theta_1 \subseteq \Theta$. The functional of interest is

$$F(\theta) = \begin{cases} 0 & \text{if } \theta \in \Theta_0 \\ 1 & \text{if } \theta \in \Theta_1. \end{cases}$$

A decision rule $\hat{F} : \Omega \to \{0, 1\}$ is called a *test*.

Definition 6.1. Given a test $\hat{F}$, the *rejection region* of $\hat{F}$ is the set

$$R \coloneqq \{\omega \in \Omega : \hat{F}(\omega) = 1\}.$$

Note that there is a one-to-one correspondence between tests and (Borel) subsets of Ω: we can associate any test $\hat{F}$ with its rejection region and, conversely, given a set $R \subseteq \Omega$, we can define an associated test $\hat{F} = \mathbb{1}_{\omega \in R}$ with rejection region R.

The most common loss function is

$$\ell(F, \hat{F}) = \mathbb{1}_{F \neq \hat{F}}$$

or, more generally,

$$\ell(F, \hat{F}) = c_1 \mathbb{1}_{F=0, \hat{F}=1} + c_2 \mathbb{1}_{F=1, \hat{F}=0} \tag{6.1}$$

for $c_1, c_2 \geq 0$.

Using this loss, the risk satisfies

$$\mathcal{R}_\theta(\hat{F}) = \begin{cases} c_1 \mathbb{P}_\theta(R) & \text{if } \theta \in \Theta_0 \\ c_2(1 - \mathbb{P}_\theta(R)) & \text{if } \theta \in \Theta_1, \end{cases} \tag{6.2}$$

where R is the rejection region of $\hat{F}$. To evaluate the risk of a test, it therefore suffices to know $\mathbb{P}_\theta(R)$ for all $\theta \in \Theta_0 \cup \Theta_1$.

Definition 6.2. The *power function* of a test with rejection region R is

$$\beta(\theta) \coloneqq \mathbb{P}_\theta(R).$$

The *size* of the test is

$$\alpha \coloneqq \sup_{\theta \in \Theta_0} \beta(\theta).$$

We say a test is *level* α if its size is at most α.

By examining (6.2), we see that the risk is minimized when the power function is large for $\theta \in \Theta_1$ but small for $\theta \in \Theta_0$. Specifying the size or level of a test guarantees that the power function is uniformly small when $\theta \in \Theta_0$. These goals are in tension with each other: by making the set R larger, we increase the power function, which is good when $\theta \in \Theta_1$ but bad when $\theta \in \Theta_0$. Typically, one imagines *fixing* a desired level α, and then choosing a test to make $\beta(\theta)$ as large as possible for $\theta \in \Theta_1$.

Definition 6.3. A test $\hat{F}$ with power function β is *more powerful* than a test $\hat{G}$ with power function β' if

$$\beta(\theta) \geq \beta'(\theta) \quad \forall \theta \in \Theta_1 \,.$$

Given two tests with the same size, (6.2) indicates that we should prefer the more powerful test.

Mathematically, the roles of Θ_0 and Θ_1 are exactly symmetric—nothing is lost if we rename Θ_0 as Θ_1, and vice-versa. However, traditionally, statisticians viewed these two sets very differently, and it is important to understand this distinction even though it is not mathematically relevant. The assumption $\theta \in \Theta_0$ is called the *null hypothesis* (written H_0), and is typically a very special, "default" assumption on the value of the parameter. The assumption $\theta \in \Theta_1$, called the *alternative* (written H_1), can sometimes be implicit, and represents a deviation from the null hypothesis, which the statistician may want to detect.

Example 6.4 (Lady tasting tea)**.** The first formal treatment of hypothesis testing is due to Fisher in 1935. A woman known to him claimed to be able to tell the difference between the taste of tea when the milk was added to the cup *before* or *after* the tea. Fisher proposed serving the woman eight cups of tea in a random order, four of which had milk added before tea and four of which had milk added after. The woman was able to label all eight cups correctly. Fisher's insight was that, under the "null" hypothesis that the woman had no ability to distinguish whether milk was added before or after, the probability that she would label all eight cups correctly was at most $1/\binom{8}{4} = 1/70 \approx .014$. Since this probability is quite small (smaller than $1/20 = .05$), Fisher says that this result represents a "significant discrepancy" from the null hypothesis.

Note that in this example, there is no mention of the alternative hypothesis. It is therefore not meaningful to speak of the power function in general, since we do not even model the set of possible alternatives. Nevertheless, we can still speak about the size of the test, since we know the probability of the event R under the null hypothesis that the woman cannot distinguish between the types of tea. There is an important asymmetry here, in that our test can only offer evidence *against* the null hypothesis. Says Fisher,

> In relation to any experiment we may speak of this hypothesis as the "null hypothesis," and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation.

For this reason, statisticians often speak of "rejecting the null" (when the outcome lies in the rejection region R) and "retaining the null" (when the outcome does not lie in R.)

The second fact to note is that the choice of cutoff $1/20$ is essentially arbitrary. Says Fisher,

> It is open to the experimenter to be more or less exacting in respect of the smallness of the probability [he or she] would require before [he or she] would be willing to admit that is observations have demonstrated a positive result.

The traditional name for a "false positive" (i.e., a rejection of H_0 when H_0 is true) is a *Type I error*. A "false negative" (i.e., retaining the null when H_0 is false) is called a *Type II error*.

6.2 Test statistics and p-values

How to design a test? The most common way is by choosing a test statistic. That is, given a real-valued statistic T and $c \in \mathbb{R}$, we consider the test with rejection region

$$R(c) = \{\omega : T(\omega) \geq c\} \,. \tag{6.3}$$

We write $R(c)$ to indicate that this actually gives a family of tests, indexed by c. We call T a *test statistic*, and c a *critical value*. Intuitively, T is chosen so that we expect it to be small under the null, and large under the alternative.

Example 6.5 (COVID-19)**.** Researchers estimate that a PCR test for COVID-19 performed three days after symptom onset has a false-negative rate of 20% and a false-positive rate of less than 1%. The PCR test functions by measuring the concentration of a piece of RNA associated with the SARS-CoV-2 virus.

In this example, the null hypothesis is "the patient does not have COVID" and the alternative hypothesis is "the patient does have COVID." Each of these two scenarios induces a distribution on the concentration of the viral RNA pieces, and the rejection region of the PCR test corresponds to particular choice of threshold value, above which the test is positive (i.e., the null is rejected). In other words, we are considering as a test statistic the concentration of viral RNA, and the experimenter has the freedom to set the critical value (i.e., the threshold level of viral RNA above which someone is declared to be COVID positive).

Writing P_{θ_0} and P_{θ_1} for the distributions under the null and alternative, respectively, a false-negative rate of 20% indicates that

$$\beta(\theta_1) = P_{\theta_1}(R) = .8\,,$$

i.e., that COVID-positive individuals obtain a positive test with 80% probability. A false-negative rate of at most 1% indicates that

$$\alpha = \beta(\theta_0) = P_{\theta_0}(R) \leq .01\,,$$

i.e., that the test has level .01.

In Example 6.5, by lowering the critical value, the experimenter can increase $\beta(\theta_1)$ at the price of possibly increasing α. Whether this tradeoff is justified depends on the costs associated with incorrectly labeling an infected person uninfected and an uninfected person infected. The fact that these two outcomes may not be equally bad is the reason for the freedom to choose $c_1 \neq c_2$ in (6.1).

To make this more precise, given a test statistic T, we may write the power function and size as a function of c alone. That is

$$\beta(\theta, c) = \mathbb{P}_\theta(R(c)) = \mathbb{P}_\theta(T \geq c) \tag{6.4}$$

and

$$\alpha(c) = \sup_{\theta \in \Theta_0} \beta(\theta, c)\,.$$

Clearly $\beta(\theta, c)$ is non-increasing in c.

6.2.1 p-values

It is often desirable to say something more precise than "reject H_0" or "retain H_0." The concept of a p-value can be used to quantify this distinction. If two experimenters agree on a test statistic, but disagree on which critical value to use to form the rejection region, they can still extract potentially meaningful information from the value of the test statistic evaluated on their data.

To put this in a slightly more general context, we note that the regions obtained in (6.3) are nested, in the sense that if regions R_α and $R_{\alpha'}$ have level α and α', respectively, then

$$R_\alpha \subseteq R_{\alpha'} \quad \text{if } \alpha \leq \alpha'. \tag{6.5}$$

We may therefore

Definition 6.6. Let $\{R_\alpha\}_{\alpha \in [0,1]}$ be a family of nested rejection regions, as in (6.5). The *p-value* is

$$p(\omega) \coloneqq \inf\{\alpha : \omega \in R_\alpha\}\,.$$

Note that the p-value is a statistic (i.e., a random variable). It represents the smallest α at which we would reject H_0. The magnitude of the p-value allows a researcher to assess the magnitude of evidence against H_0: if $p < .01$, this can be considered strong evidence against H_0, since it indicates that there is at most a 1% chance under H_0 of observing the given outcome. On the other hand, if p is large, say $p > .1$, then there is little evidence against H_0. Often, in applied statistics, a p-value less than .05 is deemed "significant."

In the context of regions based on test statistics, we employ the following simpler expression.

Definition 6.7. Let $\{R(c)\}_{c\in\mathbb{R}}$ be a family of tests based on a test statistic T. The *p-value* is

$$p(\omega) = \alpha(T(\omega))\,.$$

It can be shown that these two definitions agree when $\alpha(c)$ is a strictly decreasing continuous function of c. As we will see below, both definitions lead to what is known as a "valid p-value."

Warning #1 The p-value is *not* the probability that the null hypothesis is true—this statement is nonsensical in the absence of a prior distribution on the parameter space.

Warning #2 The p-value does *not* measure the "size" of an effect. It measures only how likely or not the observed outcome would be under the null hypothesis.

Since the p-value is a random variable, we can ask about its distribution under different assumptions on the data. The next result indicates that under the null hypothesis, this distribution can be precisely specified.

Proposition 6.8. *Let $p = p(\omega)$ be defined as in Definition 6.6 or Definition 6.7. Then*

$$\mathbb{P}_\theta\{p \le u\} \le u \quad \forall u \in [0,1], \theta \in \Theta_0\,. \tag{6.6}$$

This proposition indicates that, under the null, p is "well spread" over the interval $[0,1]$, so under the null, we are *unlikely* to see values of p that are very small. By contrast, under H_1, if the tests are powerful enough, the distribution of p will concentrate near 0.

Proof. If $u = 1$, the claim obviously holds. In the setting of Definition 6.6, for any $u \in [0,1)$, $p \le u$ implies that $\omega \in R_{u'}$ for all $u' > u$. Therefore, if $\theta \in \Theta_0$,

$$\mathbb{P}_\theta\{p \le u\} \le \mathbb{P}_\theta\{R_{u'}\} \le u' \quad \forall u' > u\,,$$

and letting $u' \to u$ yields the claim. In the setting of Definition 6.7, let us define a pseudo-inverse $\alpha^{[-1]} : [0,1] \to \mathbb{R} \cup \{\pm\infty\}$ by

$$\alpha^{[-1]}(u) = \sup\{c \in \mathbb{R} : \alpha(c) \ge u\}\,.$$

First, note that $\alpha(c) < u$ if and only if $c > \alpha^{[-1]}(u)$. Indeed, if $c > \alpha^{[-1]}(u)$, then $\alpha(c) < u$ by definition. In the other direction, we use the fact that the function α is left-continuous, i.e., for any sequence c_n approaching c from below, we have $\lim_{c_n \nearrow c} \alpha(c_n) = \alpha(c)$. This implies that if $\alpha(c) < u$, then $c > \alpha^{[-1]}(u)$.

Now, let $u' > u$ be arbitrary. We have

$$\begin{aligned}
\mathbb{P}_\theta\{\alpha(T(\omega)) \le u\} &\le \mathbb{P}_\theta\{\alpha(T(\omega)) < u'\} \\
&= \mathbb{P}_\theta\left\{T(\omega) > \alpha^{[-1]}(u')\right\} \\
&= \lim_{\varepsilon \searrow 0} \mathbb{P}_\theta\left\{T(\omega) \ge \alpha^{[-1]}(u') + \varepsilon\right\} \\
&\le \lim_{\varepsilon \searrow 0} \alpha(\alpha^{[-1]}(u') + \varepsilon) \le u'\,,
\end{aligned}$$

where the last step uses the fact that $\alpha^{[-1]}(u') + \varepsilon > \alpha^{[-1]}(u')$. Letting $u' \to u$ proves the claim. □

A statistic satisfying Eq. (6.6) is called a *valid p-value*. The following converse to Proposition 6.8 shows that, if $\hat{p}$ is a valid p-value, then it can be used to construct a set of nested rejection regions corresponding to tests of prescribed levels

Proposition 6.9. *Let $p = p(\omega)$ be any statistic such that (6.6) holds.*
Then for any $\alpha \in [0, 1]$, the test with rejection region

$$R_\alpha = \{\omega : p(\omega) \leq \alpha\}$$

has level α.

Proof. For any $\theta \in \Theta_0$,

$$\mathbb{P}_\theta \{R_\alpha\} = \mathbb{P}_\theta \{p(\omega) \leq \alpha\} \leq \alpha$$

by (6.6). Therefore the test associated with R_α has level α, as claimed. □

Propositions 6.8 and 6.9 make no assumptions on the alternative H_1, so there is no guarantee that the tests in question have good power under any alternative hypothesis. However, they do guarantee good behavior under the null.

6.3 Likelihood ratio tests

In estimation, a semi-canonical role is played by the maximum likelihood estimator. A similarly semi-canonical role is played in the theory of testing by *likelihood ratio tests*. The likelihood ratio test uses a version of the likelihood function as the test statistic.

Definition 6.10. The likelihood ratio test with critical value c is the test with rejection region

$$R_{\mathrm{LR}}(c) = \left\{\omega : 2 \sup_{\theta \in \Theta} \ell(\theta \mid \omega) - 2 \sup_{\theta \in \Theta_0} \ell(\theta \mid \omega) \geq c\right\}.$$

This is called a likelihood ratio test because the rejection region can also be written

$$\left\{\omega : \frac{\sup_{\theta \in \Theta} \mathcal{L}(\theta \mid \omega)}{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta \mid \omega)} \geq e^{c/2}\right\},$$

so that we can see more clearly that it is based on a ratio of likelihood functions.

The factor of 2 in the likelihood ratio test is a matter of mathematical convenience, and obviously does not affect the family of tests in question. Note that the test statistic $2 \sup_{\theta \in \Theta} \ell(\theta \mid \omega) - 2 \sup_{\theta \in \Theta_0} \ell(\theta \mid \omega)$ is always nonnegative (since the second supremum is over a smaller set). A large value of this statistic implies that the observations are much better explained by a $\theta \notin \Theta_0$ than one in Θ_0, giving evidence against the null.

Note also that $\sup_{\theta \in \Theta} \ell(\theta \mid \omega)$ is nothing but $\ell(\hat{\theta} \mid \omega)$, where $\hat{\theta}$ is the maximum likelihood estimator. A benefit of this test is that it is sometimes possible to compute the asymptotic distribution of the test statistic, under suitable regularity conditions. We will not pursue this direction here, because the regularity properties are often too strong to apply to interesting models, and are difficult to apply in practice.

A remarkably simple proposal from Wasserman, Ramdas, and Balakrishnan shows that a variant of this test actually is valid in the finite sample regime. It requires splitting the data $(X_1, \ldots, X_n)$ into two disjoint sets (D_0, D_1), which for convenience we take to be the same size, and write $\mathcal{L}_0(\theta) = \prod_{i \in D_0} \mathcal{L}(\theta \mid X_i)$ for the likelihood corresponding just to the samples in D_0.

Proposition 6.11. *Let $\hat{\theta}_1$ be* any *estimator constructed from D_1, and let*

$$\hat{\theta}_0 = \operatorname*{argmax}_{\theta \in \Theta_0} \mathcal{L}_0(\theta).$$

Then the test with rejection region

$$R_\alpha = \left\{\omega : \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)} \geq \frac{1}{\alpha}\right\}$$

has level α.

Proof. Fix $\theta \in \Theta_0$. We aim to bound

$$\mathbb{P}_\theta \{R_\alpha\} = \mathbb{P}_\theta \left\{ \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)} \geq \frac{1}{\alpha} \right\} .$$

Since the likelihood is nonnegative, Markov's inequality implies

$$\mathbb{P}_\theta \left\{ \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)} \geq \frac{1}{\alpha} \right\} \leq \alpha \mathbb{E}_\theta \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)} \leq \alpha \mathbb{E}_\theta \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\theta)} ,$$

where the second inequality follows from the fact that $\hat{\theta}_0$ maximizes the likelihood over Θ_0. We now use the crucial fact that for any fixed $\theta' \in \Theta$, we have

$$\mathbb{E}_\theta \frac{\mathcal{L}_0(\theta')}{\mathcal{L}_0(\theta)} = \mathbb{E}_\theta \frac{\prod_{i \in D_0} p_{\theta'}(X_i)}{\prod_{i \in D_0} p_\theta(X_i)} = \int_A \frac{\prod_{i \in D_0} p_{\theta'}(x_i)}{\prod_{i \in D_0} p_\theta(x_i)} \prod_{i \in D_0} p_\theta(x_i) \, \mathrm{d}x_i = \int_A \prod_{i \in D_0} p_{\theta'}(x_i) \, \mathrm{d}x_i \leq 1 ,$$

where A denotes the support of $\mathbb{P}_\theta$. Since D_1 is independent of D_0, we therefore obtain

$$\mathbb{E}_\theta \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\theta)} = \mathbb{E}_\theta \mathbb{E}_\theta \left[\frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\theta)} \Big| D_1 \right] \leq 1 ,$$

which proves the claim. □

6.3.1 Neyman-Pearson Lemma

In this section, we specialize to a situation we can show explicitly that the likelihood ratio test is actually optimal.

Definition 6.12. A set of distributions is called *simple* if it consists of a single distribution, otherwise it is called *composite*.

When the null and alternative hypotheses are simple, it is possible to completely specify the optimal test.

The following fundamental result says that for testing problems with simple hypotheses, tests based on the likelihood ratio are optimal.

Theorem 6.13 (Neyman-Pearson Lemma)**.** *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. For any critical value $c \geq 0$, suppose that the set*

$$R = \left\{ \omega : \frac{\mathcal{L}(\theta_1 \mid \omega)}{\mathcal{L}(\theta_0 \mid \omega)} > c \right\}$$

satisfies

$$P_{\theta_0}(R) = \alpha.$$

Then the test with rejection region R is the most powerful level α test. That is, any test with level α cannot have more power than R on the alternative.

Proof. Let S be the rejection region of any test with level α, i.e., a set such that

$$P_{\theta_0}(S) \leq \alpha .$$

It suffices to show that $P_{\theta_1}(S) \leq P_{\theta_1}(R)$.

Let us assume that P_{θ_0} and P_{θ_1} are continuous with densities p_0 and p_1. (The case where they are discrete is the same up to a change in notation.)

We claim that

$$\int_R (p_1(x) - cp_0(x))\,\mathrm{d}x \geq \int_S (p_1(x) - cp_0(x))\,\mathrm{d}x\,.$$

Indeed, by subtracting $\int_{R\cap S}(p_1(x) - cp_0(x))\,\mathrm{d}x$ from both sides, it suffices to show that

$$\int_{R\setminus S} (p_1(x) - cp_0(x))\,\mathrm{d}x \geq \int_{S\setminus R} (p_1(x) - cp_0(x))\,\mathrm{d}x\,. \tag{6.7}$$

But for $x \in R$, we have

$$\frac{p_1(x)}{p_0(x)} = \frac{\mathcal{L}(\theta_1 \mid x)}{\mathcal{L}(\theta_0 \mid x)} > c\,,$$

so that $p_1(x) - cp_0(x) > 0$. Conversely, for $x \notin R$, the same argument shows $p_1(x) - cp_0(x) \leq 0$. Therefore, the integrand on the left side of (6.7) is nonnegative and the integrand on the right is nonpositive, so the claimed inequality holds.

Rearranging, we have shown

$$\int_R p_1(x)\,\mathrm{d}x - \int_S p_1(x)\,\mathrm{d}x \geq c\left(\int_R p_0(x)\,\mathrm{d}x - \int_S p_0(x)\right)\,.$$

The right side of this expression is

$$c(P_{\theta_0}(R) - P_{\theta_0}(S)) = c(\alpha - P_{\theta_0}(S)) \geq 0$$

by assumption; hence

$$P_{\theta_1}(R) - P_{\theta_1}(S) \geq 0\,,$$

as claimed. □

Theorem 6.13 implies that when looking for optimal tests for simple-vs-simple testing problems, we can restrict our attention to tests based on the likelihood ratio.

6.4 Exercises

1. Let $X_1, \ldots, X_n \sim \text{Uniform}([0, \theta])$ be i.i.d., with $\theta > 0$ unknown. Let $Y = \max\{X_1, \ldots, X_n\}$. We want to test

$$\begin{aligned} \mathrm{H}_0 &: \theta = 1 \\ \mathrm{H}_1 &: \theta > 1\,. \end{aligned}$$

   For any $c \in \mathbb{R}$, consider the test with rejection region $\{Y \geq c\}$.

   (a) Find the power function, as a function of θ and c.

   (b) For what value of c does the test have size α?

   (c) Show how to calculate p values on the basis of this test.

   (d) Construct a $1-\alpha$ confidence set for θ. What relationship does this set have to the test constructed above?

2. This exercise investigates asymptotic tests.

   (a) Fix $\theta_0 \in \Theta$. Suppose that T_n is an estimator satisfying $\sqrt{n}(T_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma_0^2)$ under $\mathbb{P}_{\theta_0}$. The *Wald test* corresponding to the null hypothesis $\Theta_0 = \{\theta_0\}$ is the test with rejection region

$$R = \{\omega : |T_n(\omega) - \theta_0| \geq \sigma_0 z_{\alpha/2}/\sqrt{n}\}\,,$$

where for $\alpha \in (0,1)$, the real number $z_{\alpha/2}$ satisfies

$$\mathbb{P}\left\{|Z| \geq z_{\alpha/2}\right\} = \alpha\,, \qquad Z \sim \mathcal{N}(0,1)\,.$$

Show that this test has asymptotic level α, i.e.,

$$\mathbb{P}_{\theta_0}\left\{R\right\} \to \alpha\,.$$

(b) Write $\Phi(t) = \mathbb{P}\left\{Z \leq t\right\}$ where $Z \sim \mathcal{N}(0,1)$. Show that $p(\omega) := 2(1 - \Phi(\sqrt{n}/\sigma_0 |T_n(\omega) - \theta_0|))$ is an asymptotic p-value, i.e.,

$$\limsup_{n\to\infty} \mathbb{P}_{\theta_0}\left\{p(\omega) \leq u\right\} \leq u \qquad \forall u \in [0,1]\,.$$

(c) In 1000 tosses of a coin, 530 heads and 470 tails appear. Using the Wald test, assess the plausibility of the hypothesis that the coin is fair.

(d) Using Proposition 1.15, design a non-asymptotic test of the same hypothesis.

Chapter 7

Regularization

7.1 The bias-variance tradeoff revisited

We recall the basic decomposition we saw in Proposition 5.6: for any estimator $\hat{\theta}$ of θ, we have

$$\mathbb{E}_\theta(\hat{\theta} - \theta)^2 = \mathrm{Var}_\theta(\hat{\theta}) + (\theta - \mathbb{E}_\theta\hat{\theta})^2 \,.$$

For an unbiased estimator, this mean squared error is dominated by the variance. When this variance is larger on some parts of the parameter space than others, it is natural to ask if there is a smarter way to make this tradeoff. Balancing these two terms can yield an estimator with better overall performance.

Example 7.1. Let $X_1, \ldots, X_n \sim \mathrm{Bern}(p)$ be i.i.d., for some unknown $p \in [0,1]$. The maximum likelihood estimator for p is the sample average $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Let us examine the bias and variance of this estimator:

$$\mathbb{E}\bar{X} - p = \frac{\mathbb{1}}{n}\sum_{i=1}^n \mathbb{E}X_i - p = 0$$
$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{p(1-p)}{n} \,.$$

Clearly, the estimator $\bar{X}$ is unbiased. Its variance is a quadratic function of p, which takes its maximum value when $p = 1/2$ and decreases to 0 as $p \to 0$ and $p \to 1$. By Proposition 5.6, the behavior of this estimator is best when $p = 0$ or $p = 1$ and is worst at $p = 1/2$. If we are interested in obtaining a minimax-optimal estimator, however, it makes sense that we should modify this estimator so that it performs better when $p = 1/2$, even if this means slightly worsening the performance when $p = 0$ or $p = 1$. A natural way to do this is to bias the estimator towards $1/2$. For $\lambda \in [0,1]$, consider

$$\bar{X}_\lambda \coloneqq (1-\lambda)\bar{X} + \lambda \cdot \frac{1}{2} \,.$$

Intuitively, $\bar{X}_\lambda$ is a version of the sample average which we have pulled back towards $1/2$. Of course, this will help the performance when $p = 1/2$, but will possibly hurt this estimator's performance at other values of p. We calculate

$$\mathbb{E}\bar{X}_\lambda - p = (1-\lambda)p + \lambda \cdot \frac{1}{2} - p = \lambda\left(\frac{1}{2} - p\right)$$
$$\mathrm{Var}(\bar{X}_\lambda) = (1-\lambda)^2\,\mathrm{Var}(\bar{X}) = (1-\lambda)^2\frac{p(1-p)}{n} \,.$$

We see that this estimator is biased (i.e., $\mathbb{E}\bar{X}_\lambda - p \neq 0$) whenever $\lambda > 0$ and $p \neq 1/2$; on the other hand, when $\lambda > 0$, the variance of $\bar{X}_\lambda$ is strictly smaller than the variance of $\bar{X}$.

Applying Proposition 5.6, we obtain

$$\begin{aligned}\mathcal{R}_p(\bar{X}_\lambda) &= (1-\lambda)^2 \frac{p(1-p)}{n} + \lambda^2 \left(\frac{1}{2} - p\right)^2 \\ &= (1-\lambda)^2 \frac{p(1-p)}{n} - \lambda^2 p(1-p) + \frac{\lambda^2}{4} \\ &= \left(\frac{(1-\lambda)^2}{n} - \lambda^2\right) p(1-p) + \frac{\lambda^2}{4}\,.\end{aligned}$$

Depending on whether $\left(\frac{(1-\lambda)^2}{n} - \lambda^2\right)$ is positive or negative, this expression is maximized either at $p = 1/2$ (where it equals $\frac{(1-\lambda)^2}{4n}$) or at $p \in \{0, 1\}$ (where it equals $\frac{\lambda^2}{4}$). The worst-case risk is therefore minimized when

$$\frac{(1-\lambda)^2}{n} = \lambda^2\,,$$

i.e., when $\lambda = (\sqrt{n}+1)^{-1}$.

In other words, we can improve the worst-case risk of the sample average by biasing it slightly (by $\mathcal{O}(n^{-1/2})$) towards $1/2$.

This is a toy example, but it demonstrates an important principle: biasing an estimator towards a certain part of the parameter space can improve the performance of the estimator near that parameter, while hurting it at other points.

7.2 Regularization as a variance-reduction strategy

The philosophy of regularization is simple: *introduce bias to reduce the variance at parts of the parameters space*. This goal makes sense when the performance of a naive estimator varies over the parameter space, *or* when there is reason to believe that some parts of the parameter space are "more natural" than others. We have already seen an example of the first phenomenon, and we will now see an example of the second.

To see how this works, we consider an example in the context of linear regression. Recall the setting of Example 5.17. Let us consider a linear model without intercept, namely:

$$Y_i = \beta^\top X_i + \varepsilon_i\,, \qquad \mathbb{E}[\varepsilon_i | X_i] = 0\,. \tag{7.1}$$

You will show on your homework that we can employ this model without loss of generality by augmenting the covariates. One can show that in this case the least squares estimator of β is given by

$$\hat{\beta} = \hat{\Sigma}_X^{-1} \hat{\Sigma}_{XY}\,,$$

where

$$\begin{aligned}\hat{\Sigma}_X &= \frac{1}{n}\sum_{i=1}^n X_i X_i^\top \\ \hat{\Sigma}_{XY} &= \frac{1}{n}\sum_{i=1}^n Y_i X_i^\top\,,\end{aligned}$$

so long as Σ_X^{-1} is invertible.

Let us begin by evaluating the bias and variance of this estimator.

Proposition 7.2. *Assume that (X_i, Y_i) are i.i.d. copies of (X, Y) satisfying (7.1), and let*

$$\sigma^2 = \mathbb{E}[\varepsilon_i^2 \mid X_i]\,.$$

The least squares estimator $\hat{\beta}$ satisfies

$$\mathbb{E}[\hat{\beta}] = \beta$$

and

$$\mathbb{E}[\|\hat{\beta} - \beta\|^2] = \frac{\sigma^2}{n} \operatorname{tr}(\mathbb{E}[\hat{\Sigma}_X^{-1}])$$

Proof. We will in fact prove the stronger representations:

$$\mathbb{E}[\hat{\beta} \mid X_1, \dots, X_n] = \beta$$
$$\mathbb{E}[\|\hat{\beta} - \beta\|^2 \mid X_1, \dots, X_n] = \frac{\sigma^2}{n} \operatorname{tr}(\hat{\Sigma}_X^{-1}) \,.$$

The claim then follows upon taking expectations of the above representations.

Note that (7.1) implies

$$Y_i = X_i^\top \beta + \varepsilon_i \,.$$

Thus

$$\begin{aligned}
\hat{\beta} = \hat{\Sigma}_X^{-1} \hat{\Sigma}_{XY} &= \hat{\Sigma}_X^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i (X_i^\top \beta + \varepsilon_i) \right) \\
&= \hat{\Sigma}_X^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \beta \right) + \hat{\Sigma}_X^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i \varepsilon_i \right) \\
&= \hat{\Sigma}_X^{-1} \hat{\Sigma}_X^\top \beta + \hat{\Sigma}_X^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i \varepsilon_i \right) \\
&= \beta + \hat{\Sigma}_X^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i \varepsilon_i \right) .
\end{aligned}$$

Since $\mathbb{E}[\varepsilon_i \mid X_i] = 0$, we also have

$$\mathbb{E}[X_i \varepsilon_i \mid X_i] = 0 \,. \tag{7.2}$$

Therefore

$$\mathbb{E}[\hat{\beta} \mid X_1, \dots, X_n] = \beta \,.$$

Now, we turn to the variance. We have

$$\|\hat{\beta} - \beta\|^2 = \left\| \frac{1}{n} \sum_{i=1}^n \hat{\Sigma}_X^{-1} X_i \varepsilon_i \right\|^2 .$$

By assumption, the pairs (X_i, Y_i) (and therefore the pairs (X_i, ε_i)) are independent, which implies that

$$\mathbb{E}[\varepsilon_i \varepsilon_j \mid X_1, \dots, X_n] = \mathbb{E}[\varepsilon_i \varepsilon_j \mid X_i, X_j] = \mathbb{E}[\varepsilon_i \mid X_i] \mathbb{E}[\varepsilon_j \mid X_j] = 0 \,.$$

Therefore

$$\begin{aligned}
\mathbb{E}[\|\hat{\beta} - \beta\|^2 \mid X_1, \dots, X_n] &= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \left[\left\| \hat{\Sigma}_X^{-1} X_i \varepsilon_i \right\|^2 \mid X_i \right] \\
&= \frac{1}{n^2} \sum_{i=1}^n \left\| \hat{\Sigma}_X^{-1} X_i \right\|^2 \mathbb{E}[\varepsilon_i^2 \mid X_i] \\
&= \frac{\sigma^2}{n^2} \sum_{i=1}^n \left\| \hat{\Sigma}_X^{-1} X_i \right\|^2
\end{aligned}$$

To proceed, we will use two properties of the trace function: that it is linear (i.e., that $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$) and that for any vector $v \in \mathbb{R}^p$, $\|v\|^2 = \mathrm{tr}(vv^\top)$. Continuing from above, we have

$$\begin{aligned}
\mathbb{E}[\|\hat{\beta} - \beta\|^2 \mid X_1, \ldots, X_n] &= \frac{\sigma^2}{n^2} \sum_{i=1}^n \mathrm{tr}(\hat{\Sigma}_X^{-1} X_i X_i^\top \hat{\Sigma}_X^{-1}) \\
&= \frac{\sigma^2}{n} \,\mathrm{tr}\left(\hat{\Sigma}_X^{-1} \cdot \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \cdot \hat{\Sigma}_X^{-1}\right) \\
&= \frac{\sigma^2}{n} \,\mathrm{tr}\left(\hat{\Sigma}_X^{-1} \hat{\Sigma}_X \hat{\Sigma}_X^{-1}\right) \\
&= \frac{\sigma^2}{n} \,\mathrm{tr}(\hat{\Sigma}_X^{-1})\,.
\end{aligned}$$

□

In order for this estimator to be well-defined, we assumed that the empirical covariance matrix $\hat{\Sigma}_X$ was invertible. What happens if this assumption fails?

First, let's begin by trying to say when exactly this situation occurs. As you will have seen in your linear algebra classes, the matrix $\hat{\Sigma}_X$ is not invertible if it has nontrivial kernel, i.e., if there exists a nonzero $v \in \mathbb{R}^p$ such that

$$\hat{\Sigma}_X v = 0\,.$$

Simpler still, we can reduce this to a scalar equation by noting that it implies

$$v^\top \hat{\Sigma}_X v = 0\,.$$

By plugging in the explicit form of $\hat{\Sigma}_X$, we see that this is equivalent to

$$\frac{1}{n} \sum_{i=1}^n (X_i^\top v)^2 = 0\,.$$

Each term $(X_i^\top v)^2$ is nonnegative, so the only way their sum can be zero is if each term is zero. Therefore, if $\hat{\Sigma}_X$ is not invertible, then there exists a nonzero vector v such that

$$X_i^\top v = 0 \quad \forall i = 1, \ldots, n\,. \tag{7.3}$$

Conversely, if (7.3) holds, then clearly $\hat{\Sigma}_X v = 0$ holds. Therefore, the existence of a nonzero v satisfying (7.3) is equivalent to $\hat{\Sigma}_X$ failing to be invertible.

How should (7.3) be interpreted? We will give two ways: one geometric, one statistical. Geometrically, (7.3) says that there exists a vector v which is orthogonal to each covariate vector—in other words, the covariate vectors all lie on a hyperplane in $\mathbb{R}^p$. It is easiest to understand this when $p = 2$, when this reduces to saying that the covariate vectors are *colinear*. This means that some of the covariates are redundant, in the sense that they are deterministic functions of the other variables. Therefore, they offer no additional explanatory power in the regression analysis.

A more statistical perspective says that this is a failure of identifiability. If v is a nonzero vector orthogonal to all of the covariate vectors, then

$$Y_i = \beta^\top X_i + \varepsilon_i = (\beta + v)^\top X_i + \varepsilon_i\,,$$

so that β is not uniquely determined by the data. In this situation, there is no way to say that we have recovered the "true" β, because what's to say that the true β isn't $\beta + v$?

Everything we've said still holds in the situation where $\hat{\Sigma}_X$ is invertible, but has a very small eigenvalue. If the smallest eigenvalue of $\hat{\Sigma}_X$ is η, then the corresponding eigenvector v satisfies

$$\frac{1}{n} \sum_{i=1}^n (X_i^\top v)^2 = \eta\,,$$

and our argument from before establishes that if η is very small, it must be the case that each of the terms in this sum is small, so that the covariates are *approximately* orthogonal to v. In this setting, β is "nearly" un-identifiable, since we again have that

$$Y_i = \beta^\top X_i + \varepsilon_i \approx (\beta + v)^\top X_i + \varepsilon_i \,.$$

This will manifest in the variance of our estimator $\hat{\beta}$ becoming very large. An important fact about the trace of a matrix is that it is also equal to the sum of the eigenvalues of the matrix. Moreover, one can check that the eigenvalues of $\hat{\Sigma}_X$ are all nonnegative, which in particular implies

$$\operatorname{tr}(\hat{\Sigma}_X^{-1}) \geq \eta^{-1} \,,$$

so that the variance explodes as $\eta \to 0$.

There are a few possible fixes for this situation. First, if we observe or suspect colinearity in our covariates, we can simply drop covariates with no explanatory power. This will reduce the dimension of our space and eliminate the kernel of $\hat{\Sigma}_X$, which will at least ensure that the resulting covariance is invertible.

However, if the covariates are merely *approximately* colinear, we may not wish to drop variables. We are therefore in a situation which we covered earlier in the course: we would like to reduce the variance of an estimator, and a natural way to do so is to regularize. This may increase the bias, but it will be worth it if this increase is only minor compared to the decrease in the variance.

The classic way to do this for regression is to introduce an ℓ_2 penalty to the least squares objective:

$$\hat{\beta}_\lambda \coloneqq \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (\beta^\top X_i - Y_i)^2 + \lambda \|\beta\|^2 \,,$$

for some $\lambda > 0$. The resulting procedure is known as *ridge regression*. Implicitly, we are biasing ourselves towards solutions with smaller norm, because we view β closer to the origin as "more natural" than those which are very far from the origin.

Differentiating the objective with respect to β, we obtain that $\hat{\beta}_\lambda$ satisfies

$$\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \hat{\beta}_\lambda + \lambda \hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^n X_i Y_i \,,$$

or, equivalently,

$$(\hat{\Sigma}_X + \lambda I)\hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^n X_i Y_i \implies \hat{\beta}_\lambda = (\hat{\Sigma}_X + \lambda I)^{-1} \hat{\Sigma}_{XY}$$

Therefore, ridge regression is equivalent to modifying the empirical covariance $\hat{\Sigma}_X$ by adding a multiple of the identity function. If we mimic our earlier analysis of the least squares estimator, we can write this estimator as

$$\hat{\beta}_\lambda = (\hat{\Sigma}_X + \lambda I)^{-1} \hat{\Sigma}_X \beta + (\hat{\Sigma}_X + \lambda I)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i \varepsilon_i \right) .$$

By analyzing these two terms, we can obtain the bias and variance. Since $(\hat{\Sigma}_X + \lambda I)^{-1} \hat{\Sigma}_X \neq I$ if $\lambda > 0$, the expectation of this estimator will not be β. For the bias, we have

$$\|\mathbb{E}_\beta \hat{\beta}_\lambda - \beta\|^2 = \|((\hat{\Sigma}_X + \lambda I)^{-1} \hat{\Sigma}_X - I)\beta\|^2 \,.$$

For the variance, we have

$$\mathbb{E}[\|\hat{\beta}_\lambda - \mathbb{E}\hat{\beta}_\lambda\|^2 \mid X_1, \ldots, X_n] = \frac{\sigma^2}{n} \operatorname{tr}((\hat{\Sigma}_X + \lambda I)^{-1} \hat{\Sigma}_X (\hat{\Sigma}_X + \lambda I)^{-1}) \,.$$

To understand these terms, let us first estimate the variance. The eigenvalues of the matrix $(\hat{\Sigma}_X + \lambda I)^{-1}\hat{\Sigma}_X(\hat{\Sigma}_X + \lambda I)^{-1}$ are all at most λ^{-1}, We obtain that

$$\frac{\sigma^2}{n}\operatorname{tr}((\hat{\Sigma}_X + \lambda I)^{-1}\hat{\Sigma}_X(\hat{\Sigma}_X + \lambda I)^{-1}) \le \frac{\sigma^2}{n} \cdot \frac{p}{\lambda}\,.$$

This quantity decreases as λ increases, indicating that introducing the ridge penalty indeed gives control over the variance.

The bias is slightly more complicated to understand. Let us assume that $\hat{\Sigma}_X$ is a diagonal matrix. (Though this assumption seems restrictive, it can be achieved by a change of basis.) Then we can write $\hat{\Sigma}_X = \operatorname{diag}(\eta)$, where $\eta = (\eta_1, \dots, \eta_n)$ is a vector of eigenvalues of $\hat{\Sigma}_X$. Then we have explicitly that $(\hat{\Sigma}_X + \lambda I)^{-1}\hat{\Sigma}_X$ is a diagonal matrix with entries

$$\left((\hat{\Sigma}_X + \lambda I)^{-1}\hat{\Sigma}_X\right)_{ii} = \frac{\eta_i}{\eta_i + \lambda}\,.$$

To interpret this, we see that when $\eta_i \gg \lambda$, this quantity is approximately 1, so that this matrix is approximately the identity on subspaces corresponding to large eigenvalues. By contrast, when $\eta_i \ll \lambda$, this quantity is approximately $\frac{\eta_i}{\lambda}$, so that this matrix dampens the eigenspaces corresponding to small egienvalues. Though this introduces bias in these subspaces, but when η_i is very small it is more than compensated for by the decrease in variance, so long as λ is chosen properly. Note that the behavior of this estimator can be seen as a "smoothed" version of the procedure which deletes subspaces corresponding to small eigenvalues.

If we let η be the smallest eigenvalue of $\hat{\Sigma}_X$, then one can show

$$\|((\hat{\Sigma}_X + \lambda I)^{-1}\hat{\Sigma}_X - I)\beta\|^2 \le \max_i \frac{\lambda^2}{(\eta + \lambda)^2}\|\beta\|^2 \le \|\beta\|^2\,,$$

and that the first inequality is tight if β lies in the subspace corresponding to the eigenvector with smallest eigenvalue. Note that if $\lambda > 0$, this expression is unbounded as $\beta \to \infty$—we have introduced a significant amount of bias for very large β. But if $\|\beta\|$ is not too large, we may have gained a lot in terms of variance.

7.3 Smoothness in the Gaussian sequence model

Here is another example of regularization, which seeks to exploit an idea about which parts of the parameter space are more natural than others.

Let us consider a different version of this problem, where the means are allowed to differ across observations. Let $\theta_1, \dots, \theta_p \in \mathbb{R}$ be unknown parameters, and consider observations of the form

$$Y_i = \theta_i + \varepsilon_i\,, \tag{7.4}$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$ are i.i.d. Equivalently, we can write this in vector form as

$$\mathbf{Y} = \theta + \mathcal{E}\,, \tag{7.5}$$

where $\theta \in \Theta = \mathbb{R}^p$ and $\mathcal{E} \sim \mathcal{N}(0, I)$. The natural loss function for estimating θ is the *normalized* square loss:

$$\ell(\theta, \hat{\theta}) = \frac{1}{p}\|\theta - \hat{\theta}\|^2 \coloneqq \frac{1}{p}\sum_{i=1}^{p}(\theta_i - \hat{\theta}_i)^2\,. \tag{7.6}$$

The normalizing factor $\frac{1}{p}$ is used to make an apples-to-apples comparison between this model for different values of p.

The formulation (7.5) is known as the *Gaussian sequence model*. The maximum likelihood estimator of θ in this model is $\mathbf{Y}$, whose risk is

$$\mathcal{R}_\theta(\mathbf{Y}) = \mathbb{E}_\theta\frac{1}{p}\|\theta - \mathbf{Y}\|^2 = \mathbb{E}_\theta\frac{1}{p}\sum_{i=1}^{p}\varepsilon_i^2 = 1\,.$$

That is, as p grows, this problem does not get any easier, even though there are multiple observations. And this is quite intuitive, because the parameter space $\mathbb{R}^p$ is getting bigger and bigger. Moreover, since $\mathbf{Y}$ is unbiased, all the difficulty stems from variance. And indeed one can show that if the parameter space is all of $\mathbb{R}^p$, this issue is unavoidable.

Now, let us see what happens when we change the parameter space. Suppose that we interpret (7.4) as modeling a situation where we obtain noisy observations of a certain process (say, the temperature at a given location) at fixed times. It is then reasonable to expect that θ_i and θ_{i+1} are relatively close together. To formalize this, we assume that θ lies in the restricted parameter space of *Lipschitz* vectors:

$$\Theta_{\mathrm{Lip}}(L) \coloneqq \{\theta \in \mathbb{R}^p : |\theta_i - \theta_{i+1}| \le L/p \quad \forall i = 1, \dots, p-1\}\,.$$

The normalization L/p is reasonable here because we imagine that increasing p corresponds to denser observations of our underlying process. (Say, observing the temperature every minute instead of every hour.)

If we believe that our parameter lies in $\Theta_{\mathrm{Lip}}(L)$ as the parameter space, then it seems plausible that the risk should be lower. To exploit this, we bias our estimator by forcing it to be "smooth" (specifically, we require that it be piecewise constant, with a small number of jumps.

Though $\mathbf{Y}$ is still a valid unbiased estimator for θ, its variance is still too large. To achieve a better result, it is necessary to focus on a biased estimator. The key insight is that the structure of the parameter space implies that we can share information between coordinates.

Proposition 7.3. *There exists an estimator in the model* (7.4) *which achieves*

$$\sup_{\theta \in \Theta_{\mathrm{Lip}}(L)} \mathcal{R}_\theta(\hat\theta) \lesssim \max\left\{\frac{L^{2/3}}{p^{2/3}}, \frac{1}{p}\right\}.$$

The notation $a \lesssim b$ means that $a \le Cb$ for some positive constant C.

Proof. We give an upper bound by constructing an estimator based on local averaging. Let us fix h, and assume for simplicity that h divides p. For $\ell = 1, \dots, p/h$, write I_ℓ for the set $\{(\ell-1)h+1, \dots, \ell h\}$. We define an estimator $\hat\theta^{(h)}$ which takes local averages:

$$\hat\theta_i^{(h)} = \frac{1}{h}\sum_{j \in I_\ell} \mathbf{Y}_j \quad \forall i \in I_\ell\,.$$

Let us also define $\bar\theta^{(h)}$ for the averaged version of θ:

$$\bar\theta_i^{(h)} = \frac{1}{h}\sum_{j \in I_\ell} \theta_j \quad \forall i \in I_\ell\,.$$

Let us compute the risk of $\hat\theta^{(h)}$. We have

$$\mathbb{E}\hat\theta^{(h)} = \bar\theta^{(h)}\,.$$

and

$$\mathrm{Var}(\hat\theta_i^{(h)}) = \mathbb{E}(\hat\theta_i^{(h)} - \bar\theta_i^{(h)})^2 = \mathbb{E}\left(\frac{1}{h}\sum_{j \in I_\ell} \varepsilon_j\right)^2 = \frac{1}{h} \quad \forall i = 1, \dots, p\,.$$

Therefore Proposition 5.6 yields

$$\mathcal{R}_\theta(\hat\theta^{(h)}) = \frac{1}{h} + \frac{1}{p}\|\theta - \bar\theta^{(h)}\|^2\,.$$

To evaluate the second term, we claim that

$$\frac{1}{h}\sum_{i \in I_\ell} (\bar\theta_i^{(h)} - \theta_i)^2 \le \frac{(Lh)^2}{p^2} \quad \forall \ell = 1, \dots, p/h\,.$$

Indeed, this follows from the fact that the definition of $\Theta_{\mathrm{Lip}}(L)$ implies that $|\max_{i\in I_\ell}\theta_i - \min_{j\in I_\ell}\theta_j| \le \frac{Lh}{p}$.

Therefore

$$\frac{1}{p}\|\theta - \bar{\theta}^{(h)}\|^2 = \frac{h}{p}\sum_{\ell=1}^{p/h}\frac{1}{h}\sum_{i\in I_\ell}(\bar{\theta}_i^{(h)} - \theta_i)^2 \le \frac{(Lh)^2}{p^2}\,.$$

We obtain

$$\mathcal{R}_\theta(\hat{\theta}^{(h)}) \le \frac{1}{h} + \frac{(Lh)^2}{p^2}\,.$$

Note the presence of a bias-variance trade off: as h increases, the variance decreases, but the bias increases. To minimize the risk, we choose $h \asymp \min\{(p/L)^{2/3}, p\}$, yielding

$$\mathcal{R}_\theta(\hat{\theta}^{(h)}) \lesssim \max\left\{\frac{L^{2/3}}{p^{2/3}}, \frac{1}{p}\right\}$$

□

Note that we have given up on doing well for θ that do not satisfy the smoothness assumption. But in return, we have turned an impossible problem into one we can solve.

7.4 Exercises

1. Suppose we augment the linear model model by replacing the covariate vectors $X_i \in \mathbb{R}^p$ by an augmented vectors $\tilde{X}_i \in \mathbb{R}^{p+1}$ defined by

$$(\tilde{X}_i)_0 = 1$$
$$(\tilde{X}_i)_j = (X_i)_j \quad \forall j = 1, \ldots, p\,.$$

   Show that the estimator obtained by solving

$$\operatorname*{argmin}_{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n(\beta_0 + \beta^\top X_i - Y_i)^2$$

   agrees (in a sense you should specify) with the predictor obtained by solving

$$\operatorname*{argmin}_{\tilde{\beta}\in\mathbb{R}^{p+1}} \frac{1}{n}\sum_{i=1}^n(\tilde{\beta}^\top \tilde{X}_i - Y_i)^2\,.$$

   Conclude, upon replacing p by $p+1$, we can always reduce to the no intercept model in (7.1) without loss of generality.

2. This exercise will show that the risk of $\hat{\beta}$ in the model (7.1) typically scales like $\sigma^2 p/n$.

   (a) Show that $\mathrm{tr}(\hat{\Sigma}_X^{-1})\,\mathrm{tr}(\hat{\Sigma}_X) \ge p^2$. (Hint: if $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\hat{\Sigma}_X$, then $\mathrm{tr}(\hat{\Sigma}_X) = \sum_{i=1}^p \lambda_i$ and $\mathrm{tr}(\hat{\Sigma}_X^{-1}) = \sum_{i=1}^p \lambda_i^{-1}$. It may be helpful to use the Cauchy–Schwarz inequality: for any $x_1, \ldots, x_p$ and $y_1, \ldots, y_p$, $\sum_{i=1}^p x_i y_i \le \left(\sum_{i=1}^p x_i^2\right)^{1/2}\left(\sum_{i=1}^p y_i^2\right)^{1/2}$.)

   (b) Show that $\mathrm{tr}(\hat{\Sigma}_X) = \frac{1}{n}\sum_{i=1}^n \|X_i\|^2$.

   (c) Conclude that if the covariate vectors are normalized such that each entry of X_i is of magnitude at most 1, then the variance of $\hat{\beta}$ is at least $\sigma^2 p/n$.

3. Ridge regression is an example of what is known as Tikhonov regularization, which is a generic regularization technique given by

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, \theta) + \alpha R(\theta) \,,$$

where $R : \Theta \to \mathbb{R}$ is a quadratic function.

(a) Show that the estimator $\bar{X}_\lambda$ defined in Example 7.1 minimizes

$$p \mapsto \frac{1}{n} \sum_{i=1}^{n} (X_i - p)^2 + \alpha \left(\frac{1}{2} - p \right)^2$$

with $\alpha = \frac{\lambda}{1-\lambda}$.

(b) Fix a parametric model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ where $\mathbb{E}_\theta X = \theta$. Consider the estimator

$$\hat{\theta} = \operatorname*{argmin}_{y \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (X_i - y)^2 + \alpha y^2 \,.$$

Show that for any $\alpha > 0$, $\hat{\theta}$ is a biased estimator. However, show that if $0 < \mathrm{Var}_\theta(X) < \infty$ for all $\theta \in \Theta$, then $\hat{\theta}$ has smaller variance than the sample average $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ for all $\theta \in \Theta$. This shows that, depending on the assumptions on Θ, the estimator $\hat{\theta}$ may be better than $\bar{X}$ under the square loss.

4. Suppose that instead of Lipschitz smoothness in the Gaussian sequence model, we consider *Hölder smoothness*, that is, we consider the set

$$\Theta(\beta, L) \coloneqq \{\theta \in \mathbb{R}^p : |\theta_i - \theta_j| \le L |i - j|^\beta / p^\beta \quad \forall i, j = 1, \dots, p\}$$

for some $\beta \in (0, 1]$ and $L \ge 0$.

(a) Show that $\Theta(1, L) = \Theta_{\mathrm{Lip}}(L)$.

(b) Show that if $\beta \le \beta'$, then

$$\Theta(\beta, L) \supseteq \Theta(\beta', L).$$

That is, the assumption that $\theta \in \Theta(\beta, L)$ becomes stronger as β increases. In light of this fact, do you expect estimation over the class $\Theta(\beta, L)$ for $\beta < 1$ to be easier or harder than estimation in the class $\Theta_{\mathrm{Lip}}(L)$?

(c) Show that if $\theta \in \Theta(\beta, L)$, the the bias of the estimator $\hat{\theta}^{(h)}$ is at most $L^2 (h/p)^{2\beta}$.

(d) Assuming for simplicity that $L = 1$, conclude that there exists an estimator in the model (7.4) which achieves

$$\sup_{\theta \in \Theta(\beta, L)} \mathcal{R}_\theta(\hat{\theta}) \lesssim p^{-\frac{2\beta}{1+2\beta}} \,.$$

Compare this with the rate obtained in Proposition 7.3.

Chapter 8

Monte-Carlo methods

8.1 Monte Carlo p-values

Another means of obtaining a good test can be used when it is possible to efficiently generate samples from the null distribution P_0. Though these tests require computational resources, they are *exact*—meaning that they produce valid p-values without asymptotic assumptions. An interesting feature of these tests are that they are randomized, meaning that they require the statistician to produce their own random variables as part of the testing procedure. They are useful for goodness-of-fit testing, which is a test with hypotheses of the form

$$\begin{aligned} \mathrm{H}_0 &: \mathbb{P} = \mathbb{P}_0 \\ \mathrm{H}_1 &: \mathbb{P} \neq \mathbb{P}_0 \,. \end{aligned}$$

Example 8.1. The website `random.org` claims to generate sequences of independent random bits. Suppose we wish to test this claim. Specifically, we observe a sequence $S \in \{0,1\}^n$, and wish to perform a goodness-of-fit test to determine whether the entries of S are i.i.d. Bern(1/2) random variables. Here is an output just generated from `random.org`:

$$S = (1,0,0,1,0,1,0,0,1,1,0,1,0,0,1,1,1,1,1,1) \,.$$

A natural test statistic is L_n, the length of the longest consecutive run of digits. In the example S above, $L_n = 6$.

However, to design a test, we need to know the distribution of L_n under the null, and in fact this is not known exactly. A natural thing to try is to generate some sequences of truly Bern(1/2) random variables (e.g., by flipping coins we know to be fair), compute L_n on these sequences, and see how the value of 6 compares.

The above idea is an example of a Monte Carlo method, and it can be used to produce valid p values.

Proposition 8.2. *Let T be any statistic, and let $T_1, \ldots, T_m$ be i.i.d copies of T under $\mathbb{P}_0$. Then the quantity*

$$\frac{|\{i : T_i \geq T(\omega)\}| + 1}{m+1}$$

is a valid p value.

Proof. We need to show that

$$\mathbb{P}_0 \left\{ \frac{|\{i : T_i \geq T\}| + 1}{m+1} \leq u \right\} \leq u \quad \forall u \in [0,1] \,.$$

Since the $\frac{|\{i:T_i \geq t\}|+1}{m+1}$ only takes the values $0, 1/(m+1), \ldots, 1$, it is enough to check this when $u = k/(m+1)$ for some integer k. Equivalently, we must show that for any $k = 0, \ldots, m+1$, we have

$$\mathbb{P}_0 \left\{ |\{i : T_i \geq T\}| \leq k-1 \right\} \leq \frac{k}{m+1} \,.$$

The key point is that, if T is indeed distributed according to $\mathbb{P}_0$, then the $m+1$ random variables $\{T_1, \ldots, T_m, T\}$ are i.i.d. Let us write $S_{(1)} \leq \cdots \leq S_{(m+1)}$ for the order statistics of these $m+1$ random variables. The event $\{|\{i : T_i \geq T\}| \leq k-1\}$ corresponds precisely to the situation when $T > S_{(m+1-k)}$. But since $\{T_1, \ldots, T_m, T\}$ are i.i.d. when T is distributed according to $\mathbb{P}_0$, it holds that

$$\mathbb{P}_0 \left\{ T > S_{(m+1-k)} \right\} = \mathbb{P}_0 \left\{ T_i > S_{(m+1-k)} \right\} \quad \forall i = 1, \ldots, m \,.$$

Therefore

$$\begin{aligned}
\mathbb{P}_0 \left\{ T > S_{(m+1-k)} \right\} &= \frac{1}{m+1} \left(\mathbb{P} \left\{ T > S_{(m+1-k)} \right\} + \sum_{i=1}^{m} \mathbb{P} \left\{ T_i > S_{(m+1-k)} \right\} \right) \\
&= \mathbb{E}[(m+1)^{-1} (\mathbb{1}_{T > S_{(m+1-k)}} + \sum_{i=1}^{m} \mathbb{1}_{T_i > S_{(m+1-k)}})] \\
&\leq \frac{k}{m+1} \,,
\end{aligned}$$

where we have used the fact that at most k elements of the set $\{T_1, \ldots, T_m, T\}$ are greater than $S_{(m-k+1)}$. (There can be fewer than k such elements if there are ties.) □

8.2 Permutation tests

Another exact method is the permutation test. This is useful in the situation where we wish to do a harder task than goodness-of-fit testing known as *two-sample testing*. Here, we have random variables $X_1, \ldots, X_n \sim P$ and $Y_1, \ldots, Y_m \sim Q$, and we wish to test the hypotheses

$$\begin{aligned}
\mathrm{H}_0 &: P = Q \\
\mathrm{H}_1 &: P \neq Q \,.
\end{aligned}$$

Note that this is harder than goodness of fit testing, because we have not specified the distributions P and Q under the null—we merely ask whether they are equal or not. Suppose we choose some statistic $T(X_1, \ldots, X_n; Y_1, \ldots, Y_m)$ which we design to be large if $X_1, \ldots, X_n$ are very different from $Y_1, \ldots, Y_m$; perhaps it's the difference in their sample means:

$$T(X_1, \ldots, X_n; Y_1, \ldots, Y_m) = \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{m} \sum_{i=1}^{m} Y_i \right| .$$

The idea of the permutation test is to note again, crucially, that under H_0, the random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are all i.i.d. Let us concatenate the lists and rename them $Z_1, \ldots, Z_{n+m}$. If π is any permutation, then under the null distribution $Z_{\pi(1)}, \ldots, Z_{\pi(n+m)}$ has the same distribution as $Z_1, \ldots, Z_{n+m}$. This leads to the following idea: let $\pi_1, \ldots, \pi_N$ be random permutations of the set $\{1, \ldots, m+n\}$, and consider the statistic

$$\frac{|\{i : T(Z_{\pi_i(1)}, \ldots, Z_{\pi_i(n+m)}) \geq T(\omega)\}| + 1}{N+1} \,. \tag{8.1}$$

Proposition 8.3. *The quantity in* (8.1) *is a valid p value.*

You will show this on your homework.

In fact, it is possible to show that the expression in (8.1) is a valid p value even if we condition on the ordered list of observed quantities. This makes the permutation test an example of a *conditional inference* procedure.

8.3 Rejection sampling

In discussing Monte Carlo p values, we have made the assumption that it is easy to sample from the null distribution. How do we accomplish this on a computer?

We begin with a simple setting. Suppose we wish to sample uniformly from a compact set $S \subseteq [0,1]^d$. Consider the following procedure: we take samples uniformly from $[0,1]^d$ one at a time until we get one that lands in S, and then return that last sample.

Proposition 8.4. *The above procedure yields a sample from the uniform distribution on S.*

Proof. Denote the sample obtained by the above procedure by Y. We need to show that for any subset $A \subseteq S$, the probability that $Y \in A$ is equal to $|A|/|S|$, where $|\cdot|$ denotes the volume (Lebesgue measure). Letting X be uniformly distributed on $[0,1]^d$, we note that

$$Y \stackrel{d}{=} X \mid X \in S\,.$$

Therefore

$$\mathbb{P}\{Y \in A\} = \mathbb{P}\{X \in A \mid X \in S\} = \frac{\mathbb{P}\{X \in A \cap S\}}{\mathbb{P}\{X \in S\}} = \frac{\mathbb{P}\{X \in A\}}{\mathbb{P}\{X \in S\}} = \frac{|A|}{|S|}\,,$$

where we have used the assumption that X is uniformly distributed on $[0,1]^d$. □

This is known as *rejection sampling*: we produce *proposals* from an easy-to-sample from distribution, and reject those that do not fit our target distribution.

This procedure can be extended to settings where a non-uniform distribution is desired. Suppose that we can sample easily from a continuous distribution with density p_0 on $\mathbb{R}^d$, and would like to produce samples from a distribution with density p. Let

$$c = \sup_{x \in \mathbb{R}^d} \frac{p(x)}{p_0(x)}\,.$$

We perform the following procedure:

- Sample $X \sim p_0$ and $U \sim \mathrm{Unif}([0,1])$ independently.
- If $U \leq p(X)/cp_0(X)$, then return X. Otherwise, continue.

Proposition 8.5. *The above procedure yields a sample from p.*

Proof. Write E for the event $U \leq p(X)/cp_0(X)$. As above, if we let Y be the output fo the procedure, then

$$Y \stackrel{d}{=} X \mid E\,.$$

For any set $A \subseteq \mathbb{R}^d$, we therefore have

$$\mathbb{P}\{Y \in A\} = \mathbb{P}\{X \in A \mid E\} = \frac{\mathbb{P}\{(X \in A) \text{ and } E\}}{\mathbb{P}\{E\}} = \frac{\int_A \mathbb{P}\{E \mid X = x\}\, p_0(x)\,\mathrm{d}x}{\mathbb{P}\{E\}}\,.$$

For each $x \in \mathbb{R}^d$, we have

$$\mathbb{P}\{E \mid X = x\} = \mathbb{P}\{U \leq p(x)/cp_0(x)\} = p(x)/cp_0(x)\,,$$

so that the numerator is

$$\int_A \frac{p(x)}{cp_0(x)} p_0(x)\,\mathrm{d}x = \frac{1}{c}\int_A p(x)\,\mathrm{d}x$$

By the same reasoning, the denominator is $\int \mathbb{P}\{E \mid X = x\}\, p_0(x)\,\mathrm{d}x = \int \frac{p(x)}{cp_0(x)} p_0(x)\,\mathrm{d}x = \frac{1}{c}$. Therefore the probability is $\int_A p(x)\,\mathrm{d}x$, so that the distribution of Y has density p, as desired. □

Rejection sampling is sensible in low dimension, but can scale poorly in high dimension.

Example 8.6. Suppose we wish to sample uniformly from the region $S = \{x \in [0,1]^d : \sum_{i=1}^d x_i \leq 1\}$. It is possible to show that the volume of this region is $1/d!$, which is exponentially small in d. Therefore, we are *exponentially* unlikely to produce a sample from S when we take a sample from $[0,1]^d$, so producing even a single sample will require waiting a very long time.

8.4 Exercises

1. Above, we claimed that two-sample testing is harder than goodness-of-fit testing. Prove this, by showing that any two-sample test with level α can be adapted to obtain a level α randomized goodness-of-fit test for any null distribution P_0 from which it is possible to generate i.i.d. samples.

2. Let $\pi_1, \ldots, \pi_N$ be independent random permutations of the set $\{1, \ldots, n+m\}$, selected uniformly from the set of all such permutations. Mimicking the proof of Proposition 8.2, show that (8.1) is a valid p-value. (Hint: write $T = T(Z_1, \ldots, Z_{n+m})$ and $T_i = T(Z_{\pi_i(1)}, \ldots, Z_{\pi_i(n+m)})$ for $i = 1, \ldots, N$. While it is not true that the random variables $\{T_1, \ldots, T_m, T\}$ are independent, they are *exchangeable*, meaning that their joint distribution is the same as the joint distribution of $\{T'_1, \ldots, T'_{m+1}\}$, where $T'_1, \ldots, T'_{m+1}$ is any reordering of $\{T_1, \ldots, T_m, T\}$.)

3. Show how to extend Proposition 8.4 to the setting where S does not lie in $[0,1]^d$. (Assume that $|S| < \infty$.)

4. Show that the probability that we accept a sample in the setting of Proposition 8.5 is $1/c$. Conclude that the expected number of samples from p_0 required to obtain a single sample from p is c.

Chapter 9

Model selection and cross validation

9.1 The basic problem

Throughout this course, we have assumed from the outset that we have a statistical model in mind for a problem at hand. However, it is usually the case that the statistician has a *choice* of which model to use, and this choice can affect performance substantially. How to choose the best option? One way to think about this problem is that it is another example of the bias-variance tradeoff we have seen many times in the course: if we adopt a more complicated model, then our estimators will typically be less biased, but have greater variance. The key question is how to optimally balance these two concerns.

Example 9.1. Let's begin with a familiar idea: linear regression. Recall that our basic linear model reads

$$Y_i = \beta^\top x_i + \varepsilon_i \,,$$

where ε_i is uncorrelated noise. The vector x_i contains covariates, which we hope will help us predict Y_i. However, in practice we have the choice of which covariates to include.

Let us consider two models:

$$\begin{aligned} \mathcal{P} &: Y_i = \beta^\top x_i + \varepsilon_i \,, \qquad x_i \in \mathbb{R}^p \\ \mathcal{P}' &: Y_i = (\beta')^\top x_i' + \varepsilon_i \,, \qquad x_i' \in \mathbb{R}^{p'} \,, \end{aligned}$$

where $p' > p$ and the covariate vector x_i' consists in augmenting x_i with extra covariates. The second model is larger than the first, in the sense that if we force the extra coordinates of β' (those corresponding to covariates appearing in $\mathcal{P}'$ but not in $\mathcal{P}$) to be zero in the second model, then we recover the first model.

A modification of our argument in Proposition 7.2 shows that if ε_i has variance σ^2, then the ordinary least squares estimators in the two models satisfy

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathrm{Var}(\hat{\beta}^\top x_i) &= \frac{\sigma^2 p}{n} \\ \frac{1}{n} \sum_{i=1}^n \mathrm{Var}((\hat{\beta}')^\top x_i') &= \frac{\sigma^2 p'}{n} \,. \end{aligned}$$

Therefore, the variance of the predictions in the larger model is larger, *whether or not the extra covariates actually help in predicting* Y_i. How should we select which covariates to include?

Example 9.2. Polynomial regression is a generalization of linear regression. Given a covariate $X \in \mathbb{R}$, we model the optimal regression function as

$$r(x) = \mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p \,.$$

We need to choose the order of the polynomial. As above, we have a sequence of nested models: if $p' \geq p$, then the order-p' model includes the order-p model as a special case.

Example 9.3. A *time series* is a sequence of random variables $Y_1, Y_2, \ldots$ which are assumed to have some temporal structure. In this setting is silly to assume that the Y_i are i.i.d. A common model is the autoregressive model, which models interactions between the terms:

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_k Y_{t-k} + \varepsilon_t \,,$$

where ε_t is independent noise. Again, we need to choose the order k of the model.

Example 9.4. Suppose we have decided to use ridge regression. Then our estimator is given by

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (x_i^\top \beta - Y_i)^2 + \lambda \|\beta\|^2 \,,$$

How should we choose λ?

This is a model selection problem. Indeed the general theory of Lagrange multipliers indicates that this family of optimization problems is equivalent to the family

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^p, \|\beta\| \leq \tau} \frac{1}{n} \sum_{i=1}^{n} (x_i^\top \beta - Y_i)^2 \,.$$

We can therefore view the question of finding the best regularization parameter as equivalent to the question of selecting the best parameter space over which to optimize.

How should we select the error? A point to emphasize is that we cannot simply check which model gives the smallest error on our data. For instance, suppose that we observe data $Y_1, \ldots, Y_n$, and we have access to corresponding covariates $x_1, \ldots, x_n \in \mathbb{R}^p$ or their augmented versions $x_1', \ldots, x_n' \in \mathbb{R}^{p'}$. The ordinary least squares estimator using the p-dimensional covariates is

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (x_i^\top \beta - Y_i)^2 \tag{9.1}$$

and the same estimator using the p'-dimensional covariates is

$$\hat{\beta}' = \operatorname*{argmin}_{\beta' \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} ((x_i')^\top \beta' - Y_i)^2 \,. \tag{9.2}$$

We can compute both estimators on our data. A naïve approach to select one of them is to see which one of them minimizes the ℓ_2 loss (the "training" loss):

$$\frac{1}{n} \sum_{i=1}^{n} (x_i^\top \hat{\beta} - Y_i)^2 \overset{?}{\gtrless} \frac{1}{n} \sum_{i=1}^{n} ((x_i')^\top \hat{\beta}' - Y_i)^2 \,.$$

But it is obvious from (9.1) and (9.2) that the loss associated with the second estimator will always be smaller, because if we augment $\hat{\beta}$ by adding zeroes in coordinates corresponding to the additional covariates, then the resulting vector is feasible in (9.2), so the loss associated with $\hat{\beta}'$ is only smaller than this. Therefore this approach always selects the more complicated model.

The first lesson is **using the same data to compute the estimator and select the model is biased in favor of more complicated models.** Therefore the methods we propose will either use different data to compute the estimator and select the model (cross-validation) or penalize more complicated models to counteract the bias just described (AIC and BIC).

We have two goals in mind:

- Find the model that gives the best prediction (without assuming that any of the models are correct).
- Assuming that one of the models is actually true, find the true one.

As we will see, cross-validation and AIC work well for the first goal, but they do not work well for the second.

9.2 Cross validation and AIC

The most straightforward approach to model selection uses sample-splitting: we divide the data set into two groups (what in machine-learning speak are called "training sets" and "validation sets"), and use the first group to construct our estimators and the second group to select a model.

The mathematically cleanest formulation of cross-validation splits the data into two sets of equal size. However, in practice, it is more common to split the data into K groups, $K > 2$, and for $k = 1, \dots, K$, use all but the kth group for training, and use the kth group for validation. This formulation, though common, is much more difficult to analyze and the justification is not always rigorous.

We fix the following framework. Assume that we have k different models $\mathcal{P}_1, \dots, \mathcal{P}_k$ given by

$$\mathcal{P}_j = \{\mathrm{P}_\theta : \theta \in \Theta_j\},$$

for some parameter sets $\Theta_1, \dots, \Theta_k$. For example, in the context of the linear models described above, we have $\Theta = \mathbb{R}^p$ for the first model and $\Theta' = \mathbb{R}^{p'}$ for the second. We assume we have access to $2n$ i.i.d. samples, which we label $X_1, \dots, X_n$ and $X'_1, \dots, X'_n$. We will use the first set to construct our estimators, and the second to select which estimator is best.

We focus on maximum likelihood estimation. We can compute estimators

$$\hat{\theta}_j = \operatorname*{argmax}_{\theta \in \Theta_j} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta | X_i).$$

If the data come from P_{θ^*}, we recall that we can view this maximum likelihood procedure as minimizing

$$\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} \ell(\theta^* | X_i) - \ell(\theta | X_i),$$

which is the empirical counterpart of

$$\mathbb{E}[\ell(\theta^* | X_i) - \ell(\theta | X_i)] = \mathrm{D}(\mathrm{P}_{\theta^*} \,\|\, \mathrm{P}_\theta) = \int p_{\theta^*}(x) \log p_{\theta^*}(x) \, \mathrm{d}x - \int p_{\theta^*}(x) \log p_\theta(x) \, \mathrm{d}x.$$

We can therefore evaluate the quantity of the estimator $\hat{\theta}_j$ by estimating

$$R_j \coloneqq \int p_{\theta^*}(x) \log p_{\hat{\theta}_j}(x) \, \mathrm{d}x$$

As noted above, the naïve estimator (which we might call the "training error")

$$\frac{1}{n} \sum_{i=1}^{n} \ell(\hat{\theta}_j | X_i) \tag{9.3}$$

is badly biased, because the estimator $\hat{\theta}_j$ was constructed in order to minimize this criterion in the first place. In some special cases, such as for Gaussian models, it is possible to "de-bias" this estimator to obtain a more accurate assessment of the true risk. (You will show this on your homework.)

Cross validation simply constructs an unbiased estimator of R_j by using the validation set:

$$\hat{R}_j = \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{\theta}_j | X'_i).$$

Note that

$$\mathbb{E}[\hat{R}_j | X_1, \dots, X_n] = R_j.$$

We simply select the estimator for which $\hat{R}_j$ is the largest.

If we assume for simplicity that ℓ is bounded, then it is easy to show that cross-validation has good performance.

Proposition 9.5. *Suppose the log-likelihood satisfies $|\ell(\theta|x)| \le B$ for all θ and x. Let $\hat{j} = \operatorname{argmax}_j \hat{R}_j$ be the model with the best cross-validation performance, and let $j^* = \operatorname{argmax}_j R_j$ be the model with the best true performance. Then with probability at least $1-\delta$,*

$$R_{\hat{j}} \ge R_{j^*} - 2B\sqrt{\frac{2\log(2k/\delta)}{n}}\,.$$

Proof. The proof is the same as in Section 2.2.1. Since ℓ is bounded, we have by Proposition 2.2

$$\max_j |\hat{R}_j - R_j| \le B\sqrt{\frac{2\log(2k/\delta)}{n}}$$

with probability at least $1-\delta$. Using the definition of $\hat{j}$, we then obtain

$$R_{\hat{j}} \ge \hat{R}_{\hat{j}} - B\sqrt{\frac{2\log(2k/\delta)}{n}} \ge \hat{R}_{j^*} - B\sqrt{\frac{2\log(2k/\delta)}{n}} \ge R_{j^*} - 2B\sqrt{\frac{2\log(2k/\delta)}{n}}\,.$$

□

The Akaike Information Criterion (AIC) is an approximation for the above procedure, which doesn't require sample splitting. It is useful if the number of samples is too small for sample splitting to be viable, or if cross-validation is too computationally intensive. The AIC simply subtracts the bias of the naïve estimator to get an unbiased estimator of R_j.

To motivate it, we focus on the Gaussian linear model. Recall that the likelihood is

$$\ell(\beta|y,x) = -\frac{1}{2\sigma^2}(x^\top\beta - y)^2\,.$$

If we assume that $Y_i = x_i^\top\beta + \varepsilon_i$, where ε_i have variance σ^2, then given a candidate estimator $\hat{\beta}_j \in \mathbb{R}^p$, we have

$$R_j = -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i^\top\hat{\beta}_j - \mathbb{E}Y_i)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n \operatorname{Var}(Y_i) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i^\top\hat{\beta}_j - x_i^\top\beta^*)^2 - \frac{n}{2}\,.$$

On the other hand, if we simply compute the average log-likelihood of our estimator on our data, we get

$$-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i^\top\hat{\beta}_j - Y_i)^2 = -\frac{1}{2\sigma^2}\sum_{i=1}^n \left[(x_i^\top\hat{\beta}_j - x_i^\top\beta)^2 + \sigma^2 - 2(x_i^\top\hat{\beta}_j - x_i^\top\beta)\varepsilon_i\right].$$

The first two terms are exactly what we want to appear; the bias arises from the fact that the expectation of the third term is not zero. If we recall the explicit expression for the ordinary least squares estimator given in Proposition 7.2, we have that

$$(x_i^\top\hat{\beta}_j - x_i^\top\beta)\varepsilon_i = x_i^\top\Sigma_X^{-1}\left(\frac{1}{n}\sum_{j=1}^n x_j\varepsilon_j\right)\varepsilon_i\,,$$

where

$$\Sigma_X = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top\,.$$

Using the fact that the noise terms ε_i are i.i.d. mean-zero Gaussian random variables with variance σ^2, we obtain

$$\mathbb{E}(x_i^\top\hat{\beta}_j - x_i^\top\beta)\varepsilon_i = \frac{\sigma^2}{n}x_i^\top\Sigma_X^{-1}x_i\,.$$

We obtain that

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathbb{E}2(x_i^\top\hat{\beta}_j - x_i^\top\beta)\varepsilon_i = -\frac{1}{2n}\sum_{i=1}^{n}x_i^\top\Sigma_X^{-1}x_i = -p\,.$$

We obtain the second lesson: **in the Gaussian model, the bias of the training error is proportional to the dimension of the parameter space.**

If we remove this bias, we are led to the following model selection procedure: given candidate ordinary least squares estimators $\hat{\beta}_1, \ldots, \hat{\beta}_k$, lying in $\mathbb{R}^{p_1}, \ldots, \mathbb{R}^{p_k}$, select an estimator by solving

$$\operatorname*{argmax}_j \left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i^\top\hat{\beta}_j - Y_i)^2\right) - 2p_j\,.$$

Note that this proposal has the effect of penalizing models with higher complexity (larger dimension).

Akaike showed that the above calculations remain asymptotically valid in more general situations; therefore, he proposed the following general approach:

- For each model $\mathcal{P}_j$, compute the maximum likelihood estimator $\hat{\theta}_j \in \Theta_j$.
- Return the estimator $\hat{\theta}_{\hat{j}}$, where

$$\hat{\theta}_j = \operatorname*{argmax}_j 2\ell(\hat{\theta}_j) - 2\dim(\Theta_j)\,.$$

It can be shown that asymptotically, this proposal has essentially the same behavior as cross-validation.

9.3 Model selection consistency and BIC

So far, we have been focusing on finding a model which offers the best prediction accuracy. Now, we turn to the question of selecting a the "true" model, given that one is actually correct. Assuming that our models are nested, $\mathcal{P}_1 \subseteq \cdots \subseteq \mathcal{P}_k$, we can define the true model $\mathcal{P}_{j^*}$ as the smallest model which contains the distribution from which our data is drawn. An important third lesson, which we will demonstrate now, is that **cross-validation does not correctly identify the true model, even when $n \to \infty$.**

To see this, let us look at a very simple example. We observe i.i.d. samples $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$, and consider two different parameter spaces for θ:

$$\Theta_1 = \{0\}$$
$$\Theta_2 = \mathbb{R}\,.$$

The corresponding models are nested. Let us assume that $\theta = 0$, so that the first model is correct. Our goal is to estimate θ under the square loss.

These two models lead to two different maximum likelihood estimates

$$\hat{\theta}_1 = 0$$
$$\hat{\theta}_2 = \bar{X}\,.$$

First, note that, as observed above, if we simply consider training error, then we will always select the second model, since

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\theta}_1)^2 > \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\theta}_2)^2$$

with probability 1. Therefore using the training error alone always selects the wrong model.

Surprisingly, cross validation can also fail to select the right model, even when n is large. Let us assume that we have $2n$ samples, evenly split between training and validation. We write $\bar{X}$ and $\bar{X}'$ for the sample average of the two sets of samples. We select the second model whenever

$$\frac{1}{n}\sum_{i=1}^{n}(X_i' - \hat{\theta}_1)^2 > \frac{1}{n}\sum_{i=1}^{n}(X_i' - \hat{\theta}_2)^2\,,$$

or equivalently whenever

$$(\bar{X}')^2 > (\bar{X}' - \bar{X})^2\,.$$

This holds if and only if

$$(2\bar{X}' - \bar{X})\bar{X} > 0\,.$$

But note that, no matter how large n is, this happens with probability at least $1/4$. (To see this, note that $\sqrt{n}\bar{X}'$ and $\sqrt{n}\bar{X}$ are independent, mean zero Gaussians.) So, no matter how large n is, the probability that we make a mistake does not tend to 0.

AIC will also make mistakes in this setting. The two models correspond to 0 and 1 dimensional parameter spaces, respectively. Therefore, the AIC criterion selects the second model if

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\theta}_1)^2 > \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\theta}_2)^2 + \frac{2}{n}\,,$$

which happens if and only if

$$\bar{X}^2 > \frac{2}{n}\,.$$

Since $\sqrt{n}\bar{X}$ is a mean-zero standard Gaussian, this happens with probability that does not vanish in the $n \to \infty$ limit. We say that both cross-validation and AIC are inconsistent for model selection.

A fix for this problem is a different penalization approach, known as the Bayesian information criterion (BIC). The procedure is very similar to AIC:

- For each model $\mathcal{P}_j$, compute the maximum likelihood estimator $\hat{\theta}_j \in \Theta_j$.
- Return the estimator $\hat{\theta}_{\hat{j}}$, where

$$\hat{\theta}_j = \operatorname*{argmax}_j 2\ell(\hat{\theta}_j) - \log(n)\dim(\Theta_j)\,.$$

Note that BIC penalizes larger models more harshly than AIC does. Moreover, it can be shown that BIC is consistent for model selection. (I.e., it selects the correct model with probability approaching 1 as $n \to \infty$.) For example, in the above setup, BIC selects model 2 if and only if

$$\bar{X}^2 > \frac{\log n}{n}\,.$$

If $\theta = 0$, so that the first model is true, this event happens with probability tending to 0.

While giving a justification for BIC is outside the scope of this course, it can be justified by showing that it is an asymptotic approximation to a Bayesian approach which puts a prior distribution on the different models.

9.4 Structural risk minimization and oracle inequalities

The AIC and BIC methods were both motivated and justified by asymptotic considerations. There is a another penalization method with a slightly different flavor known as *structural risk minimization* which offers finite sample guarantees and is easier to apply in many machine learning contexts.

Let us recall the empirical risk minimization procedure for classification. We observe covariates and responses (X, Y) and we would like to minimize the classification error

$$R(h) \coloneqq \mathbb{P}\{h(X) \neq Y\}$$

To actually perform estimation when we have observed n i.i.d. samples (X_i, Y_i) for $i = 1, \dots n$, we can minimize the empirical risk

$$\hat{R}(h) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{h(X_i) \neq Y_i} .$$

Explicitly, for a class of candidate functions $\mathcal{H}$, we can consider the estimator

$$\hat{h} \coloneqq \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}(h) .$$

As we know, it will matter a lot which class of candidate functions we choose. (Recall from the exercises to Chapter 2 that if we let $\mathcal{H}$ be the set of all functions from $\mathbb{R}^p \to \{0, 1\}$, then $\hat{h}$ will never be a good classifier.) However, in Section 2.2.1 we argued that we could control the quality of this empirical risk minimizer by the inequality

$$R(\hat{h}) \leq \min_{h \in \mathcal{H}} R(h) + 2 \max_{h \in \mathcal{H}} |\hat{R}(h) - R(h)| .$$

Let us interpret these two terms. The first term will get smaller as $\mathcal{H}$ gets bigger. The second term will get bigger as $\mathcal{H}$ gets bigger, since it is harder to uniformly control deviations over larger classes. You can think of these two terms as representing a bias-variance tradeoff, just as we have seen in other aspects of model selection.

We have shown that, when $\mathcal{H}$ is finite, for any $\delta \in (0, 1)$,

$$\max_{h \in \mathcal{H}} |\hat{R}(h) - R(h)| \leq \sqrt{\frac{\log(|\mathcal{H}|) + \log(2/\delta)}{n}} \quad \text{with probability at least } 1 - \delta.$$

In fact, for many sets $\mathcal{H}$ of interest (even infinite ones) it possible to prove a bound that looks quite similar:

$$\max_{h \in \mathcal{H}} |\hat{R}(h) - R(h)| \leq \sqrt{\frac{c(\mathcal{H}) + \log(2/\delta)}{n}} \quad \text{with probability at least } 1 - \delta, \tag{9.4}$$

where $c_n(\mathcal{H})$ is some complexity measure of the class $\mathcal{H}$. When $\mathcal{H}$ is finite, we can take $c(\mathcal{H}) = \log |\mathcal{H}|$, but such a statement can hold for infinite classes as well. We saw one example of this when we used bracketing numbers to prove Theorem 2.8.

All in all, the standard bound for empirical risk minimization reads

$$R(\hat{h}) \leq \min_{h \in \mathcal{H}} R(h) + 2\sqrt{\frac{c(\mathcal{H}) + \log(2/\delta)}{n}} \qquad \text{with probability at least } 1 - \delta. \tag{9.5}$$

Everything which we have said so far holds for a single class $\mathcal{H}$. What happens when we move to a model selection context? Now, let's imagine that we have a sequence of nested classes, $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \dots$. This sequence may even be infinite. For each class, we can compute an empirical risk minimizer

$$\hat{h}_j \coloneqq \min_{h \in \mathcal{H}_j} \hat{R}(h) .$$

The model selection problem asks which estimator we should pick.

As we have seen, if we just pick the j for which $\hat{R}(\hat{h}_j)$ is the smallest, *we will always pick the largest (most complicated) class.* Motivated by the AIC and BIC procedures, we should attempt to penalize the larger classes in such a way that we select a model of the right size. Examining (9.5), we have for each class $\mathcal{H}_j$ that

$$R(\hat{h}_j) \leq \min_{h \in \mathcal{H}_j} R(h) + 2\sqrt{\frac{c(\mathcal{H}_j) + \log(2/\delta)}{n}} \qquad \text{with probability at least } 1 - \delta. \tag{9.6}$$

As we have seen, the second term represents the "price" that we pay for more complicated classes. This motivates the following idea, called *structural risk minimization* (SRM). Let $w_1, w_2, \dots$ be a sequence of positive numbers such that $\sum_{j=1}^{\infty} w_j \le 1$. For example, we can let $w_j = \frac{6}{\pi^2 j^2}$. The SRM procedure is:

1. Compute the empirical risk minimizer $\hat{h}_j$ for each class $\mathcal{H}_j$
2. Compute
$$\hat{j} \coloneqq \operatorname*{argmin}_{j \ge 1} \hat{R}(\hat{h}_j) + \sqrt{\frac{c(\mathcal{H}_j) + \log(2/\delta w_j)}{n}}\,.$$
3. Return $\hat{h}_{\hat{j}}$.

The SRM estimator enjoys the following bound.

Proposition 9.6. *For any $\delta \in (0,1)$, the SRM estimator defined above satisfies*

$$R(\hat{h}_{\hat{j}}) \le \min_{j \ge 1} \left\{ \min_{h \in \mathcal{H}_j} R(h) + 2\sqrt{\frac{c(\mathcal{H}_j) + \log(2/\delta w_j)}{n}} \right\} \tag{9.7}$$

with probability at least $1 - \delta$.

The bound (9.7) is known as an *oracle inequality*. The idea is that this estimator performs essentially as well as an "oracle" estimator that knows the best model to select in advance. Amazingly, up to the appearance of w_j in the logarithm, (9.7) shows that there is no price to pay for model selection in this context.

Proof. For convenience, let us write

$$\varepsilon_j \coloneqq \sqrt{\frac{c(\mathcal{H}_j) + \log(2/\delta w_j)}{n}}\,.$$

First, we claim that, with probability at least $1 - \delta$, simultaneously for all $j \ge 1$, we have

$$\max_{h \in \mathcal{H}_j} |R(h) - \hat{R}(h)| \le \varepsilon_j\,. \tag{9.8}$$

The proof is a simple application of the union bound. Indeed, we know by (9.4) that the probability that this inequality is violated for some fixed j is bounded above by $w_j \delta$. Therefore, by a union bound, the probability that it is violated for *any* j is bounded above by

$$\sum_{j \ge 1} w_j \delta \le \delta\,,$$

which proves the claim.

Now, let us restrict our attention to the event where (9.8) holds for all $j \ge 1$. Let us fix a j and compare the performance of $\hat{h}_{\hat{j}}$ with $h_j^* \coloneqq \operatorname{argmin}_{h \in \mathcal{H}_j} R(h)$. Our goal is to show

$$R(\hat{h}_{\hat{j}}) \le R(h_j^*) + 2\varepsilon_j\,.$$

We have

$$\begin{aligned} R(\hat{h}_{\hat{j}}) - R(h_j^*) &= R(\hat{h}_{\hat{j}}) - \hat{R}(\hat{h}_{\hat{j}}) + \hat{R}(\hat{h}_{\hat{j}}) - \hat{R}(h_j^*) + \hat{R}(h_j^*) - R(h_j^*) \\ &\le \varepsilon_{\hat{j}} + \hat{R}(\hat{h}_{\hat{j}}) - \hat{R}(h_j^*) + \varepsilon_j\,. \end{aligned}$$

But by definition of the SRM estimator,

$$\hat{R}(\hat{h}_{\hat{j}}) + \varepsilon_{\hat{j}} \le \hat{R}(\hat{h}_j) + \varepsilon_j \le \hat{R}(h_j^*) + \varepsilon_j\,.$$

We obtain

$$R(\hat{h}_{\hat{j}}) - R(h_j^*) \le 2\varepsilon_j\,,$$

as claimed. Since $j \ge 1$ was arbitrary, this proves the proposition. □

9.5 Exercises

1. This exercise will show that even though AIC is inconsistent for model selection, it can still lead to consistent estimates. Consider again the model where $X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$ i.i.d., with parameter spaces $\Theta_1 = \{0\}$ and $\Theta_2 = \mathbb{R}$, and define $\hat{\theta}_1$ and $\hat{\theta}_2$ to be the maximum-likelihood estimators over these spaces. Write $A_1 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\theta}_1)^2$ and $A_2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\theta}_2)^2 + \frac{2}{n}$. Consider the estimator

$$\hat{\theta} \coloneqq \begin{cases} \hat{\theta}_1 & \text{if } A_1 \le A_2, \\ \hat{\theta}_2 & \text{if } A_1 > A_2. \end{cases}$$

   (a) Show that if $\theta = 0$ (i.e., $\theta \in \Theta_1$), then $\hat{\theta} \xrightarrow{p} 0$ as $n \to \infty$.

   (b) Show that if $\theta \neq 0$ (i.e., $\theta \notin \Theta_1$), then $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.

   Conclude that AIC yields consistent estimators of θ, even if it is inconsistent for model selection.

2. This problem shows that performing inference procedures *after* model selection can be dangerous. Consider the linear model $Y = \beta^\top X + \varepsilon$, where $\beta \in \mathbb{R}^2$ is an unknown parameter and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. samples from this model.

   (a) Show that conditioned on $X_1, \dots, X_n$, the maximum likelihood estimator $\hat{\beta}$ satisfies

$$\hat{\beta} \sim \mathcal{N}(\beta, \frac{\sigma^2}{n} \cdot \hat{\Sigma}_X^{-1}).$$

   (Hint: recall that we can write $\hat{\beta} = \beta + \hat{\Sigma}_X^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i \varepsilon_i\right)$, and use the fact that ε_i are i.i.d. Gaussians.)

   (b) Assume for convenience that $\hat{\Sigma}_X = I$. Show that for $j = 1, 2$, the set $(\hat{\beta}_j - z_{\alpha/2}\sigma/\sqrt{n}, \hat{\beta}_j + z_{\alpha/2}\sigma/\sqrt{n})$ is a valid $1 - \alpha$ confidence set for β_j, where $z_{\alpha/2}$ is as usual the quantity satisfying $\mathbb{P}\left\{|Z| \ge z_{\alpha/2}\right\} = \alpha$ when $Z \sim \mathcal{N}(0, 1)$.

   (c) Suppose now that we attempt to do model selection only by retaining the *larger* of the two coordinates of the maximum likelihood estimator $\hat{\beta}$. That is, let $\hat{j} = \operatorname{argmax}_{j=1,2} |\hat{\beta}_j|$, where $\hat{\beta}$ is the maximum likelihood estimator. Is $(\hat{\beta}_{\hat{j}} - z_{\alpha/2}\sigma/\sqrt{n}, \hat{\beta}_{\hat{j}} + z_{\alpha/2}\sigma/\sqrt{n})$ a valid $1 - \alpha$ confidence set for $\beta_{\hat{j}}$? Why or why not? (Hint: if you are not sure, consider running a computer experiment to check!)

3. This exercise develops the *Stein Unbiased Risk Estimator*, which is a technique for obtaining an unbiased estimate of the risk for Gaussian models.

   (a) Let f be any differentiable function, and let $Z \sim \mathcal{N}(0, 1)$. Show that

$$\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)]$$

   by writing

$$\mathbb{E}[Zf(Z)] = \frac{1}{\sqrt{2\pi}} \int z f(z) e^{-z^2/2} \, \mathrm{d}z$$

   and integrating by parts.

   (b) Let f be any differentiable function, and let $X \sim \mathcal{N}(\theta, \sigma^2)$. Show that

$$\frac{1}{\sigma^2}\mathbb{E}[(X - \theta)f(X)] = \mathbb{E}[f'(X)].$$

   (Hint: define $Z = (X - \theta)/\sigma$, and write $f(X) = f(\sigma Z + \theta)$ to reduce to the previous case.)

(c) Let $Y = \theta + \sigma Z$, where $Z \sim \mathcal{N}(0,1)$ and $\theta \in \mathbb{R}$ is unknown. We assume for simplicity that σ is known. Show that for any estimator $\hat{\theta} = \hat{\theta}(Y)$ of θ, the mean square error of $\hat{\theta}$ satisfies

$$R_\theta(\hat{\theta}) = \mathbb{E}_\theta \|\theta - \hat{\theta}\|^2 = -\sigma^2 + \mathbb{E}\|Y - \hat{\theta}\|^2 + 2\mathbb{E}[(Y-\theta)\hat{\theta}]\,.$$

(d) Combining the results of the above two parts, show that

$$R_\theta(\hat{\theta}) = -\sigma^2 + \mathbb{E}\|Y - \hat{\theta}\|^2 + 2\sigma^2\mathbb{E}[\hat{\theta}'(Y)]\,.$$

Conclude that

$$\hat{R} \coloneqq -\sigma^2 + \|Y - \hat{\theta}\|^2 + 2\sigma^2\hat{\theta}'(Y)$$

is an unbiased estimator of the risk of the estimator $\hat{\theta}$.

Chapter 10

Non-parametric estimation

10.1 Kernel density estimation

Consider the set $\mathcal{P} = \{\mathrm{P} : \mathrm{P} \text{ is a continuous distribution on } \mathbb{R}\}$. In contrast to many of the models we have considered this semester, $\mathcal{P}$ is *non-parametric*: we cannot label the elements of $\mathcal{P}$ by a finite-dimensional parameter θ. In parametric contexts, we are used to being able to consistently estimate parameters of our model, but in non-parametric contexts this is not always true.

Example 10.1. Suppose that $X_1, \ldots, X_n \sim \mathrm{P}$ are i.i.d. samples from an unknown continuous distribution P with density f. Suppose we are interested in estimating $f(x_0)$—i.e., the value of f at a single, pre-specified point. It turns out the minimax risk of estimating this quantity is infinite. We will not give a rigorous justification for this fact, but an intuitive explanation is that I can make f extremely spiky at x_0 without affecting the distribution P at all, which makes it impossible to get any information about $f(x_0)$ from samples alone.

Note, however, that the Glivenko-Cantelli theorem (Theorem 2.8) shows that it *is* possible to estimate the CDF consistently. It is worth thinking about why the "spike" problem for estimating the density does not cause a problem when estimating the CDF.

Is there any non-parametric model for which we can consistently estimate $f(x_0)$? The spike problem indicates a possible remedy: if we restrict our attention to the class of continuous distributions whose densities are sufficiently smooth (e.g., Lipschitz continuous), then we will be able to perform consistent estimation.[1]

We will focus on a procedure known as kernel density estimation.[2] The idea is to plot the observed samples on a line and to *smooth* them so that they look like a density.

Concretely, let $K : \mathbb{R} \to \mathbb{R}$ be a nonnegative function satisfying

$$\int_{\mathbb{R}} K(x)\,\mathrm{d}x = 1\,, \quad \int_{\mathbb{R}} xK(x)\,\mathrm{d}x = 0\,.$$

In words, these requirements express the fact that K should be a centered bump, normalized to have unit mass. Given such a function, a positive *bandwidth* h, and a point $x \in \mathbb{R}$, the function

$$\frac{1}{h} K\left(\frac{\cdot - x}{h}\right)$$

is a rescaled version of K centered at x. When the bandwidth is small, this function is narrow and tall; when the bandwidth is large it is short and wide.

[1] This is the same situation we encountered when discussing the Gaussian sequence model.

[2] A warning that these "kernels" are different from the "kernels" discussed in machine learning in the context of RKHS's.

Given i.i.d. samples $X_1, \dots, X_n$ from an unknown continuous distribution, we will use as our estimate of the density a sum of scaled bumps centered at the observed data:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1} K\left(\frac{x - X_i}{h}\right) .$$

The key question, of course, is how to choose h. Once again, there is a bias-variance trade-off. It turns out that when h is properly chosen, this estimator achieves a small *mean integrated square error*:

$$\mathbb{E} \int (\hat{f}(x) - f(x))^2 \, \mathrm{d}x \, .$$

Proposition 10.2. *Suppose that $X_1, \dots, X_n \sim \mathrm{P}$, where P has twice-differentiable density f. Then*

$$\mathbb{E} \int (\hat{f}(x) - f(x))^2 \, \mathrm{d}x = O(\frac{1}{nh}) + O(h^4) \, ,$$

where the implicit constants depend on $\int f''(x)^2 \, \mathrm{d}x$, $\int K^2(x) \, \mathrm{d}x$, and $\int x^2 K^2(x) \, \mathrm{d}x$. In particular, choosing the optimal bandwidth $h \asymp n^{-1/5}$ yields

$$\mathbb{E} \int (\hat{f}(x) - f(x))^2 \, \mathrm{d}x \lesssim n^{-4/5} \, .$$

It can be shown that the rate $n^{-4/5}$ is unimprovable.

Proof. Interchanging the order of expectation and using a bias-variance decomposition, we obtain

$$\mathbb{E} \int (\hat{f}(x) - f(x))^2 \, \mathrm{d}x = \int \mathbb{E}(\hat{f}(x) - f(x))^2 \, \mathrm{d}x = \int \mathrm{Var}(\hat{f}(x)) \, \mathrm{d}x + \int (\mathbb{E}\hat{f}(x) - f(x))^2 \, \mathrm{d}x \, .$$

For the first term, the fact that $X_1, \dots, X_n$ are i.i.d. implies

$$\begin{aligned} \mathrm{Var}(\hat{f}(x)) &= \frac{1}{nh^2} \mathrm{Var}\left(K\left(\frac{x - X}{h}\right)\right) \\ &\le \frac{1}{nh^2} \mathbb{E} K\left(\frac{x - X}{h}\right)^2 && X \sim \mathrm{P} \\ &= \frac{1}{nh^2} \int K\left(\frac{x - y}{h}\right)^2 f(y) \, \mathrm{d}y \, , \end{aligned}$$

Therefore

$$\int \mathrm{Var}(\hat{f}(x)) \, \mathrm{d}x \le \frac{1}{nh^2} \int \int K\left(\frac{x - y}{h}\right)^2 f(y) \, \mathrm{d}y \, \mathrm{d}x \, .$$

Interchanging the order of integration and performing a change of variables, we have for all $y \in \mathbb{R}$,

$$\int K\left(\frac{x - y}{h}\right)^2 \mathrm{d}x = h \int K(u)^2 \, \mathrm{d}u \, .$$

We obtain

$$\int \mathrm{Var}(\hat{f}(x)) \, \mathrm{d}x \le \frac{1}{nh} \int f(y) \, \mathrm{d}y \int K(u)^2 \, \mathrm{d}u = \frac{\int K(u)^2}{nh} \, .$$

We see that the variance increases as h gets smaller, which makes sense because in that case the estimator is more spikey.

To evaluate the bias, we perform a Taylor expansion. Taylor's theorem with remainder implies for any $t \in \mathbb{R}$ that

$$f(x + t) = f(x) + t f'(x) + t^2 \int_0^1 f''(x + ts)(1 - s) \, \mathrm{d}s \, .$$

We obtain

$$\begin{aligned}
\mathbb{E}\hat{f}(x) &= \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)\,\mathrm{d}y \\
&= \int K(u) f(x-uh)\,\mathrm{d}u \\
&= \int K(u)(f(x) - uhf'(x) + (uh)^2 \int_0^1 f''(x-uhs)(1-s)\,\mathrm{d}s)\,\mathrm{d}u \\
&= f(x) + h^2 \int\int_0^1 K(u)u^2 f''(x-uhs)(1-s)\,\mathrm{d}s\,.
\end{aligned}$$

Therefore

$$\begin{aligned}
(\mathbb{E}\hat{f}(x) - f(x))^2 &= h^4 \left(\int\int_0^1 K(u)u^2 f''(x-uhs)(1-s)\,\mathrm{d}s\,\mathrm{d}u\right)^2 \\
&\leq h^4 \left(\int u^2 K(u)\,\mathrm{d}u\right)\left(\int\int u^2 K(u) f''(x-uhs)^2(1-s)^2\,\mathrm{d}s\,\mathrm{d}u\right),
\end{aligned}$$

where the inequality uses the Cauchy-Schwarz inequality.[3] Integrating with respect to x we obtain

$$\begin{aligned}
\int (\mathbb{E}\hat{f}(x) - f(x))^2\,\mathrm{d}x &\leq h^4 \left(\int u^2 K(u)\,\mathrm{d}u\right)\left(\int\int u^2 K(u) \int f''(x-uhs)^2\,\mathrm{d}x(1-s)^2\,\mathrm{d}s\,\mathrm{d}u\right) \\
&= \frac{1}{3} h^4 \left(\int u^2 K(u)\,\mathrm{d}u\right)^2 \int f''(x)^2\,\mathrm{d}x\,.
\end{aligned}$$

Note that the bias becomes smaller when the bandwidth gets smaller, which reflects the fact that we are blurring the function less.

This proves the claim. □

Choosing the correct bandwidth is an *extremely* important question in practice. A good approach is to use leave-one-out cross validation.

10.2 Kernel regression estimators

Kernel methods can also be used in the regression context. Suppose again that we have covariate-response pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, and we wish to predict the response on the basis of the covariate. We have shown that the optimal estimator for the squared loss is the optimal regression function

$$r(x) = \mathbb{E}[Y \mid X = x]\,.$$

Previously, we adopted a parametric model for r, (e.g., the linear model $r(x) = \beta_0 + \beta x$). However, we can also adopt a nonparametric approach. If f is the joint density of x and y, then the definition of conditional expectation implies

$$r(x) = \frac{\int y f(x,y)\,\mathrm{d}y}{\int f(x,y)\,\mathrm{d}y}\,.$$

[3]The form we use states for any two functions g and h

$$\left|\int\int_0^1 g(u,s)h(u,s)\,\mathrm{d}s\,\mathrm{d}u\right|^2 \leq \left|\int\int_0^1 g(u,s)^2\,\mathrm{d}s\,\mathrm{d}u\right| \left|\int\int_0^1 h(u,s)^2\,\mathrm{d}s\,\mathrm{d}u\right|.$$

We have chosen $g(u,s) \coloneqq uK(u)^{1/2}$ and $h(u,s) \coloneqq uK(u)^{1/2} f''(x-uhs)(1-s)$.

A natural estimator for the joint density is the kernel density estimator:

$$\hat{f}(x,y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h}\right).$$

(This is the natural bivariate extension of the univariate KDE derived above.) If we assume that K satisfies $\int K(x)\,\mathrm{d}x = 1$ and $\int xK(x)\,\mathrm{d}x = 0$, we obtain the *Nadaraya-Watson estimator*

$$\begin{aligned}
\hat{r}(x) &= \frac{\int y\hat{f}(x,y)\,\mathrm{d}y}{\int \hat{f}(x,y)\,\mathrm{d}y} \\
&= \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \int yK\left(\frac{y-Y_i}{h}\right)\,\mathrm{d}y}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \int K\left(\frac{y-Y_i}{h}\right)\,\mathrm{d}y} \\
&= \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}.
\end{aligned}$$

Writing this more succinctly, we have

$$\hat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i\,,$$

where

$$w_i(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)}.$$

Note that assuming K is nonnegative, these weights are nonnegative and sum to 1. In other words, this estimator for $r(x)$ is nothing but a weighted average of the Y_i, where the weight according to the proximity of x to the sample points X_i.

To prove a bound on the quality of this estimator, we make the further assumptions that K is compactly supported, i.e. there exists an M such that $K(x) = 0$ if $|x| \geq M$, and that K is bounded, i.e., that there exists a B such that $K(x) \leq B$ for all x. In particular, this implies that

$$|x - X_i| \geq Mh \implies w_i(x) = 0$$

and

$$w_i(x) \leq \frac{B}{nh} \cdot \frac{1}{\frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)}.$$

Under this assumption, we can show that $\hat{r}(x)$ is a good estimator at any fixed point as long as r is Lipschitz.

Proposition 10.3. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and satisfy the relation*

$$Y_i = r(X_i) + \varepsilon_i\,,$$

where $\mathbb{E}[\varepsilon_i | X_i] = 0$ and $\mathbb{E}[\varepsilon_i^2 | X_i] \leq \sigma^2$.

For any $x_0 \in \mathbb{R}$, if r is L-Lipschitz then

$$\mathbb{E}(\hat{r}(x_0) - r(x_0))^2 \leq \frac{B\sigma^2}{nh} \cdot \mathbb{E}\left[\frac{1}{\frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x_0-X_j}{h}\right)}\right] + (LMh)^2\,.$$

In particular, with the optimal choice $h \asymp n^{-1/3}$, we obtain

$$\mathbb{E}(\hat{r}(x_0) - r(x_0))^2 \lesssim n^{-2/3}\,.$$

To get intuition for the quantity $\mathbb{E}\left[\frac{1}{\frac{1}{nh}\sum_{j=1}^n K\left(\frac{x_0-X_j}{h}\right)}\right]$, we note that the average appearing in the denominator will be a consistent estimator for $f(x_0)$, where f is the density of the covariates, so we can expect this term to be of moderate size in places where the density is large, and large in places where the density is small. (Can you see intuitively why this behavior should be expected?)

Proof. We will conduct the analysis conditioned on the covariates $X_1, \ldots, X_n$, and show

$$\mathbb{E}[(\hat{r}(x_0) - r(x_0))^2 | X_1, \ldots, X_n] \leq \frac{B\sigma^2}{nh} \cdot \frac{1}{\frac{1}{nh}\sum_{j=1}^n K\left(\frac{x_0-X_j}{h}\right)} + (MLh)^2 \,.$$

As usual, we first employ a bias-variance decomposition:

$$\mathbb{E}[(\hat{r}(x_0) - r(x_0))^2 | X_1, \ldots, X_n] = (\mathbb{E}[\hat{r}(x_0) | X_1, \ldots, X_n] - r(x_0))^2 + \mathrm{Var}(\hat{r}(x_0) \mid X_1, \ldots, X_n) \,.$$

To evaluate the variance, note since $Y_1, \ldots, Y_n$ are conditionally independent, we have

$$\mathrm{Var}(\hat{r}(x_0) \mid X_1, \ldots, X_n) = \sum_{i=1}^n \mathrm{Var}(Y_i \mid X_i) w_i(x_0)^2 \leq \sigma^2 \sum_{i=1}^n w_i(x_0)^2 \,.$$

Since the weights w_i are nonnegative and sum to one, we have

$$\sigma^2 \sum_{i=1}^n w_i(x_0)^2 \leq \sigma^2 \max_j w_j(x_0) \sum_{i=1}^n w_i(x_0) \leq \frac{B\sigma^2}{nh} \cdot \frac{1}{\frac{1}{nh}\sum_{j=1}^n K\left(\frac{x_0-X_j}{h}\right)} \,.$$

This proves the desired bound on the variance.

For the bias, we have

$$\begin{aligned}
\mathbb{E}[\hat{r}(x) - r(x) | X_1, \ldots, X_n]] &= \sum_{i=1}^n (r(X_i) - r(x)) w_i(x) \\
&\leq \sum_{i=1}^n L|X_i - x| w_i(x) \\
&\leq LMh \sum_{i=1}^n w_i(x) = LMh \,,
\end{aligned}$$

where we have used the assumptions that r is Lipschitz and that w_i vanishes when $|X_i - x| \geq Mh$. This proves the desired bound on the bias, and the claim. □

10.3 Local linear regression

We now give a different interpretation of the Nadaraya-Watson estimator. We have shown that it can be written

$$\hat{r}(x) = \sum_{i=1}^n Y_i w_i(x) \,,$$

where w_i are weights depending on the covariates. As noted above, this means that $\hat{r}(x)$ is nothing but a weighted of the responses. Specifically, for a fixed $x \in \mathbb{R}$, the estimator agrees with the solution to

$$\operatorname*{argmin}_{\beta_0} \sum_{i=1}^n (Y_i - \beta_0)^2 K\left(\frac{x - X_i}{h}\right) \,.$$

In other words, this estimator is trying to find the constant which best fits the responses corresponding to the covariates in a neighborhood of x. For this reason, the Nadaraya-Watson estimator is generally understood as trying to approximate r by a function which is "locally constant." A natural thing to try is to combine push this one step further, and assume that instead of the zeroth-order approximation $r(u) \approx \beta_0$ near x, we use a first-order (linear) approximation $r(u) \approx \beta_0 + \beta_1(u - x)$, which suggests to examine

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1(X_i - x))^2 K\left(\frac{x - X_i}{h}\right). \tag{10.1}$$

This is called "local linear regression," because it attempts to perform linear regression using the weights $K\left(\frac{x-X_i}{h}\right)$. These weights give larger weights to points near x.

Let us define

$$u_i(x) = \begin{pmatrix} 1 \\ x - X_i \end{pmatrix}.$$

We obtain the following.

Proposition 10.4. *The solution to* (10.1) *is*

$$\hat{\beta}(x) \coloneqq \begin{pmatrix} \hat{\beta}_0(x) \\ \hat{\beta}_1(x) \end{pmatrix} = \left(\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) u_i(x) u_i(x)^\top\right)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) u_i(x) Y_i \,.$$

Proof. In matrix form, we see that $\hat{\beta}$ satisfies

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^2} \sum_{i=1}^{n} (Y_i - \beta^\top u_i(x))^2 K\left(\frac{x - X_i}{h}\right),$$

or, equivalently,

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^2} \beta^\top \left(\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) u_i(x) u_i(x)^\top\right) \beta - 2 \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \beta^\top u_i(x) Y_i \,.$$

By examining the first-order optimality conditions, we obtain that $\hat{\beta}$ satisfies

$$\left(\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) u_i(x) u_i(x)^\top\right) \hat{\beta} = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) u_i(x) Y_i \,,$$

which implies the claim. □

Since this model is based on the heuristic $r(u) \approx \beta_0 + \beta_1(u - x)$, the local linear regression estimator is

$$\hat{r}(x) = \hat{\beta}_0(x) \,.$$

Note that if we write $H = \left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) u_i(x) u_i(x)^\top\right)$, then this estimator is

$$\hat{r}(x) = \sum_{i=1}^{n} \bar{w}_i(x) Y_i \,,$$

where

$$\bar{w}_i(x) = \left(H^{-1} K\left(\frac{x - X_i}{h}\right) u_i(x)\right)_1 .$$

It is not too hard to check that $\sum_{i=1}^{n} \bar{w}_i(x) = 1$, so this procedure is still a weighted average of the Y_i, but just with different weights. It turns out that when r is twice differentiable, this estimator has smaller bias than the Nadaraya-Watson estimator

10.4 Exercises

1. This exercise will prove consistency of a kernel density estimate under weaker assumptions than are used in Proposition 10.2. Let K be the *boxcar kernel*:

$$K(x) \coloneqq \mathbb{1}_{x \in [-1/2, 1/2]} .$$

Let $\hat{f}$ be a kernel density estimate constructed using K.

(a) Show that

$$\mathbb{E}\hat{f}(x) = \frac{1}{h} \int_{x-h/2}^{x+h/2} f(y) \, \mathrm{d}y .$$

(b) Show that

$$\mathrm{Var}(\hat{f}(x)) = \frac{1}{nh^2} \left(\int_{x-h/2}^{x+h/2} f(y) \, \mathrm{d}y - \left(\int_{x-h/2}^{x+h/2} f(y) \, \mathrm{d}y \right)^2 \right) .$$

(c) Show that if f is continuous at x, $h \to 0$ and $nh \to \infty$, then $\hat{f}(x) \xrightarrow{p} f(x)$.

2. As mentioned in the text, choosing the bandwidth properly is very important for constructing good kernel density estimates. This exercise will explore a leave-one-out cross validation approach. Fix a kernel K, and for each choice of bandwidth denote by $\hat{f}_h$ the kernel density estimate constructed using that bandwidth. Write $R_f(h) = \mathbb{E} \int (f(x) - \hat{f}_h(x))^2 \, \mathrm{d}x$ for the risk associated with the kernel density estimator constructed using bandwidth h.

(a) Letting $J_f(h) = \mathbb{E} \int \hat{f}_h(x)^2 \, \mathrm{d}x - 2\mathbb{E} \int \hat{f}_h(x) f(x) \, \mathrm{d}x$, show that $R_f(h) - J_f(h)$ does not depend on h. Therefore, to choose the bandwidth with the smallest risk, it suffices to find a bandwidth minimizing $J_f(h)$.

(b) Suppose that $X'_1, \ldots, X'_n$ is a validation set consisting of independent samples from the same distribution which generated the data $X_1, \ldots, X_n$. Show that

$$\hat{J}_f(h) = \int \hat{f}_h(x)^2 \, \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_h(X'_i)$$

is an unbiased estimator of $J_f(h)$.

(c) In the interest of using the data as efficiently as possible, it is also worthwhile to consider a *leave-one-out* cross validation scheme. Let $\hat{f}_{h,(-i)}$ be the kernel density estimate constructed from all the data points *except* X_i. Show that the leave-one-out estimator

$$\hat{J}_f^{\mathrm{LOO}}(h) = \int \hat{f}_h(x)^2 \, \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{h,(-i)}(X_i)$$

is an unbiased estimator of $J_f(h)$.

3. This exercise generalizes kernel density estimation to higher dimensions.

(a) Given a (univariate) kernel $K : \mathbb{R} \to \mathbb{R}$ satisfying $\int K(x) \, \mathrm{d}x = 1$ and $\int xK(x) \, \mathrm{d}x = 0$, show that the multivariate function $K_d : \mathbb{R}^d \to \mathbb{R}$ given by

$$K_d(x) \coloneqq K_d(x_1, \ldots, x_d) = \prod_{i=1}^{d} K(x_i)$$

satisfies $\int K_d(x) \, \mathrm{d}x = 1$ and $\int x_i K_d(x) \, \mathrm{d}x = 0$ for $i = 1, \ldots, d$.

(b) Given d-dimensional observations $X_1, \ldots, X_n$ and a positive bandwidth h, define

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K_d \left(\frac{x - X_i}{h} \right) .$$

Show that $\int \hat{f}(x) \, \mathrm{d}x = 1$.

(c) Mimicking the proof in the one dimensional case, show that $\int \mathrm{Var}(\hat{f}(x)) \, \mathrm{d}x \leq \frac{\left(\int K^2(u) \, \mathrm{d}u \right)^d}{nh^d}$.

(d) One can show that if f has bounded second derivatives, then $\int (\mathbb{E}\hat{f}(x) - f(x))^2 \, \mathrm{d}x \lesssim h^4$. (You do not need to prove this, but may do so if you wish.) Conclude that by choosing $h \asymp n^{-\frac{1}{d+4}}$, we obtain an estimator with mean integrated squared error of order $n^{-\frac{4}{d+4}}$.

(e) Suppose that we wish to obtain an estimator whose mean integrated squared error is at most .01. Assuming that the estimator above has risk at most $cn^{-\frac{4}{d+4}}$, say how many samples are needed (as a function of c and d) to obtain an estimator of the desired accuracy. The fact that this number scales exponentially in d is known as the *curse of dimensionality*.

(f) One can show that if f possesses s bounded derivatives, then a kernel density estimator achieves the rate $n^{-\frac{2s}{d+2s}}$. Based on this fact, how many derivatives does a density need to possess in order for a kernel density estimator to achieve the rate $n^{-1/2}$, as a function of d?

Chapter 11

High-dimensional linear regression

11.1 Sparsity in the Gaussian sequence model

In this section, we return to the Gaussian sequence model, which we first studied in Section 7.3. Recall that we can write this model as

$$Y_i = \theta_i + \varepsilon_i \,, \tag{11.1}$$

where $\theta = (\theta_1, \dots, \theta_p) \in \mathbb{R}^p$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2/n)$ is i.i.d. noise. As usual, we can write this in vector form as

$$\mathbf{Y} = \theta + \mathcal{E} \,, \tag{11.2}$$

where $\mathcal{E} \sim \mathcal{N}(0, \frac{\sigma^2}{n} I)$. The presence of the $1/n$ in the variance of $\mathcal{E}$ reflects that we may imagine that we have n independent copies of $\mathbf{Y}$, and by averaging these copies we can reduce the variance by a factor of n. This will facilitate the comparison with linear regression in the following lectures. Even though this model doesn't include covariates, a lot of intuition about high-dimensional linear regression—and other high-dimensional models—can be developed by studying this model.

By mimicking our earlier analysis of the gaussian sequence modelwe obtain the following result.

Proposition 11.1. *The maximum likelihood estimator for $\theta \in \Theta = \mathbb{R}^p$ in the model* (11.2) *is* $\mathbf{Y}$*. This estimator achieves*

$$\mathbb{E}_\theta \|\theta - \mathbf{Y}\|^2 = \frac{\sigma^2 p}{n} \,. \tag{11.3}$$

When $p \gg n$ (the high-dimensional regime), the risk in (11.3) is large, and even if $n \to \infty$, the risk does not vanish so long as $p \to \infty$ at least as fast. This phenomenon indicates that we need to make additional assumptions on the parameter space if we are going to be able to consistently estimate θ. In Section 7.3, we focused on *smoothness*: by assuming that the coordinates of θ vary in a Lipschitz way, we were able to consistently estimate θ, albeit at a slower-than-parametric rate. In this lecture, we will focus on a different assumption, *sparsity*. Specifically, we consider the parameter space

$$\Theta_{\ell_0}(k) \coloneqq \left\{ \theta \in \mathbb{R}^p : \sum_{i=1}^p \mathbb{1}_{\theta_i \neq 0} \leq k \right\}$$

of vectors with a small number of nonzero coordinates. Such vectors are called "k-sparse," or just "sparse." (The notation ℓ_0 will be explained soon.)

There are several justifications for sparsity, from both mathematical and practical perspectives. Practically speaking, it is quite reasonable to assume in a number of applications that most of the quantities in question have a zero or minimal impact on a particular outcome of interest, and that most of the contribution comes from a small number of important factors—though, of course, *which* factors are important is

unknown. This is essentially a sparsity assumption. Perhaps remarkably, this assumption on the parameter space significantly improves the statistical properties of the problem as well.

If we know *a priori* that the parameter θ is sparse, then it is natural to estimate θ by a vector which is itself sparse. This leads to a simple idea: if the observed entry Y_i is large, then it was unlikely that $\theta_i = 0$, since this would mean that the noise was very large. On the other hand, if Y_i is near zero, then we might expect that this corresponds to an entry of θ which was originally zero.

Definition 11.2. Given a positive threshold τ, the *hard thresholding estimator* $\hat{\theta}_\tau$ is given by

$$(\hat{\theta}_\tau)_i = Y_i \mathbb{1}_{|Y_i| \geq \tau}\,.$$

In other words, the hard thresholding estimator preserves entries of Y_i that are far from zero, and zeroes out entries which are close to zero. Our main result in this lecture is that the hard thresholding estimator achieves good performance.

Proposition 11.3. *Let $\tau = 2\sqrt{\frac{2\sigma^2 \log(2p/\delta)}{n}}$. Then for any $\theta \in \Theta_{\ell_0}(k)$*

$$\|\hat{\theta}_\tau - \theta\|^2 \leq 18\frac{\sigma^2 k \log(2p/\delta)}{n} \qquad \textit{with probability at least } 1-\delta.$$

Before giving the proof, we make three remarks about Proposition 11.3. First, note that we have given a bound on $\|\hat{\theta}_\tau - \theta\|^2$ which holds with high probability (rather than in expectation). By being slightly more careful, it is possible to show that a similar estimator (with a good choice of δ) leads to a bound in expectation as well, but we will not pursue this direction here. Second, note that the quantity $\frac{\sigma^2 k \log p}{n}$ now depends on p only logarithmically. In other words, this yields an *exponential* improvement over (11.3). Third, note the remarkable fact that implementing this estimator *does not require knowing* k. This property is known as *adaptivity*, and we say that the hard thresholding estimator adapts to the right value of k.

Proof of Proposition 11.3. Consider the event $\mathcal{A} \coloneqq \{|\varepsilon_i| \leq \frac{\tau}{2} \quad \forall i = 1, \dots, p\}$. By Proposition 2.1, this event holds with probability at least $1-\delta$. We will show that, on this event, $\|\hat{\theta}_\tau - \theta\|^2 \leq 18\frac{\sigma^2 k \log(2p/\delta)}{n}$.

For the remainder of the proof, assume $\mathcal{A}$ holds. Under this assumption,

$$\begin{aligned} |Y_i| \geq \tau &\implies |\theta_i| \geq |Y_i| - |\varepsilon_i| \geq \frac{\tau}{2} \\ |Y_i| < \tau &\implies |\theta_i| < |Y_i| + |\varepsilon_i| < \frac{3\tau}{2} \end{aligned}$$

Combining these bounds, we obtain

$$\begin{aligned} |(\hat{\theta}_\tau)_i - \theta_i| &= |\varepsilon_i| \mathbb{1}_{|Y_i| \geq \tau} + |\theta_i| \mathbb{1}_{|Y_i| < \tau} \\ &\leq \frac{\tau}{2} \mathbb{1}_{|Y_i| \geq \tau} + |\theta_i| \mathbb{1}_{|Y_i| < \tau}\,. \end{aligned}$$

We now analyze three cases:

1. If $|\theta_i| < \frac{\tau}{2}$, then $\frac{\tau}{2}\mathbb{1}_{|Y_i| \geq \tau} + |\theta_i| \mathbb{1}_{|Y_i| < \tau} = |\theta_i| = \min\{\tau/2, |\theta_i|\}$.
2. If $|\theta_i| \geq \frac{3\tau}{2}$, then $\frac{\tau}{2}\mathbb{1}_{|Y_i| \geq \tau} + |\theta_i| \mathbb{1}_{|Y_i| < \tau} = \tau/2 = \min\{\tau/2, |\theta_i|\}$.
3. If $\frac{\tau}{2} \leq |\theta_i| < \frac{3\tau}{2}$, then $\frac{\tau}{2}\mathbb{1}_{|Y_i| \geq \tau} + |\theta_i| \mathbb{1}_{|Y_i| < \tau} \leq \max\{\tau/2, |\theta_i|\} \leq 3\min\{\tau/2, |\theta_i|\}$.

We obtain

$$\|\hat{\theta}_\tau - \theta\|^2 \leq \sum_{i=1}^{p} 9\min\{\tau^2/4, \theta_i^2\} \leq 18\frac{\sigma^2 k \log(2p/\delta)}{n}\,,$$

as claimed. □

11.2 ℓ_0 and ℓ_1 norms

It is possible to show that

$$\hat{\theta}_\tau \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{Y} - \theta\|^2 + \frac{\tau^2}{2} \sum_{i=1}^{p} \mathbb{1}_{\theta_i \neq 0} \,. \tag{11.4}$$

(The notation $\hat{\theta}_\tau \in \operatorname{argmin}$ indicates that $\hat{\theta}_\tau$ may not be the unique minimizer of the right side.) The quantity $\sum_{i=1}^{p} \mathbb{1}_{\theta_i \neq 0}$ is called the ℓ_0 "norm," and denoted

$$\|\theta\|_0 = \sum_{i=1}^{p} \mathbb{1}_{\theta_i \neq 0} \,.$$

We put "norm" in quotes because $\|\theta\|_0$ does not satisfy the definitions of a norm, since $\|ax\|_0 \neq |a|\|x\|_0$ in general. Nevertheless, this "norm" has many norm-like properties, which justify the name.

The formulationn in (11.4) suggests that, if we want to find sparse solutions for models other than the Gaussian sequence model, we can simply perform estimation with an additional penalty. For instance, for the case of maximum-likelihood estimation, we are led to consider estimators obtained as solutions to

$$\operatorname*{argmin}_{\theta \in \mathbb{R}^p} -\ell(\theta) + \frac{\tau^2}{2}\|\theta\|_0 \,. \tag{11.5}$$

The problem with focusing on equations of this type is that $\|\theta\|_0$ is not a convex function of θ. While this does not mean that it is impossible to solve (11.5)—after all, we were able to solve it in the case of the Gaussian sequence model—it does mean that we cannot in general apply the machinery of convex optimization to (11.5), even when ℓ itself is well behaved.

A common approach, which you may have seen in other contexts, is to replace a nonconvex function by a convex function with similar properties. This is very often done in the context of high-dimensional statistics problems by replacing $\|\theta\|_0$ by the ℓ_1 norm:

$$\|\theta\|_1 \coloneqq \sum_{i=1}^{p} |\theta_i| \,.$$

Unlike the ℓ_0 norm, the function $\theta \mapsto \|\theta\|_1$ is convex, which makes it attractive from an optimization perspective.

It turns out that the ℓ_1 norm is a very sensible regularizer in a variety of contexts. For example, it is possible to show that

$$\hat{\theta}^{\ell_1}_\tau \coloneqq \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{Y} - \theta\|^2 + \tau\|\theta\|_1 \,, \qquad \tau \geq 0$$

has the following simple form:

$$(\hat{\theta}^{\ell_1}_\tau)_i = \begin{cases} Y_i - \tau & \text{if } Y_i > \tau \\ 0 & \text{if } Y_i \in [-\tau, \tau] \\ Y_i + \tau & \text{if } Y_i < -\tau \,. \end{cases} \tag{11.6}$$

In other words, this estimator shrinks Y_i back towards zero in a continuous fashion. (This is sometimes called "soft thresholding.")

While a full investigation of the properties of the ℓ_1 norm is outside the scope of this course, we will survey a few ideas that indicate that $\|\theta\|_1$ is a good substitute for the ℓ_0 norm. The first gives a connection between the ℓ_1 norm, the ℓ_0 norm, and the standard ℓ_2 (Euclidean) norm. It shows that, when the ℓ_0 norm is small, the ℓ_1 norm is not too large.

Proposition 11.4. *If* $\|\theta\| = (\sum_{i=1}^{p} \theta_i^2)^{1/2} = 1$, *then*

$$\|\theta\|_1 \leq \sqrt{\|\theta\|_0} \,.$$

Proof. We have

$$\begin{aligned}\|\theta\|_1 &= \sum_{i=1}^{p} |\theta_i| \\ &= \sum_{i:\theta_i \neq 0} |\theta_i| \\ &= \|\theta\|_0 \cdot \left(\frac{1}{\|\theta\|_0} \sum_{i:\theta_i \neq 0} |\theta_i| \right) \\ &\leq \|\theta\|_0 \cdot \left(\frac{1}{\|\theta\|_0} \sum_{i:\theta_i \neq 0} |\theta_i|^2 \right)^{1/2} ,\end{aligned}$$

where the last step follows from Jensen's inequality. □

The second, more interesting connection gives a partial converse to Proposition 11.4, and shows that, if $\|\theta\|_1$ is small, then θ is *approximately* sparse, in the sense that it is close to a sparse vector.

Proposition 11.5. *For any positive integer k and vector $\theta \in \mathbb{R}^p$, there exists a vector $\varphi \in \mathbb{R}^p$ such that $\|\theta - \varphi\| \leq \|\theta\|_1/\sqrt{k}$ and $\|\varphi\|_0 \leq k$.*

Proof. The claim is obvious if $\theta = 0$, so we assume that $\theta \neq 0$. By rescaling, we may assume that $\|\theta\|_1 = 1$ without loss of generality. The clever idea is to view the entries of θ as describing a probability distribution on the set $\{1, \ldots, p\}$: we assign mass $|\theta_i|$ to i. Since $\|\theta\|_1 = 1$, this describes a valid probability distribution.

Let us now draw k i.i.d. samples $I_1, \ldots, I_k$ from this distribution, and construct the vector ϕ given by

$$\phi_i = \frac{|\{j : I_j = i\}|}{k} \operatorname{sign}(\theta_i) .$$

Note that $\|\phi\|_0 \leq k$. Moverover, $\mathbb{E}\phi_i = \frac{1}{k} \sum_{j=1}^{k} \mathbb{P}\{I_j = i\} \operatorname{sign}(\theta_i) = \operatorname{sign}(\theta_i)|\theta_i| = \theta_i$ and

$$\operatorname{Var}(\phi_i) = \frac{1}{k} \mathbb{P}\{I_j = i\})(1 - \mathbb{P}\{I_j = i\}) < \frac{|\theta_i|}{k} .$$

We obtain

$$\mathbb{E}\|\phi - \theta\|^2 = \sum_{i=1}^{p} \mathbb{E}(\phi_i - \theta_i)^2 = \sum_{i=1}^{p} \frac{|\theta_i|}{k} = \frac{1}{k} .$$

We now come to the crucial step: since $\|\phi\|_0 \leq k$ with probability 1, we have

$$\min_{\varphi : \|\varphi\|_0 \leq k} \|\varphi - \theta\|^2 \leq \mathbb{E}\|\phi - \theta\|^2 \leq \frac{1}{k} .$$

This proves the claim. □

One interesting implication of Proposition 11.5, which we will not pursue further here, is that it can be used to show that estimators which work well on the parameter space $\Theta_{\ell_0}(k)$ also tend to work well on "ℓ_1 balls" of the form

$$\Theta_{\ell_1}(\lambda) \coloneqq \{\theta \in \mathbb{R}^p : \|\theta\|_1 \leq \lambda\} .$$

11.3 The Lasso

The idea of ℓ_1 regularization has proven most influential in its application to high-dimensional linear regression. Let us return to the linear model (7.1), and let us allow for the possibility that $p \gg n$, so that we are truly in the high-dimensional regime. To avoid issues of identifiability, we will evaluate an estimator $\hat{\beta}$ by its "in-sample error":

$$\frac{1}{n}\sum_{i=1}^{n}(X_i^\top \hat{\beta} - X_i^\top \beta)^2 \,.$$

The justification for the term $\frac{1}{n}$ comes from the idea that we would like our average error across the n observations to be small.

Given $\tau \geq 0$, the *Lasso* estimator $\hat{\beta}_{\mathcal{L}}$ is defined to be

$$\hat{\beta}_{\mathcal{L}} \coloneqq \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 + \tau\|\beta\|_1 \,.$$

Like the hard thresholding estimator considered above, the Lasso achieves good performance, even when $p \gg n$, so long as $\|\beta\|_1$ is bounded.

Proposition 11.6. *Consider the linear regression model* (7.1) *with Gaussian noise of norm* σ^2*, and assume that* $|(X_i)_j| \leq 1$ *for all* $i = 1, \ldots, n$*,* $j = 1, \ldots, p$*. For any* $\delta > 0$*, the Lasso estimator with regularization parameter* $\tau = \sqrt{\frac{2\sigma^2 \log(2p/\delta)}{n}}$ *achieves*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i^\top \hat{\beta} - X_i^\top \beta)^2 \leq 4\|\beta\|_1\sqrt{\frac{2\sigma^2 \log(2p/\delta)}{n}} \qquad \textit{with probability at least } 1-\delta.$$

Proof. By definition of $\hat{\beta}_{\mathcal{L}}$, we have

$$\frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^\top \hat{\beta}_{\mathcal{L}})^2 + \tau\|\hat{\beta}_{\mathcal{L}}\|_1 \leq \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 + \tau\|\beta\|_1 \,. \tag{11.7}$$

By expanding the square, we have

$$\begin{aligned}(Y_i - X_i^\top \hat{\beta}_{\mathcal{L}})^2 &= (Y_i - X_i^\top \beta + X_i^\top \beta - X_i^\top \hat{\beta}_{\mathcal{L}})^2 \\ &= (Y_i - X_i^\top \beta)^2 + (X_i^\top \beta - X_i^\top \hat{\beta}_{\mathcal{L}})^2 + 2(Y_i - X_i^\top \beta)(X_i^\top \beta - X_i^\top \hat{\beta}_{\mathcal{L}}) \\ &= (Y_i - X_i^\top \beta)^2 + (X_i^\top \beta - X_i^\top \hat{\beta}_{\mathcal{L}})^2 + 2\varepsilon_i(X_i^\top \beta - X_i^\top \hat{\beta}_{\mathcal{L}}) \,.\end{aligned}$$

Rearranging (11.7), this yields

$$\frac{1}{n}\sum_{i=1}^{n}(X_i^\top \hat{\beta} - X_i^\top \beta)^2 \leq 2\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i\right)^\top (\hat{\beta}_{\mathcal{L}} - \beta) + 2\tau(\|\beta\|_1 - \|\hat{\beta}_{\mathcal{L}}\|_1) \,.$$

Now, consider the vector $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i \in \mathbb{R}^p$. Since each entry of X_i has magnitude at most 1, each entry of $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i \in \mathbb{R}^p$ is a Gaussian with mean 0 and variance at most σ^2/n. Therefore, by Proposition 2.1, with probability at least $1-\delta$ we have

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i\right)_j\right| \leq \sqrt{\frac{2\sigma^2 \log(2p/\delta)}{n}} = \tau \,, \quad \forall j$$

By Hölder's inequality, which you will verify on this form in your homework, for any two vectors $v, u \in \mathbb{R}^p$, it holds

$$v^\top u \leq \max_{j=1,\ldots,p}|v_j| \cdot \|u\|_1 \,. \tag{11.8}$$

Applying this to the vectors $\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i$ and $\hat{\beta}_{\mathcal{L}} - \beta$ yields

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (X_i^\top \hat{\beta} - X_i^\top \beta)^2 &\le 2\tau\|\hat{\beta}_{\mathcal{L}} - \beta\|_1 + 2\tau(\|\beta\|_1 - \|\hat{\beta}_{\mathcal{L}}\|_1) \\ &\le 2\tau(\|\hat{\beta}_{\mathcal{L}}\|_1 + \|\beta\|_1) + 2\tau(\|\beta\|_1 - \|\hat{\beta}_{\mathcal{L}}\|_1) \\ &= 4\|\beta\|_1\tau\,,\end{aligned} \tag{11.9}$$

as claimed. □

11.4 Slow vs. fast rate

A possibly surprising fact about Proposition 11.6 is that the squared in-sample error converges to zero at the rate $n^{-1/2}$. If we recall other results proven in this class so far, we have typically seen the square loss of estimators converging at the rate n^{-1}. For example, Proposition 11.3 shows that the hard thresholding estimator achieves the rate n^{-1}. Why does the estimator in Proposition 11.6 converge more slowly?

The different between the rate $n^{-1/2}$ and n^{-1} can be understood partially as a failure of strong convexity. As we described before when understanding the asymptotic normality of M-estimators, minimizing a strongly convex loss function will lead to favorable convergence properties, since it implies that errors are more localized. By contrast, this problem exhibits a "slow rate" of $n^{-1/2}$. Can we find a form of strong convexity to prove that this estimator actually converges at the fast rate n^{-1}?

Recall that our object of interest is the in-sample prediction error

$$\hat{\beta} \mapsto \frac{1}{n}\sum_{i=1}^n (X_i^\top \hat{\beta} - X_i^\top \beta)^2\,.$$

We say that this loss is λ-strongly convex if

$$\frac{1}{n}\sum_{i=1}^n (X_i^\top \hat{\beta} - X_i^\top \beta)^2 \ge \lambda\|\hat{\beta} - \beta\|^2\,, \quad \forall \hat{\beta} \in \mathbb{R}^p$$

or, equivalently, if

$$\frac{1}{n}\sum_{i=1}^n (X_i^\top v)^2 \ge \lambda\|v\|^2\,, \quad \forall v \in \mathbb{R}^p\,. \tag{11.10}$$

The first observation is that in the high-dimensional regime when $p \gg n$, this error is *never* strongly convex—indeed, if $v \in \mathbb{R}^p$ is any vector such that $X_i^\top v = 0$ for $i = 1, \dots, n$, then this inequality obviously fails. We therefore seem to be stuck. However, it turns out that there is a restricted version of strong convexity, which *can* hold in the high-dimensional regime, and which allows us to recover the fast rate for the Lasso.

To proceed, we first note that we do not actually need (11.10) to hold for all v—it will be enough if it holds, at least approximately, for $v = \hat{\beta}_{\mathcal{L}} - \beta$. The key observation is that, so long as β is sparse, $\hat{\beta}_{\mathcal{L}} - \beta$ is likely to be approximately a sparse vector. Indeed, if β is sparse and $\hat{\beta}$ is close to β, then $\hat{\beta}_{\mathcal{L}}$ should be approximately sparse too.

Let us formalize this notion of approximate sparsity by the following condition.

Definition 11.7. Given a set $S \subseteq \{1, \dots, p\}$, let

$$C(S) \coloneqq \{v \in \mathbb{R}^p : \|v_{S^C}\|_1 \le 3\|v_S\|_1\}\,,$$

where for any subset $T \subseteq \{1, \dots, p\}$ the notation v_T means the subvector of v consisting only of the coordinates in T.

We say that $(X_1, \dots, X_n)$ satisfy the *restricted eigenvalue condition* for S if there exists $\lambda_{\mathrm{RE}} > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i^\top v)^2 \geq \lambda_{\mathrm{RE}}\|v\|^2\,, \quad \forall v \in C(S)\,.$$

The important point is that since the set $C(S)$ is much smaller than $\mathbb{R}^p$, it is possible for the restricted eigenvalue condition to hold even when $p \gg n$.[1]

We now show that if $(X_1, \dots, X_n)$ satisfy the restricted eigenvalue condition for the support of β (i.e., the set of coordinates of β which are nonzero), then the Lasso attains the fast rate.

Proposition 11.8. *Consider the setting of Proposition 11.6, with* $\tau = 2\sqrt{\frac{2\sigma^2\log(2p/\delta)}{n}}$. *Assume that* $(X_1, \dots, X_n)$ *satisfy the restricted eigenvalue condition for* S, *where* S *is the support of* β. *Then with probability at least* $1-\delta$, *the vector* $\hat{\beta}_{\mathcal{L}} - \beta$ *lies in* $C(S)$, *and*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\hat{\beta} - X_i^\top\beta)^2 \leq 72\|\beta\|_0\frac{\sigma^2\log(2p/\delta)}{\lambda_{\mathrm{RE}}n}\,.$$

Proof. Write $\Delta = \hat{\beta}_{\mathcal{L}} - \beta$. By the same argument as led to (11.9), we have that with probability at least $1-\delta$,

$$0 \leq \frac{1}{n}\sum_{i=1}^{n}(X_i^\top\Delta)^2 \leq 2\max_j\left|\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i\right)_j\right|\|\Delta\|_1 + 2\tau(\|\beta\|_1 - \|\hat{\beta}_{\mathcal{L}}\|_1) \leq \tau\|\Delta\|_1 + 2\tau(\|\beta\|_1 - \|\hat{\beta}_{\mathcal{L}}\|_1)\,.$$

Now, note that

$$\|\beta\|_1 - \|\hat{\beta}_{\mathcal{L}}\|_1 = \|\beta_S\| - \|(\hat{\beta}_{\mathcal{L}})_S\| - \|\Delta_{S^C}\|_1 \leq \|\Delta_S\|_1 - \|\Delta_{S^C}\|_1\,.$$

We obtain that

$$0 \leq \frac{1}{n}\sum_{i=1}^{n}(X_i^\top\Delta)^2 \leq \tau\|\Delta\|_1 + 2\tau(\|\Delta_S\|_1 - \|\Delta_{S^C}\|_1) \leq 3\tau\|\Delta_S\|_1 - \tau\|\Delta_{S^C}\|_1\,.$$

Therefore $\hat{\beta}_{\mathcal{L}} - \beta$ lies in $C(S)$, as claimed. Moreover, dropping the term $\tau\|\Delta_{S^C}\|$ yields

$$\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\Delta)^2 \leq 3\tau\|\Delta_S\|_1 \leq 3\tau\sqrt{|S|}\|\Delta_S\|\,.$$

Squaring both sides and applying the restricted eigenvalue assumption yields

$$\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\Delta)^2\right)^2 \leq 9\tau^2|S|\|\Delta_S\|^2 \leq \frac{9}{\lambda_{\mathrm{RE}}}\tau^2|S|\cdot\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\Delta)^2\,,$$

and dividing through by $\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\Delta)^2$ yields the claim. □

Is the dependence on λ_{RE} necessary? It turns out that if we replace the ℓ_1 penalty in the definition of the Lasso by an ℓ_0 penalty, then the resulting estimator achieves the fast rate without any dependence on λ_{RE}. However, this estimator is computationally intractable. A remarkable result due to Zhang et al. (2014) shows that, in fact, this problem evinces what is known as a *statistical-computational gap*, and that *no* computationally efficient estimator can avoid dependence on λ_{RE}.

[1] Indeed, it can be shown that if the vectors X_i are suitably generic—say, they are drawn i.i.d. from the Gaussian distribution—then the restricted eigenvalue condition holds with high probability.