Statistical Estimation

Lecture Notes for MATH 6262 at Georgia Tech

Cheng Mao

School of Mathematics, Georgia Tech

May 2, 2023

This set of notes, based on [LC06, Kee11, Ber18, RH19, Tsy08] and other sources, is provided to the students in the course MATH 6262 at Georgia Tech as a complement to the lectures. It is not meant to be a complete introduction to the subject.

Contents

Chapter 1

Fundamentals of statistical estimation

1.1 Background and setup

1.1.1 Review of probability

Consider a sample space $\mathcal{X}$ containing all possible outcomes of an experiment. Let μ be the reference (or natural) measure on $\mathcal{X}$. We primarily consider the following spaces:

- A finite or countable set $\mathcal{X}$ equipped with the counting measure μ. For example, when we roll a die, the outcome lies in $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$. Moreover, if we consider the number of times we throw a coin before a "heads" is observed, then this number lies in $\mathcal{X} = \{1, 2, 3, \dots\}$.
- $\mathcal{X} = \mathbb{R}^d$ equipped with the Lebesgue measure μ. For example, tomorrow's temperature is in $\mathcal{X} = \mathbb{R}$, while tomorrow's temperature and humidity jointly lie in $\mathcal{X} = \mathbb{R}^2$.

A random variable X is an experiment taking values in $\mathcal{X}$. We write $X \sim \mathcal{P}$ if X follows a distribution $\mathcal{P}$. There are several ways to describe a random variable or a distribution:

- If X is discrete, i.e., $\mathcal{X}$ is finite or countable, we can specify the probability mass function (PMF) f_X of X. For example, for the uniform random variable $X \sim \mathsf{Unif}([n])$ where $[n] := \{1, \dots, n\}$, we have $f_X(i) = \mathbb{P}\{X = i\} = 1/n$ for $i = 1, \dots, n$.
- If X is continuous, e.g., $X = \mathbb{R}$ or $\mathbb{R}^d$, we can specify the probability density function (PDF or density) f_X of X. For example, for the standard Gaussian random variable $X \sim \mathsf{N}(0, 1)$, we have $f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ for $t \in \mathbb{R}$.
- The cumulative distribution function (CDF) of a random variable X on $\mathbb{R}$ is $F_X(t) = \mathbb{P}\{X \le t\}$. We have $F_X'(t) = f_X(t)$ and $\int_{-\infty}^{t} f_X(s)\, ds = F_X(t)$. The CDF of $X = (X_1, \dots, X_d)$ on $\mathbb{R}^d$ is $F_X(t_1, \dots, t_d) = \mathbb{P}\{X_1 \le t_1, \dots, X_d \le t_d\}$.

In general, for a subset $E \subset \mathcal{X}$, the probability of the event $\{X \in E\}$ is $\mathbb{P}\{X \in E\} = \int_E f_X \, d\mu$. Examples:

- Roll a die; the outcome is $X \sim \mathsf{Unif}([6])$. The probability of seeing 2 or 3 is $\mathbb{P}\{X \in \{2, 3\}\} = \sum_{i=2}^{3} 1/6 = 1/3$.
- Consider $X \sim \mathsf{N}(0, 1)$. The probability that X is positive is $\mathbb{P}\{X > 0\} = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = 1/2$.

The expectation of X is $\mathbb{E}[X] = \int_{\mathcal{X}} t\, f_X(t)\, d\mu(t)$. Given a function $g : \mathcal{X} \to \mathbb{R}$, the expectation of $g(X)$ is $\mathbb{E}[g(X)] = \int_{\mathcal{X}} g(t) f_X(t)\, d\mu(t)$. Examples:

- For $X \sim \mathsf{Unif}([6])$, $\mathbb{E}[X] = \sum_{i=1}^{6} i \cdot \frac{1}{6} = 3.5$.
- For $X \sim \mathsf{N}(0,1)$, the variance of X is $\mathbb{E}[(X-0)^2] = \int_{-\infty}^{\infty} t^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = 1$.

1.1.2 Setup of statistical estimation

Statistical estimation is in some sense the reverse engineering of probability. Observing realizations of random variables $X_1, X_2, \ldots, X_n$, we aim to estimate certain (functions of) parameters of the underlying distribution. Let us describe the basic setup of parametric estimation using throwing a biased coin as a running example. Consider a biased coin for which we see 1 (heads) with probability $\theta \in [0,1]$ and see 0 (tails) with probability $1-\theta$. In other words, the observation follows the $\mathsf{Ber}(\theta)$ distribution. The following is a list of basic concepts in parametric estimation:

- Parameter: θ, which is typically a real number. E.g., $\theta = 0.3, 0.5$, or 0.8.
- Parameter space: the set Θ of parameters. E.g., $\Theta = [0,1]$.
- Probability distribution: $\mathcal{P}_\theta$. E.g., $\mathcal{P}_\theta = \mathsf{Ber}(\theta)$.
- Family of distributions: the set $\mathscr{P}$ containing all $\mathcal{P}_\theta$. E.g., $\mathscr{P} = \{\mathsf{Ber}(\theta) : \theta \in [0,1]\}$.
- Observations: i.i.d. $X_1, \ldots, X_n \sim \mathcal{P}_\theta$. E.g., $X_1, \ldots, X_n$ are the binary outcomes of n independent coin throws.
- Statistic: a function of the observations $X_1, \ldots, X_n$. E.g., $h(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- Estimand: a function $g(\theta)$ of the parameter θ. E.g., $g(\theta) = \theta$ or θ^2.
- Estimator: a statistic which is used to estimate the estimand, denoted by $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ or $\hat{g} = \hat{g}(X_1, \ldots, X_n)$. E.g., $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ when $g(\theta) = \theta$, or $\hat{g} = g(\hat{\theta}) = \hat{\theta}^2 = (\frac{1}{n}\sum_{i=1}^{n} X_i)^2$ when $g(\theta) = \theta^2$.
- Loss function: a bivariate function $L(g, \hat{g}) \geq 0$. E.g., the squared loss $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$.
- Risk: the expectation of the loss $\mathbb{E}[L(g, \hat{g})]$ with respect to the observations. E.g., $\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta)^2]$ for i.i.d. $X_1, \ldots, X_n \sim \mathsf{Ber}(\theta)$.

1.2 Exponential families

1.2.1 Definition and examples

Definition 1.1. *Let $\Theta \subset \mathbb{R}^d$ be a parameter space. A family $\{\mathcal{P}_\theta\}_{\theta\in\Theta}$ of probability distributions on a sample space $\mathcal{X}$ with measure μ is called an exponential family if $\mathcal{P}_\theta$ has PDF (or PMF)*

$$f(x \mid \theta) = \exp\left(\eta^\top T(x) - A(\eta)\right) \cdot h(x).$$

Here, $\eta = \eta(\theta) \in \mathbb{R}^m$ is the natural parameter, $T(x) \in \mathbb{R}^m$ is the sufficient statistic,

$$A(\eta) = \log \int_{\mathcal{X}} \exp\big(\eta^\top T(x)\big) \cdot h(x)\, d\mu(x)$$

is the log-partition function, and $h(x)$ is the base measure. Moreover, the natural parameter space is $\mathcal{E} := \{\eta : A(\eta) < \infty\} = \{\eta : \int_{\mathcal{X}} \exp\big(\eta^\top T(x)\big) \cdot h(x)\, d\mu(x) < \infty\}$.

Examples of exponential families include:

- Gaussian distribution $\mathsf{N}(\mu, \sigma^2)$: If $\eta_1 = \frac{\mu}{\sigma^2}$ and $\eta_2 = \frac{-1}{2\sigma^2}$, then

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\Big) \\ &= \exp\Big(\eta_1 x + \eta_2 x^2 + \frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log(-2\eta_2)\Big) \frac{1}{\sqrt{2\pi}}. \end{aligned}$$

- Poisson distribution $\mathsf{Poi}(\lambda)$: If $\eta = \log \lambda$, then

$$f(x \mid \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} = \exp\big(x \log \lambda - \lambda\big) \cdot \frac{1}{x!} = \exp\big(\eta x - e^\eta\big) \cdot \frac{1}{x!}.$$

- Binomial distribution $\mathsf{Bin}(n, p)$: If n is fixed and known, and $\eta = \log \frac{p}{1-p}$, then

$$f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \exp\Big(x \log \frac{p}{1-p} + n \log(1-p)\Big) = \exp\Big(\eta x - n \log(1+e^\eta)\Big) \binom{n}{x}.$$

1.2.2 Moments and cumulants

For a random variable X taking values in $\mathbb{R}^m$, consider its moments

$$\alpha_{r_1,\dots,r_m} = \mathbb{E}\big[X_1^{r_1} \cdots X_m^{r_m}\big], \quad r_i \in \mathbb{Z}_{\geq 0}.$$

The moment generating function (MGF) is defined as

$$M_X(u) := \mathbb{E}\big[\exp\big(u^\top X\big)\big], \quad u \in \mathbb{R}^m.$$

If M_X exists in a neighborhood of the origin, then all moments exist and

$$M_X(u) = \sum_{r_1,\dots,r_m} \frac{\alpha_{r_1,\dots,r_m}}{r_1! \cdots r_m!} u_1^{r_1} \cdots u_m^{r_m}.$$

Therefore, we have

$$\alpha_{r_1,\dots,r_m} = \frac{\partial^{r_1+\cdots+r_m} M_X(u)}{\partial u_1^{r_1} \cdots \partial u_m^{r_m}}\bigg|_{u=0}.$$

The cumulant generating function (CGF) is defined as $K_X(u) := \log M_X(u)$. Its power series expansion is

$$K_X(u) = \sum_{r_1,\dots,r_m} \frac{\kappa_{r_1,\dots,r_m}}{r_1! \cdots r_m!} u_1^{r_1} \cdots u_m^{r_m},$$

where we call $\kappa_{r_1,\ldots,r_m}$ the cumulants of X.

In the case that $m = 1$, i.e., the random variable is real-valued, we have

$$\alpha_1 = \kappa_1, \quad \alpha_2 = \kappa_2 + \kappa_1^2, \quad \alpha_3 = \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3, \quad \ldots$$

There is a general relation via Bell polynomials.

If $X_1, \ldots, X_n$ are independent real-valued random variables and $X := \sum_{i=1}^n X_i$, then $M_X(u) = \prod_{i=1}^n M_{X_i}(u)$ and thus $K_X(u) = \sum_{i=1}^n K_{X_i}(u)$. Therefore, $\kappa_r(X) = \sum_{i=1}^n \kappa_r(X_i)$, i.e., the rth cumulant of X is the sum of the rth cumulants of $X_1, \ldots, X_n$.

Consider $X \sim \mathcal{P}_\theta$ where $\mathcal{P}_\theta$ is an exponential family. Note that $T = T(X)$ is a random variable. Assuming mild regularity conditions, e.g., $\{\eta(\theta) : \theta \in \Theta\}$ is open, and $M_T(u)$ and $K_T(u)$ exist in a neighborhood of the origin, we have

$$M_T(u) = \mathbb{E}[\exp(u^\top T)] = \int_{\mathcal{X}} \exp\big((u+\eta)^\top T(x) - A(\eta)\big) h(x)\, d\mu(x) = \exp\big(A(\eta+u) - A(\eta)\big).$$

Then we can compute moments of T from $M_T(u)$. It also follows that

$$K_T(u) = A(\eta + u) - A(\eta).$$

From this, it can be derived that

$$\mathbb{E}[T] = \nabla A(\eta), \quad \mathsf{Cov}(T) = \nabla^2 A(\eta), \quad \ldots$$

Examples:

- $X \sim \mathsf{Poi}(\lambda)$: $T(x) = x$, $\eta = \log\lambda$, and $A(\eta) = e^\eta$. Hence

$$M_X(u) = \exp\big(e^{\eta+u} - e^\eta\big) = \exp\big(\lambda(e^u - 1)\big).$$

  From $M_X(u)$, we can compute

$$\mathbb{E}[X] = \lambda, \quad \mathbb{E}[X^2] = \lambda^2 + \lambda, \quad \mathbb{E}[X^3] = \lambda^3 + 3\lambda^2 + \lambda, \quad \ldots$$

- $X \sim \mathsf{Bin}(n, p)$: $T(x) = x$, $\eta = \log\frac{p}{1-p}$, and $A(\eta) = n\log(1 + e^\eta)$. Hence

$$M_X(u) = \Big(\frac{1 + e^{\eta+u}}{1 + e^\eta}\Big)^n = (1 - p + pe^u)^n.$$

  On the other hand, if we use the definition of the MGF, we need to compute

$$M_X(u) = \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} e^{ux}.$$

  This is slightly more involved than using the general formula for exponential families.

1.2.3 Stein's lemma

Lemma 1.2 (Stein)**.** *Let $\{\mathcal{P}_\theta : \theta \in \Theta\}$ be an exponential family. Suppose that $X \sim \mathcal{P}_\theta$ has density $f(x \mid \theta)$ for $x \in \mathbb{R}$. Let g be a differentiable function such that $\lim_{x\to\pm\infty} g(x)f(x \mid \theta) = 0$. Then we have*

$$\mathbb{E}\left[g(X)\Big(\frac{h'(X)}{h(X)} + \eta^\top T'(X)\Big)\right] = -\,\mathbb{E}[g'(X)].$$

In particular, for $X \sim \mathsf{N}(\mu, \sigma^2)$, we have

$$\mathbb{E}[g(X)(X-\mu)] = \sigma^2\,\mathbb{E}[g'(X)]. \tag{1.1}$$

Proof. By integration by parts, the RHS is equal to

$$\begin{aligned}
&-\int_{\mathbb{R}} g'(x)\exp\big(\eta^\top T(x) - A(\eta)\big)h(x)\,d\mu(x) \\
&\quad= \int_{\mathbb{R}} g(x)\Big(\eta^\top T'(x)\exp\big(\eta^\top T(x) - A(\eta)\big)h(x) + \exp\big(\eta^\top T(x) - A(\eta)\big)h'(x)\Big)\,d\mu(x) \\
&\quad= \int_{\mathbb{R}} g(x)\Big(\eta^\top T'(x) + \frac{h'(x)}{h(x)}\Big)\exp\big(\eta^\top T(x) - A(\eta)\big)h(x)\,d\mu(x)
\end{aligned}$$

which is equal to the LHS.

For $X \sim \mathsf{N}(\mu, \sigma^2)$, we have $\eta_1 = \frac{\mu}{\sigma^2}$, $\eta_2 = \frac{-1}{2\sigma^2}$, $T_1(x) = x$, $T_2(x) = x^2$, and $h(X) = \frac{1}{\sqrt{2\pi}}$, so the conclusion follows. □

For $X \sim \mathsf{N}(\mu, \sigma^2)$, setting $g(x) = 1$ gives $\mathbb{E}[X] = \mu$, and setting $g(x) = x$ gives $\mathbb{E}[X^2] = \sigma^2 + \mu^2$.

1.3 Sufficient statistics

1.3.1 Definitions and examples

For $X \sim \mathcal{P}_\theta$ and a statistic $T = T(X)$, the following are equivalent definitions or characterizations of the sufficiency of T (in which case we call T a sufficient statistic):

- The conditional distribution of X given T does not depend on θ.
- Given T, it is possible to construct a random variable X' having the same distribution as X.
- (Fisher–Neyman) There exist nonnegative functions g_θ and h such that $f(x \mid \theta) = g_\theta(T(x)) \cdot h(x)$. This is called the factorization criterion.

Remarks:

- Obviously, X itself is a sufficient statistic.
- If T is a sufficient statistic and there exists a function g such that $T = g(S)$, then S is a sufficient statistic.
- A sufficient statistic T is minimal if for any sufficient statistic S, there exists a function g such that $T = g(S)$.

- If sufficient statistics S and T are functions of each other, then we say that they are equivalent.

Examples of sufficient statistics:

- Let X have a symmetric distribution on $\mathbb{R}$. Then $T = |X|$ is a sufficient statistic.
- Let $X_1, \dots, X_n$ be i.i.d. random variables sampled from a distribution on $\mathbb{R}$. Then the set of order statistics $T = (X_{(1)}, \dots, X_{(n)})$ is sufficient.
- Consider i.i.d. $X_1, \dots, X_n \sim \mathsf{Unif}(0, \theta)$. Then $T = X_{(n)} = \max_{i \in [n]} X_i$ is a sufficient statistic by the Fisher–Neyman criterion:

$$\begin{aligned} f(x_1, \dots, x_n \mid \theta) &= (1/\theta)^n \mathbb{1}\{0 \le x_i \le \theta \text{ for all } i \in [n]\} \\ &= (1/\theta)^n \mathbb{1}\{x_{(n)} \le \theta\} \cdot \mathbb{1}\{x_i \ge 0 \text{ for all } i \in [n]\} = g_\theta(T) \cdot h(x_1, \dots, x_n). \end{aligned}$$

- Consider i.i.d. $X_1, \dots, X_n \sim \mathsf{Poi}(\lambda)$. Then $T = \sum_{i=1}^n X_i$ is a sufficient statistic, since

$$f(x_1, \dots, x_n \mid \lambda) = \lambda^{\sum_i x_i} e^{-n\lambda} / \textstyle\prod_i (x_i!).$$

- Consider i.i.d. $X_1, \dots, X_n \sim \mathsf{N}(\mu, \sigma^2)$. A set of sufficient statistics is $T = (\sum_i X_i, \sum_i X_i^2)$ or $T' = (\hat{\mu}, \hat{\sigma}^2)$ where $\hat{\mu} := \sum_i X_i / n$ and $\hat{\sigma}^2 := \sum_i (X_i - \hat{\mu})^2$, since

$$f(x_1, \dots, x_n \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(\frac{-1}{2\sigma^2} \sum_i x_i^2 + \frac{\mu}{\sigma^2} \sum_i x_i - \frac{n}{2\sigma^2}\mu^2\Big).$$

- Let $\mathcal{P}_\theta$ be from an exponential family with density $f(x \mid \theta) = \exp\big(\eta^\top T(x) - A(\eta)\big) \cdot h(x)$, where $\eta(\theta), T(x) \in \mathbb{R}^m$. For i.i.d. $X_1, \dots, X_n \sim \mathcal{P}_\theta$, the distribution of $(X_1, \dots, X_n)$ has density

$$\exp\Big(\eta^\top \big(\textstyle\sum_i T(x_i)\big) - nA(\eta)\Big) \cdot \textstyle\prod_i h(x_i),$$

  where $\sum_i T(X_i) \in \mathbb{R}^m$ is a sufficient statistic.

Although there is no loss statistically (or information-theoretically) in retaining only sufficient statistics, it may not be computationally favorable to do so. See the excellent paper [Mon15]. (It is not even obvious how to generate n i.i.d. Gaussians from the empirical mean and variance.)

1.3.2 Some results

Lemma 1.3. *Consider a family of distributions $\{\mathcal{P}_\theta : \theta \in \Theta\}$ where every $\mathcal{P}_\theta$ is absolutely continuous with respect to μ. A statistic S is sufficient if and only if for any $\theta, \theta_0 \in \Theta$, the ratio $f(x \mid \theta)/f(x \mid \theta_0)$ is a function only of $S(x)$.*

Proof. This follows immediately from the factorization criterion. □

Lemma 1.4. *Consider a finite family of distributions with densities $f_0, f_1, \dots, f_k$, all having the same support. Then the statistic $T = \big(\frac{f_1}{f_0}, \frac{f_2}{f_0}, \dots, \frac{f_k}{f_0}\big)$ is minimal sufficient.*

Proof. We need to show that for any sufficient statistic S, there exists a function g such that $T = g(S)$. This follows immediately from the previous result. □

Lemma 1.5. *Let $\mathscr{P}$ be a family of distributions with common support and $\mathscr{P}' \subset \mathscr{P}$. If a statistic T is minimal sufficient for $\mathscr{P}'$ and sufficient for $\mathscr{P}$, then it is minimal sufficient for $\mathscr{P}$.*

Proof. If S is a sufficient statistic for $\mathscr{P}$, it is a sufficient statistic for $\mathscr{P}'$. Hence there exists a function g such that $T = g(S)$. □

Consider an exponential family with density $f(x \mid \theta) = \exp\big(\eta^\top T(x) - A(\eta)\big) \cdot h(x)$, where $\theta \in \Theta$. If the interior of the set $\eta(\Theta) \subset \mathbb{R}^m$ is not empty and if T does not satisfy an affine constraint $v^\top T = c$ for nonzero $v \in \mathbb{R}^m$ and $c \in \mathbb{R}$, then the exponential family is said to be of full rank.

Theorem 1.6. *Consider an exponential family with density $f(x \mid \theta) = \exp\big(\eta^\top T(x) - A(\eta)\big) \cdot h(x)$, where $\theta \in \Theta$ and $\eta = \eta(\theta) \in \mathbb{R}^m$. Suppose that T does not satisfy an affine constraint of the form $v^\top T = c$. If there exist natural parameters $\eta^{(0)}, \eta^{(1)}, \dots, \eta^{(m)}$ such that $\{\eta^{(i)} - \eta^{(0)} : i \in [m]\}$ spans $\mathbb{R}^m$, then T is minimal sufficient.*

In particular, the sufficient statistic T in a full-rank exponential family $\mathscr{P}$ is minimal.

Proof. Let $\mathscr{P}' \subset \mathscr{P}$ be a subfamily consisting of $m+1$ distributions, with natural parameters $\eta^{(0)}, \eta^{(1)}, \dots, \eta^{(m)}$. By Lemma 1.4, a minimal sufficient statistic for $\mathscr{P}'$ is

$$\mathsf{T}' = \Big(\exp\big((\eta^{(1)} - \eta^{(0)})^\top T\big) - A(\eta^{(1)}) + A(\eta^{(0)}), \dots, \cdot\Big),$$

$$\mathsf{T} = \Big((\eta^{(1)} - \eta^{(0)})^\top T, \dots, (\eta^{(m)} - \eta^{(0)})^\top T\Big).$$

This is equivalent to T if and only if the matrix with columns $\{\eta^{(i)} - \eta^{(0)} : i \in [m]\}$ is nonsingular. Conclude using Lemma 1.5. Such a subfamily can be chosen if the exponential family is full-rank. □

1.4 Convexity, maximum entropy, Rao–Blackwell

Here are some basic facts about convex functions:

- A function $f : (a, b) \to \mathbb{R}$ is convex if and only if its epigraph is a convex set.
- If $f : (a, b) \to \mathbb{R}$ is convex, it is continuous on (a, b), and has a left and right derivative at every point in (a, b).
- If f is differentiable on (a, b), then f is convex if and only if f' is nondecreasing.
- If f is twice differentiable on (a, b), then f is convex if and only if $f'' \geq 0$.

Proposition 1.7. *Consider a real-valued convex function f on a convex open set $S \subset \mathbb{R}^n$. At each $x \in S$, there exists a vector $v \in \mathbb{R}^n$ such that $f(y) - f(x) \geq v^\top (y - x)$ for any $y \in S$. This vector v is called a subgradient of f at x.*

If f is strictly convex, then v can be chosen so that the inequality is strict unless $y = x$.

Proposition 1.8 (Jensen's inequality)**.** *Consider a real-valued convex function f on a convex set $S \subset \mathbb{R}^n$. For any $x_1, \dots, x_n \in S$ and $a_1, \dots, a_n \in [0, 1]$ such that $\sum_i a_i = 1$, we have $f(\sum_i a_i x_i) \leq \sum_i a_i f(x_i)$.*

More generally, if X is a random variable taking values in $S \subset \mathbb{R}^n$ and $\mathbb{E}[X] < \infty$, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$. If f is strictly convex, the inequality is strict unless $\mathbb{P}\{X = \mathbb{E}[X]\} = 1$.

An example is $e^{\lambda\,\mathbb{E}[X]} \le \mathbb{E}[e^{\lambda X}]$.

Proof. Define $L(y) = f(x) + v^\top(y - x) \le f(y)$ for $x = \mathbb{E}[X]$. Then $\mathbb{E}[f(X)] \ge \mathbb{E}[L(X)] = L(\mathbb{E}[X]) = f(\mathbb{E}[X])$. □

The entropy of a random variable $X \sim \mathcal{P}$ with density p is defined as

$$H(X) = H(\mathcal{P}) := \mathbb{E}_{\mathcal{P}}[-\log p(X)].$$

In the case of a continuous distribution, the entropy is also called the differential entropy.

Consider probability distributions $\mathcal{P}$ and $\mathcal{Q}$ on $\mathcal{X}$ with densities p and q respectively, such that $\mathcal{P}$ is absolutely continuous with respect to Q. The relative entropy/entropy distance/Kullback-Leibler divergence between them is

$$\mathsf{KL}(\mathcal{P}, \mathcal{Q}) = D(\mathcal{P} \,\|\, \mathcal{Q}) := \mathbb{E}_{\mathcal{P}}\Big[\log \frac{p(X)}{q(X)}\Big] = \int p(x) \log \frac{p(x)}{q(x)}\, d\mu(x) \ge 0.$$

The KL divergence is not symmetric, but is zero if and only if $p = q$ almost everywhere:

$$\mathbb{E}_{\mathcal{P}}\Big[\log \frac{p(X)}{q(X)}\Big] = \mathbb{E}_{\mathcal{Q}}\Big[\frac{p(X)}{q(X)} \log \frac{p(X)}{q(X)}\Big] = \mathbb{E}_{\mathcal{Q}}\Big[f\Big(\frac{p(X)}{q(X)}\Big)\Big] \ge f\Big(\mathbb{E}_{\mathcal{Q}}\Big[\frac{p(X)}{q(X)}\Big]\Big) = f(1) = 0,$$

where $f(x) = x \log x$ is convex.

Theorem 1.9 (Maximum entropy principle)**.** *Consider a random variable X, a vector of statistics $T = T(X) \in \mathbb{R}^m$, and a fixed vector $\mu \in \mathbb{R}^m$ (which is a value that T may take). Denote by $\mathcal{P}^*$ the solution to the optimization problem:*

$$\max_{\mathcal{P}} H(\mathcal{P}) \qquad \text{s.t. } X \sim \mathcal{P},\ \mathbb{E}[T(X)] = \mu.$$

Then $\mathcal{P}^*$ is from an exponential family with density $C \exp\big(\theta^\top T(x) - A(\theta)\big)$, where $\theta = \theta(\mu)$.

Proof. Consider $\mathcal{P}$ and $\mathcal{P}_\theta$ such that $\mathbb{E}_{\mathcal{P}}[T(X)] = \mathbb{E}_{\mathcal{P}_\theta}[T(X)] = \mu$, with densities p and $p_\theta(x) = C \exp\big(\theta^\top T(x) - A(\theta)\big)$ respectively. Then

$$\begin{aligned}
H(\mathcal{P}) &= -\,\mathbb{E}_{\mathcal{P}}[\log p(X)] = -\,\mathbb{E}_{\mathcal{P}}\Big[\log \frac{p(X)}{p_\theta(X)}\Big] - \mathbb{E}_{\mathcal{P}}[\log p_\theta(X)] \\
&= -\,\mathsf{KL}(\mathcal{P}, \mathcal{P}_\theta) - \mathbb{E}_{\mathcal{P}}\big[\theta^\top T(X) - A(\theta) + \log C\big] \\
&\le -\,\mathbb{E}_{\mathcal{P}_\theta}\big[\theta^\top T(X) - A(\theta) + \log C\big] \\
&= -\,\mathbb{E}_{\mathcal{P}_\theta}[\log p_\theta(X)] = H(\mathcal{P}_\theta),
\end{aligned}$$

where the inequality holds because $\mathsf{KL}(\mathcal{P}, \mathcal{P}_\theta) \ge 0$ and $\mathbb{E}_{\mathcal{P}}[T(X)] = \mathbb{E}_{\mathcal{P}_\theta}[T(X)] = \mu$. □

Recall that for an estimator $\hat{g}$ and a loss function $L(g, \hat{g})$, the risk is the expected loss $R(g, \hat{g}) = \mathbb{E}[L(g, \hat{g})]$.

Theorem 1.10 (Rao–Blackwell)**.** *Consider $X \sim \mathcal{P}_\theta$ from the family $\{\mathcal{P}_\theta : \theta \in \Theta\}$ and a sufficient statistic $T = T(X)$. Suppose that the loss function $L(g, \cdot)$ is convex in the second variable. Moreover, consider an estimator $\hat{g} = \hat{g}(X)$ such that $\mathbb{E}[\hat{g}(X)] < \infty$ and $R(g, \hat{g}) < \infty$. Define an estimator $\tilde{g} = \tilde{g}(T) = \mathbb{E}[\hat{g}(X) \mid T]$. Then we have $R(g, \tilde{g}) \le R(g, \hat{g})$.*

If $L(g, \cdot)$ is strictly convex in the second variable, then the above inequality is strict unless $\mathbb{P}\{\hat{g} = \tilde{g}\} = 1$.

Proof. By Jensen's inequality,

$$L(g, \tilde{g}) = L\big(g, \mathbb{E}[\hat{g}(X) \mid T]\big) \le \mathbb{E}\left[L\big(g, \hat{g}(X)\big) \mid T\right].$$

Taking the expectation on both sides with respect to T yields the result. □

1.5 Bias, variance, MVUE

1.5.1 Theory

Consider $X \sim \mathcal{P}_\theta$. Let $g(\theta)$ be an estimand, and let $\hat{g}(X)$ be an estimator of $g(\theta)$. The bias of $\hat{g}$ is $\mathbb{E}[\hat{g}(X)] - g(\theta)$. The variance of $\hat{g}$ is $\mathsf{Var}\big(\hat{g}(X)\big)$. An estimator is unbiased if $\mathbb{E}[\hat{g}(X)] = g(\theta)$. The bias-variance decomposition for the squared loss refers to the following

$$\begin{aligned}
&\mathbb{E}\left(\hat{g}(X) - g(\theta)\right)^2 \\
&= \mathbb{E}\left(\hat{g}(X) - \mathbb{E}\,\hat{g}(X) + \mathbb{E}\,\hat{g}(X) - g(\theta)\right)^2 \\
&= \mathbb{E}\left(\hat{g}(X) - \mathbb{E}\,\hat{g}(X)\right)^2 + 2\,\mathbb{E}\left(\hat{g}(X) - \mathbb{E}\,\hat{g}(X)\right)\left(\mathbb{E}\,\hat{g}(X) - g(\theta)\right) + \left(\mathbb{E}\,\hat{g}(X) - g(\theta)\right)^2 \\
&= \mathsf{Var}\left(\hat{g}(X)\right) + \mathsf{Bias}\left(\hat{g}(X)\right)^2.
\end{aligned}$$

An unbiased estimator $\hat{g}(X)$ is called the uniformly minimum-variance unbiased estimator (MVUE) of $g(\theta)$ if $\mathsf{Var}_\theta\left(\hat{g}(X)\right) \le \mathsf{Var}_\theta\left(\tilde{g}(X)\right)$ for all $\theta \in \Theta$ for any unbiased estimator $\tilde{g}(X)$.

Theorem 1.11. *Consider $X \sim \mathcal{P}_\theta$ where $\theta \in \Theta$, and let $\hat{g}(X)$ be an estimator of $g(\theta)$ such that $\mathbb{E}_\theta\,\hat{g}^2 < \infty$ for all $\theta \in \Theta$. Let $\mathcal{U}$ denote the set of $U = U(X)$ such that $\mathbb{E}_\theta\,U = 0$ and $\mathbb{E}_\theta\,U^2 < \infty$ for all $\theta \in \Theta$. Then $\hat{g}(X)$ is the MVUE if and only if it is unbiased and*

$$\mathsf{Cov}(\hat{g}, U) = \mathbb{E}_\theta\left[\hat{g}\,U\right] = 0 \text{ for all } U \in \mathcal{U} \text{ and all } \theta \in \Theta.$$

Interpretation: If U is "irrelevant" for estimating $g(\theta)$, then the MVUE $\hat{g}$ is orthogonal to U. In addition, for any estimator $\tilde{g}$, the MVUE is $\hat{g} = \tilde{g} - \tilde{U}$ where $\tilde{U}$ is the orthogonal projection of $\tilde{g}$ onto $\mathcal{U}$.

Proof. "⇒": Fix such a U and $\theta \in \Theta$. For any $\lambda \in \mathbb{R}$, $\tilde{g} = \hat{g} + \lambda U$ is an unbiased estimator. Since $\hat{g}$ is the MVUE by assumption, we have

$$\mathsf{Var}(\hat{g}) \le \mathsf{Var}(\hat{g} + \lambda U) = \mathsf{Var}(\hat{g}) + 2\lambda\,\mathsf{Cov}(\hat{g}, U) + \lambda^2\,\mathsf{Var}(U).$$

This is violated at $\lambda = -\,\mathsf{Cov}(\hat{g}, U)/\,\mathsf{Var}(U)$ unless $\mathsf{Cov}(\hat{g}, U) = 0$.

"⇐": Let $\tilde{g}(X)$ be an unbiased estimator with $\mathbb{E}\,\tilde{g}^2 < \infty$. Then $U := \hat{g} - \tilde{g}$ has zero mean and finite variance, so $\mathbb{E}[\hat{g}(\hat{g} - \tilde{g})] = 0$ by assumption. This implies

$$\begin{aligned}
\mathsf{Var}(\hat{g}) = \mathbb{E}\,\hat{g}^2 - (\mathbb{E}\,\hat{g})^2 &= \mathbb{E}[\hat{g}\tilde{g}] - (\mathbb{E}\,\hat{g})(\mathbb{E}\,\tilde{g}) = \mathbb{E}(\hat{g} - \mathbb{E}\,\hat{g})(\tilde{g} - \mathbb{E}\,\tilde{g}) \\
&\le \sqrt{\mathbb{E}(\hat{g} - \mathbb{E}\,\hat{g})^2}\sqrt{\mathbb{E}(\tilde{g} - \mathbb{E}\,\tilde{g})^2} = \sqrt{\mathsf{Var}(\hat{g})\,\mathsf{Var}(\tilde{g})}.
\end{aligned}$$

Hence $\mathsf{Var}(\hat{g}) \le \mathsf{Var}(\tilde{g})$. □

A statistic $T = T(X)$ is called complete if

$$\mathbb{E}_\theta[f(T)] = 0 \text{ for all } \theta \in \Theta \quad \text{implies} \quad f(t) = 0.$$

Theorem 1.12. *If $X \sim \mathcal{P}_\theta$ for a full-rank exponential family $\{\mathcal{P}_\theta\}$, then T is complete.*

For a proof, see Theorem 4.3.1 of [LR06]. The above result leads to an important theorem by Lehmann and Scheffé.

Theorem 1.13 (Lehmann–Scheffé)**.** *Consider $X \sim \mathcal{P}_\theta$, and let T be a complete sufficient statistic for $\{\mathcal{P}_\theta : \theta \in \Theta\}$. Suppose that $\tilde{g}(X)$ is an unbiased estimator of $g(\theta)$. Define $\hat{g}(T)$ by $\hat{g}(t) = \mathbb{E}[\tilde{g}(X) \mid T = t]$ (as in the Rao–Blackwell theorem). Then $\hat{g}(T)$*

- *is an unbiased estimator of $g(\theta)$;*
- *is the only unbiased estimator that is a function of T;*
- *uniformly minimizes the risk for any loss $L(g, \cdot)$ convex in the second variable;*
- *is the MVUE.*

Proof. We check that $\mathbb{E}[\hat{g}(T)] = \mathbb{E}[\tilde{g}(X)] = g(\theta)$, so $\hat{g}(T)$ is unbiased.

For uniqueness, let $\hat{g}(T)$ and $\tilde{g}(T)$ be two unbiased estimators. Then $\mathbb{E}[\hat{g}(T) - \tilde{g}(T)] = 0$, so by completeness we have $\hat{g} = \tilde{g}$.

The Rao–Blackwell theorem implies that $\mathbb{E}\, L\big(g(\theta), \hat{g}(T)\big) \le \mathbb{E}\, L\big(g(\theta), \tilde{g}(X)\big)$ for any $\tilde{g}(X)$.

Taking $L(x, y) = (x - y)^2$, we have $\mathsf{Var}(\hat{g}(T)) \le \mathsf{Var}(\tilde{g}(X))$. □

1.5.2 Examples

- Gaussian MVUE: Consider i.i.d. $X_1, \dots, X_n \sim \mathsf{N}(\mu, \sigma^2)$, where μ and σ are unknown. Recall that the empirical mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ and empirical vairance $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu})^2$ are jointly sufficient statistics. They are also complete by full-rankness. Since $\mathbb{E}[\hat{\mu}] = \mu$ and $\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}(X_1 - \hat{\mu})^2 = \frac{n-1}{n}\sigma^2$, we have that $(\hat{\mu}, \frac{n}{n-1}\hat{\sigma}^2)$ is the MVUE of (μ, σ^2).

- Uniform MVUE: Consider i.i.d. $X_1, \dots, X_n \sim \mathsf{Unif}(0, \theta)$ and $g(\theta) = \theta$. It can be shown that the sufficient statistic $T = X_{(n)}$ is complete. Note that $2X_1$ is an unbiased estimator of θ, so the MVUE is
$$\hat{g}(t) = \mathbb{E}[2X_1 \mid X_{(n)} = t] = \frac{1}{n} \cdot 2t + \frac{n-1}{n} \cdot t = \frac{t(n+1)}{n}.$$

- Binomial MVUE: Consider $X \sim \mathsf{Bin}(N, p)$ and $g(p) = p(1-p)$. Recall that $T(X) = X$. Then $\mathbb{E}\, \hat{g}(T) = g(p)$ says
$$\sum_{x=0}^{N} \binom{N}{x} \hat{g}(x) p^x (1-p)^{N-x} = p(1-p).$$
Let $r = p/(1-p)$. Then we have $p = r/(1+r)$ and $1 - p = 1/(1+r)$. Hence
$$\sum_{x=0}^{N} \binom{N}{x} \hat{g}(x) r^x = p(1-p)^{1-N} = r(1+r)^{N-2} = \sum_{x=1}^{N-1} \binom{N-2}{x-1} r^x,$$
which holds for $p \in (0, 1)$ or $r \in (0, \infty)$. Thus we can take
$$\hat{g}(x) = \binom{N-2}{x-1} \binom{N}{x}^{-1} = \frac{x(N-x)}{N(N-1)}.$$

1.6 Lower bounds on the variance of an unbiased estimator

Let us assume a few technical conditions throughout this section:

- $\Theta \subset \mathbb{R}$ and Θ is an open interval;
- The support $\{x : f(x \mid \theta) > 0\}$ is independent of θ;
- $\frac{\partial f(x|\theta)}{\partial \theta}$ exists and is finite for all x and θ;
- Differentiation under the integral sign works.

1.6.1 Lower bounds and the Fisher information

For an estimator $\hat{g}(X)$ with $\mathbb{E}_\theta\, \hat{g} = g(\theta)$ and $\mathbb{E}_\theta\, \hat{g}^2 < \infty$, we now provide three lower bounds on $\mathsf{Var}_\theta(\hat{g}(X))$.

1. (Cauchy–Schwarz) For any function $\phi(x, \theta)$ with $\mathbb{E}_\theta[\phi(X, \theta)^2] < \infty$,

$$\mathsf{Var}_\theta(\hat{g}) \geq \frac{\mathsf{Cov}_\theta(\hat{g}, \phi)^2}{\mathsf{Var}_\theta(\phi)}. \tag{1.2}$$

   The problem with this simple bound is that the right-hand side depends on the estimator $\hat{g}$.

2. (Hammersley–Chapman–Robbins inequality) Let us choose $\phi(x, \theta) = \frac{f(x|\theta+\delta)}{f(x|\theta)} - 1$. Then we have $\mathsf{Cov}_\theta(\hat{g}, \phi) = \mathbb{E}_\theta[\hat{g}\phi] = \mathbb{E}_{\theta+\delta}[\hat{g}] - \mathbb{E}_\theta[\hat{g}] = g(\theta + \delta) - g(\theta)$. Hence

$$\mathsf{Var}_\theta(\hat{g}) \geq \frac{\big(g(\theta + \delta) - g(\theta)\big)^2}{\mathbb{E}_\theta\left(\frac{f(X|\theta+\delta)}{f(X|\theta)} - 1\right)^2}.$$

3. (Cramér–Rao) Suppose that there exists a function $B(x, \theta)$ and $\varepsilon > 0$ such that

$$\mathbb{E}_\theta[B(X, \theta)^2] < \infty \quad \text{and} \quad \Big|\frac{f(x \mid \theta + \delta) - f(x \mid \theta)}{\delta f(x \mid \theta)}\Big| \leq B(x, \theta) \text{ for all } |\delta| \leq \varepsilon.$$

   If g is differentiable, taking the limit $\delta \to 0$ on the right-hand side of the above inequality, and applying dominated convergence, we obtain

$$\mathsf{Var}_\theta(\hat{g}) \geq \frac{\big(g'(\theta)\big)^2}{\mathbb{E}_\theta\left(\frac{\partial f(X|\theta)/\partial\theta}{f(X|\theta)}\right)^2} = \frac{\big(g'(\theta)\big)^2}{I(\theta)}.$$

   Here $I(\theta)$ is the Fisher information that X contains about θ, defined by

$$I(\theta) = \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta)\right)^2.$$

Since $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta} \log f(X \mid \theta)\right] = 0$, we have

$$I(\theta) = \mathsf{Var}_\theta\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta)\right).$$

If, in addition, $\frac{\partial^2}{\partial\theta^2}\log f(x \mid \theta)$ exists for all x and θ and differentiation under the integral sign holds, then taking the expectation of $\frac{\partial^2}{\partial\theta^2}\log f(x \mid \theta) = \frac{\frac{\partial^2}{\partial\theta^2}f(x|\theta)}{f(x|\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right)^2$ yields

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right].$$

Note that if θ is differentiable function of ξ, the Fisher information X contains about ξ is

$$\tilde{I}(\xi) = I(\theta)\cdot\big(\theta'(\xi)\big)^2.$$

1.6.2 Extensions

- (i.i.d. observations) By definition, we can check:

  Lemma 1.14. *Let X_1 and X_2 be independent random variables with densities $f_1(x \mid \theta)$ and $f_2(x \mid \theta)$ respectively. If $I_1(\theta)$, $I_2(\theta)$ and $I(\theta)$ denote the information X_1, X_2 and (X_1, X_2) contain about θ, then $I(\theta) = I_1(\theta) + I_2(\theta)$.*

  Therefore, if we observe i.i.d. $X_1, \dots, X_n \sim \mathcal{P}_\theta$, then

  $$\mathsf{Var}_\theta(\hat{g}) \geq \frac{\big(g'(\theta)\big)^2}{nI_1(\theta)}.$$

- (Biased case) If $\hat{g}$ is a biased estimator with $\mathbb{E}_\theta\,\hat{g} = g(\theta) + b(\theta)$, then the same argument yields

  $$\mathsf{Var}_\theta(\hat{g}) \geq \frac{\big(g'(\theta) + b'(\theta)\big)^2}{I(\theta)}.$$

- (Multivariate case) Consider $\theta \in \mathbb{R}^m$. Analogous to (1.2), we have the following result.

  Theorem 1.15. *Consider an unbiased estimator $\hat{g}$ and functions $\phi_i(x, \theta)$ with finite second moments where $i \in [m]$. Define $\gamma \in \mathbb{R}^m$ by $\gamma_i = \mathsf{Cov}(\hat{g}, \phi_i)$, and define $C \in \mathbb{R}^{m\times m}$ by $C_{ij} = \mathsf{Cov}(\phi_i, \phi_j)$. Then*

  $$\mathsf{Var}(\hat{g}) \geq \gamma^\top C^{-1}\gamma.$$

  Under some regularity conditions similar to the one-dimensional case, the information matrix $I \in \mathbb{R}^{m\times m}$ is defined by

  $$\begin{aligned} I_{ij}(\theta) &= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i}\log f(X \mid \theta)\cdot\frac{\partial}{\partial\theta_j}\log f(X \mid \theta)\right] \\ &= \mathsf{Cov}_\theta\left(\frac{\partial}{\partial\theta_i}\log f(X \mid \theta), \frac{\partial}{\partial\theta_j}\log f(X \mid \theta)\right) \\ &= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X \mid \theta)\right]. \end{aligned}$$

  Hence $I(\theta) = -\mathbb{E}_\theta[\nabla^2 \log f(X \mid \theta)]$.

  Theorem 1.16 (Cramér–Rao, Information Inequality)**.** *Assume mild regularity conditions (similar to the one-dimensional case) and that $I(\theta)$ is positive definite. Define $\alpha \in \mathbb{R}^m$ by $\alpha_i = \frac{\partial}{\partial\theta_i}\mathbb{E}_\theta\,\hat{g}$. Then we have*

  $$\mathsf{Var}_\theta(\hat{g}) \geq \alpha^\top I(\theta)^{-1}\alpha.$$

1.6.3 Examples

- Exponential family: $X \sim \mathcal{P}_\eta$ with $f(x \mid \eta) = \exp\big(\eta^\top T(x) - A(\eta)\big)h(x)$. Then $\nabla^2 \log f(x \mid \eta) = -\nabla^2 A(\eta)$, so $I(\eta) = \nabla^2 A(\eta)$.

- Gaussian: $X \sim \mathsf{N}(\mu, \sigma^2)$ where σ is known. Then $f(x \mid \mu) = \exp\big(\frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\sigma^2}\big)\frac{1}{\sqrt{2\pi}\sigma}$ and thus $\frac{\partial}{\partial \mu} \log f(x \mid \mu) = x/\sigma^2 - \mu/\sigma^2$. It follows $I(\mu) = 1/\sigma^2$ and $I(\mu^2) = I(\mu)\big(\frac{1}{2\mu}\big)^2 = \frac{1}{4\mu^2\sigma^2}$.

- Poisson: $X \sim \mathsf{Poi}(\lambda)$ so that $f(x \mid \lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$. Hence $\frac{\partial}{\partial \lambda} \log f(x \mid \lambda) = x/\lambda - 1$. It follows that $I(\lambda) = 1/\lambda$. However, by a change of variable, $I(\log \lambda) = I(\lambda)\big(\frac{d}{d\xi}e^\xi|_{\xi=\log\lambda}\big)^2 = \lambda$.

- Binomial: $X \sim \mathsf{Bin}(N, p)$ where N is known. Then $\frac{\partial}{\partial p} \log f(x \mid p) = x/p - (N-x)/(1-p)$, so $I(p) = Np(1-p)[1/p + 1/(1-p)]^2 = \frac{N}{p(1-p)}$. Then $\mathsf{Var}(X/N) = p(1-p)/N$ and the equality is attained in the Cramér–Rao bound.

Chapter 2

Bayesian versus minimax

2.1 Bayesian estimation

Consider $X \sim \mathcal{P}_\theta$ with density $f(x \mid \theta)$. Suppose that $\theta \sim \pi$ where π is the prior distribution with density $p(\theta)$. The marginal distribution of X has density

$$f(x) = \int f(x \mid \theta) \cdot p(\theta)\, d\mu(\theta).$$

The posterior distribution of θ refers to the conditional distribution with density

$$p(\theta \mid x) = \frac{f(x \mid \theta)}{f(x)} p(\theta).$$

2.1.1 Bayes risk and Bayes estimator

For a loss function $L(g, \hat{g})$, recall that the risk is $R(g, \hat{g}) := \mathbb{E}_{X \sim \mathcal{P}_\theta} L\big(g(\theta), \hat{g}(X)\big)$. The Bayes risk is defined as

$$R_\pi(\hat{g}) := \mathbb{E}_{\theta \sim \pi} R(g, \hat{g}).$$

An estimator $\hat{g}$ is called a Bayes (optimal) estimator if $R_\pi(\hat{g}) \le R_\pi(\tilde{g})$ for any other estimator $\tilde{g}$. A stronger condition is that the estimator $\hat{g}$ minimizes the posterior loss

$$\mathbb{E}_{\theta \sim \pi} \left[L\big(g(\theta), \hat{g}(X)\big) \mid X \right].$$

- For the squared loss $L(g, \tilde{g}) = (g - \tilde{g})^2$, the posterior mean $\hat{g}(X) := \mathbb{E}[g(\theta) \mid X]$ is a Bayes estimator because for any estimator $\tilde{g}$,

  $$\mathbb{E}[(g(\theta) - \tilde{g}(X))^2 \mid X] = \mathbb{E}[(g(\theta) - \hat{g}(X))^2 \mid X] + (\hat{g}(X) - \tilde{g}(X))^2.$$

  Note that we have

  $$\mathbb{E}[g(\theta) \mid X] = \int g(\theta) p(\theta \mid X)\, d\theta = \int g(\theta) \frac{f(X \mid \theta)}{f(X)} p(\theta)\, d\theta = \frac{\int g(\theta) f(X \mid \theta) p(\theta)\, d\theta}{\int f(X \mid \theta) p(\theta) d\theta}.$$

- For the ℓ_1 loss $L(g, \tilde{g}) = |g - \tilde{g}|$, it is not hard to check that the posterior median is a Bayes estimator.

2.1.2 Examples

Gaussian Consider $X \sim \mathsf{N}(\theta, 1)$ where $\theta \sim \mathsf{N}(0, \sigma^2)$. The Bayes estimator under the squared loss $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ is the posterior mean

$$\hat{\theta}(X) := \mathbb{E}[\theta \mid X] = \frac{\int \theta f(X \mid \theta)\, p(\theta)\, d\theta}{\int f(X \mid \theta)\, p(\theta) d\theta}.$$

One straightforward but tedious way to obtain $\hat{\theta}(X)$ is to compute

$$\int f(X \mid \theta)\, p(\theta) d\theta = \frac{1}{2\pi\sigma} \int e^{-(X-\theta)^2/2 - \theta^2/(2\sigma^2)}\, d\theta = \frac{1}{\sqrt{2\pi(1+\sigma^2)}} \exp\Big(\frac{-X^2}{2(1+\sigma^2)}\Big),$$

and

$$\int \theta f(X \mid \theta)\, p(\theta) d\theta = \frac{1}{2\pi\sigma} \int \theta e^{-(X-\theta)^2/2 - \theta^2/(2\sigma^2)}\, d\theta = \frac{\sigma^2 X}{\sqrt{2\pi(1+\sigma^2)^3}} \exp\Big(\frac{-X^2}{2(1+\sigma^2)}\Big).$$

Therefore,

$$\hat{\theta}(X) = \frac{\sigma^2}{1+\sigma^2} X.$$

Alternatively, we may note that the posterior density of θ is

$$p(\theta \mid X) = \frac{f(X \mid \theta)\, p(\theta)}{f(X)} = \frac{f(X \mid \theta)\, p(\theta)}{\int f(X \mid \theta)\, p(\theta) d\theta},$$

where

$$f(X \mid \theta)\, p(\theta) = \frac{1}{2\pi\sigma} e^{-(X-\theta)^2/2 - \theta^2/(2\sigma^2)} = h_\sigma(X) \cdot \exp\left(-\frac{\sigma^2+1}{2\sigma^2} \left(\theta - \frac{\sigma^2}{1+\sigma^2} X \right)^2 \right)$$

for some inessential function $h_\sigma(X)$. Since the above quantity is proportional to the PDF of the Gaussian $\theta \sim \mathsf{N}\left(\frac{\sigma^2}{1+\sigma^2} X, \frac{\sigma^2}{1+\sigma^2}\right)$, we conclude that this Gaussian is the posterior distribution of θ conditional on X. Therefore, the posterior mean is $\frac{\sigma^2}{1+\sigma^2} X$.

The risk of $\hat{\theta}$ is

$$\mathbb{E}_{X \sim \mathsf{N}(\theta,1)} \left(\frac{\sigma^2}{1+\sigma^2} X - \theta \right)^2 = \left(\frac{\sigma^2}{1+\sigma^2} \right)^2 + \left(\frac{1}{1+\sigma^2} \right)^2 \theta^2,$$

and the Bayes risk is

$$\mathbb{E}_{\theta \sim \mathsf{N}(0,\sigma^2)} \left[\left(\frac{\sigma^2}{1+\sigma^2} \right)^2 + \left(\frac{1}{1+\sigma^2} \right)^2 \theta^2 \right] = \left(\frac{\sigma^2}{1+\sigma^2} \right)^2 + \left(\frac{\sigma}{1+\sigma^2} \right)^2 = \frac{\sigma^2}{1+\sigma^2}.$$

i.i.d. Gaussians Suppose that we observe i.i.d. $X_1, \ldots, X_n \sim \mathsf{N}(\theta, \tau^2)$ where $\theta \sim \mathsf{N}(\mu, \sigma^2)$. By the same argument, it suffices to consider

$$\begin{aligned} \prod_{i=1}^n f(X_i \mid \theta)\, p(\theta) &= \frac{1}{(2\pi)^{(n+1)/2}\, \tau^n \sigma} \exp\left(-\sum_{i=1}^n \frac{(X_i - \theta)^2}{2\tau^2} - \frac{(\theta - \mu)^2}{2\sigma^2} \right) \\ &= h_{\mu,\tau,\sigma}(X) \cdot \exp\left(-\frac{n\sigma^2 + \tau^2}{2\sigma^2\tau^2} \left(\theta - \frac{n\sigma^2}{\tau^2 + n\sigma^2} \bar{X} - \frac{\tau^2}{\tau^2 + n\sigma^2} \mu \right)^2 \right). \end{aligned}$$

We then conclude that the posterior distribution of θ conditional on $X_1, \ldots, X_n$ is

$$\mathsf{N}\left(\frac{n\sigma^2}{\tau^2 + n\sigma^2}\bar{X} + \frac{\tau^2}{\tau^2 + n\sigma^2}\mu,\ \frac{\sigma^2\tau^2}{\tau^2 + n\sigma^2}\right).$$

(Poisson, Gamma) pair Consider $X \sim \mathsf{Poi}(\lambda)$, where $\lambda \sim \mathsf{Gamma}(a, b)$ for $a, b > 0$. Recall that the gamma distribution has density

$$p(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}, \quad \text{where } \Gamma(a) = \int_0^\infty x^{a-1}e^{-x}\,dx.$$

Then we have

$$f(x) = \int_0^\infty f(x \mid \lambda)p(\lambda)\,d\lambda = \int_0^\infty \frac{\lambda^x e^{-\lambda}}{x!}\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}\,d\lambda = \frac{b^a\Gamma(a+x)}{x!(b+1)^{a+x}\Gamma(a)}.$$

Hence, the posterior has density

$$p(\lambda \mid x) = \frac{f(x \mid \lambda)}{f(x)}p(\lambda) = \frac{(b+1)^{a+x}}{\Gamma(a+x)}\lambda^{a+x-1}e^{-(b+1)\lambda},$$

which turns out to be the density of $\mathsf{Gamma}(a + x, b + 1)$.

In this case, the prior and the posterior are conjugate distributions (i.e., in the same family of distributions), and the prior is called a conjugate prior. Other (likelihood, conjugate prior) pairs include (Binomial, Beta), (Multinomial, Dirichlet), ...

For a fixed $k \geq 0$, consider the loss $L(\lambda, \hat{\lambda}) = (\lambda - \hat{\lambda})^2/\lambda^k$. Then we have

$$\begin{aligned}
&\mathbb{E}_\pi[(\lambda - \hat{\lambda})^2/\lambda^k \mid X] = \int_0^\infty \frac{(\lambda - \hat{\lambda})^2}{\lambda^k}p(\lambda \mid X)\,d\lambda \\
&= \frac{\Gamma(a + X - k)}{\Gamma(a + X)}(b+1)^{k-2}\Big\{(b+1)\hat{\lambda}\big[(b+1)\hat{\lambda} - 2(a + X - k)\big] + (a + X - k)(a + X - k + 1)\Big\},
\end{aligned}$$

which is minimized at $\hat{\lambda} = \hat{\lambda}(X) = \frac{a+X-k}{b+1}$. Therefore, this $\hat{\lambda}$ is the Bayes estimator.

2.1.3 Hierarchical Bayes

It is sometimes useful to consider a hierarchical framework of Bayesian estimation consisting of more than one "level" of prior. For example, let X be the random observation with density $f(x \mid \theta)$, where θ is the parameter with prior density $p(\theta \mid \gamma)$ further parametrized by a hyperparameter γ; moreover, suppose that γ is a random variable with density $\phi(\gamma)$. Examples:

- In the conjugate normal hierarchy, we have $X \sim \mathsf{N}(\theta, 1)$, $\theta \sim \mathsf{N}(0, \sigma^2)$, and $\frac{1}{\sigma^2} \sim \mathsf{Gamma}(a, b)$, where a, b are known.

- For exemplary purpose, consider the following mixture model:

  – $X \sim \sum_{i=1}^k w_i\,\mathsf{N}(\theta_i, 1)$, where $\theta_1, \ldots, \theta_k$ are fixed.

– $w = (w_1, \dots, w_k)$ follows the Dirichlet distribution with parameter $\alpha > 0$. The Dirichlet distribution is defined on the probability simplex $\Delta := \{v \in \mathbb{R}^k : \sum_{i=1}^k v_i = 1, v_i \geq 0\}$ and has density $\frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \prod_{i=1}^k w_i^{\alpha-1}$.
If $\alpha = 1$, the Dirichlet distribution becomes the uniform distribution on the simplex Δ. The smaller α, the "sparser" a corresponding Dirichlet random variable. As $\alpha \to \infty$, the Dirichlet distribution converges to the point mass at $(\frac{1}{k}, \dots, \frac{1}{k})$. As $\alpha \to 0$, the Dirichlet distribution converges to the discrete distribution $\frac{1}{k} \sum_{i=1}^k \delta_{e_i}$.
– α follows the exponential distribution with density $e^{-\alpha}$, for example.

Such a hierarchical model may be useful partly because the hyperparameter space is more manageable and allows tuning of the sparsity of the mixing weights.

2.1.4 Several perspectives of estimation

Unbiased versus Bayesian In the Gaussian example $X \sim \mathsf{N}(\theta, 1)$ where $\theta \sim \mathsf{N}(0, \sigma^2)$, the Bayes estimator $\hat{\theta}(X) = \frac{\sigma^2}{1+\sigma^2} X$ is biased. The next result shows an intrinsic contradiction between unbiased and Bayesian estimation.

Theorem 2.1. *Consider $X \sim \mathcal{P}_\theta$ where $\theta \sim \pi$. Under the squared loss $L(g, \hat{g}) = (g - \hat{g})^2$, no unbiased estimator $\hat{g}$ can be a Bayes estimator, unless the Bayes risk is zero.*

Proof. For the squared loss, the Bayes estimator is the posterior mean $\hat{g}(X) = \mathbb{E}[g(\theta) \mid X]$. If $\hat{g}(X)$ is unbiased, we have $\mathbb{E}[\hat{g}(X) \mid \theta] = g(\theta)$ for any θ. Then

- $\mathbb{E}[g(\theta)\hat{g}(X)] = \mathbb{E}[\mathbb{E}[g(\theta)\hat{g}(X) \mid \theta]] = \mathbb{E}[g(\theta)\, \mathbb{E}[\hat{g}(X) \mid \theta]] = \mathbb{E}[g(\theta)^2]$;
- $\mathbb{E}[g(\theta)\hat{g}(X)] = \mathbb{E}[\mathbb{E}[g(\theta)\hat{g}(X) \mid X]] = \mathbb{E}[\mathbb{E}[g(\theta) \mid X]\, \hat{g}(X)] = \mathbb{E}[\hat{g}(X)^2]$.

Therefore $\mathbb{E}[(g(\theta) - \hat{g}(X))^2] = 0$. □

Maximum likelihood estimation Consider the likelihood $\mathcal{L}(\theta \mid x) := f(x \mid \theta)$ and the log-likelihood $\log \mathcal{L}(\theta \mid x) := \log f(x \mid \theta)$. Given i.i.d. $X_1, \dots, X_n \sim \mathcal{P}_\theta$, the maximum likehood estimator (MLE) of $g(\theta)$ is defined to be $\hat{g} := g(\hat{\theta})$ where

$$\hat{\theta} := \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^n \log f(X_i \mid \theta).$$

Let us recall the unbiased estimation for Gaussian mean and variance. Given i.i.d. $X_1, \dots, X_n \sim \mathsf{N}(\theta, \sigma^2)$ where θ and σ are unknown, the empirical mean $\hat{\theta}$ is the MVUE of θ and $\frac{n}{n-1}\hat{\sigma}^2$ is the MVUE of σ^2, where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\theta})^2$. If we do not require the estimators to be unbiased, can they achieve better risks? How about the MLEs? We have

$$(\hat{\theta}, \hat{\sigma}) = \operatorname*{argmin}_{(\theta,\sigma)} \sum_{i=1}^n \frac{(X_i - \theta)^2}{2\sigma^2} + \frac{n}{2} \log(2\pi\sigma^2),$$

so the MLEs are $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\theta})^2$. In particular, $\hat{\sigma}^2$ is a biased estimator of σ^2. One can check

$$\mathbb{E}(c\hat{\sigma}^2 - \sigma^2)^2 = \sigma^4 \Big(\frac{n^2 - 1}{n^2} c^2 - 2\frac{n-1}{n} c + 1 \Big),$$

so we have

$$\mathbb{E}(\hat{\sigma}^2 - \sigma^2)^2 < \mathbb{E}(\frac{n}{n-1}\hat{\sigma}^2 - \sigma^2)^2.$$

In fact, $\frac{n}{n+1}\hat{\sigma}^2$ is even a better choice in terms of minimizing the risk.

2.2 Bayesian Cramér–Rao, a.k.a. van Trees inequality

We establish a Bayesian version of the Cramér–Rao bound, which is also known as the van Trees inequality. In fact, this bound holds for any estimator, rather than an unbiased estimator as in the case of the Cramér–Rao bound. For simplicity, let us focus on the univariate case. See [GL95] for a multivariate version of the theorem below. As before, we assume mild regularity conditions so that we can differentiate under the integral sign.

Theorem 2.2 (Bayesian Cramér–Rao, van Trees)**.** *Let π be a distribution on $\Theta := (a, b) \subset \mathbb{R}$ with density p such that $p(\theta) \to 0$ as $\theta \to a$ or b. Consider $X \sim \mathcal{P}_\theta$ where $\theta \sim \pi$. Suppose that $f(x \mid \theta)$ is bounded. Define*

$$I(\theta) = \mathbb{E}_\theta \Big(\frac{\partial}{\partial\theta} \log f(X \mid \theta)\Big)^2 \quad \text{and} \quad \mathcal{I}(\pi) = \mathbb{E}\Big(\frac{\partial}{\partial\theta} \log p(\theta)\Big)^2.$$

For any estimator $\hat{g}(X)$ of a differentiable estimand $g(\theta)$, we have

$$\mathbb{E}\Big(\hat{g}(X) - g(\theta)\Big)^2 \geq \frac{\big(\mathbb{E}[g'(\theta)]\big)^2}{\mathbb{E}[I(\theta)] + \mathcal{I}(\pi)}.$$

Proof. By the Cauchy-Schwarz inequality, we obtain

$$\mathbb{E}\left(\hat{g}(X) - g(\theta)\right)^2 \cdot \mathbb{E}\left(\frac{\partial}{\partial\theta} \log\Big(f(X \mid \theta)p(\theta)\Big)\right)^2 \geq \left(\mathbb{E}\left[\Big(\hat{g}(X) - g(\theta)\Big) \cdot \frac{\partial}{\partial\theta} \log\Big(f(X \mid \theta)p(\theta)\Big)\right]\right)^2.$$

Let us first compute the right-hand side. Since $p(\theta) \to 0$ as $\theta \to a$ or b, we have

$$\int \frac{\partial}{\partial\theta}\Big(f(x \mid \theta)p(\theta)\Big)\, d\theta = 0$$

and

$$\int g(\theta) \cdot \frac{\partial}{\partial\theta}\Big(f(x \mid \theta)p(\theta)\Big)\, d\theta = -\int g'(\theta) \cdot f(x \mid \theta)p(\theta)\, d\theta.$$

Consequently,

$$\begin{aligned}
&\mathbb{E}\left[\Big(\hat{g}(X) - g(\theta)\Big) \cdot \frac{\partial}{\partial\theta} \log\Big(f(X \mid \theta)p(\theta)\Big)\right] \\
&= \int\int \Big(\hat{g}(x) - g(\theta)\Big) \cdot \frac{\partial}{\partial\theta} \log\Big(f(x \mid \theta)p(\theta)\Big) \cdot f(x \mid \theta)p(\theta)\, d\theta\, d\mu(x) \\
&= \int\int \Big(\hat{g}(x) - g(\theta)\Big) \cdot \frac{\partial}{\partial\theta}\Big(f(x \mid \theta)p(\theta)\Big)\, d\theta\, d\mu(x) \\
&= \int\int g'(\theta) f(x \mid \theta)p(\theta)\, d\theta\, d\mu(x) = \mathbb{E}[g'(\theta)].
\end{aligned}$$

Moreover, for the left-hand side of the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\mathbb{E}\left(\frac{\partial}{\partial\theta}\log\Big(f(X\mid\theta)p(\theta)\Big)\right)^2 \\
&= \mathbb{E}\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right)^2 + \mathbb{E}\left(\frac{\partial}{\partial\theta}\log p(\theta)\right)^2 + 2\,\mathbb{E}\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\cdot\frac{\partial}{\partial\theta}\log p(\theta)\right) \\
&= \mathbb{E}[I(\theta)] + \mathcal{I}(\pi) + 0,
\end{aligned}$$

where we used $\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X\mid\theta)\mid\theta\right]=0$. Combining everything completes the proof. □

For example, for $X\sim\mathsf{N}(\theta,1)$ where $\theta\sim\mathsf{N}(0,\sigma^2)$, we have shown that the estimator $\hat\theta=\frac{\sigma^2}{1+\sigma^2}X$ of θ achieves the Bayes risk $\frac{\sigma^2}{1+\sigma^2}$ under the squared loss. On the other hand, we have

$$g'(\theta)=1,\quad I(\theta)=1,\quad \mathcal{I}(\pi)=1/\sigma^2,$$

so that

$$\frac{\big(\mathbb{E}[g'(\theta)]\big)^2}{\mathbb{E}[I(\theta)]+\mathcal{I}(\pi)}=\frac{1}{1+1/\sigma^2}=\frac{\sigma^2}{1+\sigma^2}.$$

Therefore, the Bayes risk achieved by the estimator $\hat\theta$ matches the Bayesian Cramér–Rao lower bound, so $\hat\theta$ is optimal in this sense.

2.3 Empirical Bayes and the James–Stein estimator

2.3.1 The empirical Bayes approach

Suppose that we observe i.i.d. $X_1,\dots,X_n\sim f(x\mid\theta)$ where $\theta\sim p(\theta\mid\gamma)$. If γ is known, we can estimate θ based on the data and the prior (for example, by the posterior mean). If γ is unknown, what we can do is to first estimate the hyperparameter γ from data to obtain an approximate prior, and then use it to estimate θ. Such an approach is called empirical Bayes. For example, we may do the following:

- Note that the marginal density of $X=(X_1,\dots,X_n)$ is

  $$f(x\mid\gamma)=\int\prod_{i=1}^n f(x_i\mid\theta)p(\theta\mid\gamma)\,d\theta.$$

  We can estimate γ by, for example, the MLE

  $$\hat\gamma(X):=\operatorname*{argmax}_{\tilde\gamma} f(X\mid\tilde\gamma).$$

- For a loss function $L(g,\hat g)$, we then consider the estimator $\hat g(X)$ that minimizes the empirical posterior loss

  $$\hat g(X):=\operatorname*{argmin}_{\tilde g}\int L\big(g(\theta),\tilde g(X)\big)p\big(\theta\mid X,\hat\gamma(X)\big)\,d\theta.$$

  Compare this to the posterior loss

  $$\mathbb{E}\left[L\big(g(\theta),\tilde g(X)\big)\mid X\right]=\int L\big(g(\theta),\tilde g(X)\big)p\big(\theta\mid X,\gamma\big)\,d\theta$$

  which we would minimize if γ were known.

In the original Bayesian framework, γ is assumed to be known, while in hierarchical Bayes, γ is assumed to follow a known distribution. Here in the empirical Bayes approach, γ is estimated from data. Moreover, we remark that $\hat{\gamma}$ and $\hat{g}$ can be replaced by other estimators.

2.3.2 James–Stein estimator and its variant

Consider $X \sim \mathsf{N}(\theta, I_n)$, where $\theta \sim \mathsf{N}(0, \sigma^2 I_n)$ and $n \geq 3$. We would like to estimate $\theta \in \mathbb{R}^n$ under the squared loss

$$L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2 = \sum_{i=1}^{n} (\theta_i - \hat{\theta}_i)^2.$$

Bayes estimator and its risk We start with the case where σ^2 is known. Similar to the univariate case, the Bayes estimator, denoted by $\hat{\theta}_{\mathsf{B}}(X)$, is again the posterior mean:

$$\hat{\theta}_{\mathsf{B}}(X) = \mathbb{E}[\theta \mid X, \sigma^2] = \int \theta \cdot p(\theta \mid X, \sigma^2)\, d\theta = \int \theta \cdot \frac{f(X \mid \theta, \sigma^2)\, p(\theta \mid \sigma^2)}{f(X \mid \sigma^2)}\, d\theta.$$

The density of X marginalized over θ is

$$\begin{aligned} f(X \mid \sigma^2) &= \int f(X \mid \theta, \sigma^2)\, p(\theta \mid \sigma^2)\, d\theta \\ &= \int \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\|X-\theta\|_2^2} \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\|\theta\|_2^2}\, d\theta \\ &= \frac{1}{(2\pi(1+\sigma^2))^{n/2}} e^{-\frac{1}{2(1+\sigma^2)}\|X\|_2^2}, \end{aligned} \tag{2.1}$$

which is simply the density of $\mathsf{N}(0, (1+\sigma^2) I_n)$. Similarly, we can compute

$$\int \theta \cdot f(X \mid \theta, \sigma^2)\, p(\theta \mid \sigma^2)\, d\theta = \frac{1}{(2\pi(1+\sigma^2))^{n/2}} e^{-\frac{1}{2(1+\sigma^2)}\|X\|_2^2} \cdot \frac{\sigma^2}{1+\sigma^2} X.$$

Hence the Bayes estimator is

$$\hat{\theta}_{\mathsf{B}}(X) = \frac{\sigma^2}{1+\sigma^2} X = \Big(1 - \frac{1}{1+\sigma^2}\Big) X, \tag{2.2}$$

with Bayes risk

$$R_\pi(\hat{\theta}_{\mathsf{B}}) = \iint \|\hat{\theta}_{\mathsf{B}}(x) - \theta\|_2^2 \cdot f(x \mid \theta, \sigma^2)\, p(\theta \mid \sigma^2)\, dx\, d\theta = \frac{n\sigma^2}{1+\sigma^2}.$$

Empirical Bayes Let us now consider the empirical Bayes approach to the case where σ^2 is unknown. Note that the marginal distribution given by (2.1) is $\mathsf{N}(0, (1+\sigma^2) I_n)$. To choose an estimator of σ^2, we require the associated estimator of $\frac{1}{1+\sigma^2}$ to be unbiased, in view of (2.2). A basic fact is that, if Y follows the chi-squared distribution with n degrees of freedom, then $\mathbb{E}[\frac{1}{Y}] = \frac{1}{n-2}$. Therefore, if we let $\hat{\tau}(X) := \frac{n-2}{\|X\|_2^2}$, then

$$\mathbb{E}\left[\hat{\tau}(X) \mid \sigma^2\right] = \frac{1}{1+\sigma^2}.$$

The associated empirical Bayes estimator is therefore

$$\hat{\theta}_{\mathsf{JS}}(X) = \Big(1 - \frac{n-2}{\|X\|_2^2}\Big)X,$$

which is called the James–Stein estimator.

Risk of the James–Stein estimator We now compute the risk of the James–Stein estimator

$$R(\theta, \hat{\theta}_{\mathsf{JS}}) = \mathbb{E}\left\|\Big(1 - \frac{n-2}{\|X\|_2^2}\Big)X - \theta\right\|_2^2 = \mathbb{E}\,\|X-\theta\|_2^2 + \mathbb{E}\,\frac{(n-2)^2}{\|X\|_2^2} - 2\,\mathbb{E}\left[\frac{n-2}{\|X\|_2^2}X^\top(X-\theta)\right], \quad (2.3)$$

where the expectation is with respect to $X \sim \mathsf{N}(\theta, I_n)$. Conditional on all X_j for $j \neq i$, Stein's lemma applied to X_i with $g(X_i) = \frac{n-2}{\|X\|_2^2}X_i$ (see Lemma 1.2 and (1.1)) yields that

$$\begin{aligned}\mathbb{E}_{X_i\sim\mathsf{N}(\theta_i,1)}\left[\frac{n-2}{\|X\|_2^2}X_i(X_i-\theta_i)\right] &= \mathbb{E}_{X_i\sim\mathsf{N}(\theta_i,1)}\left[\frac{\partial}{\partial X_i}\left(\frac{n-2}{\|X\|_2^2}X_i\right)\right] \\ &= \mathbb{E}_{X_i\sim\mathsf{N}(\theta_i,1)}\left[\frac{n-2}{\|X\|_2^2} - \frac{2(n-2)X_i^2}{\|X\|_2^4}\right].\end{aligned}$$

Summing the above equation over i and taking the expectation with respect to $X \sim \mathsf{N}(\theta, I_n)$, we obtain

$$\mathbb{E}\left[\frac{n-2}{\|X\|_2^2}X^\top(X-\theta)\right] = \mathbb{E}\,\frac{n(n-2)}{\|X\|_2^2} - \mathbb{E}\,\frac{2(n-2)}{\|X\|_2^2} = \mathbb{E}\,\frac{(n-2)^2}{\|X\|_2^2}.$$

Plugging this together with $\mathbb{E}\,\|X-\theta\|_2^2 = n$ into (2.3), we conclude that

$$R(\theta, \hat{\theta}_{\mathsf{JS}}) = n - \mathbb{E}\,\frac{(n-2)^2}{\|X\|_2^2}.$$

Furthermore, the Bayes risk of the James–Stein estimator is

$$\begin{aligned}R_\pi(\hat{\theta}_{\mathsf{JS}}) &= \mathbb{E}_{\theta\sim\mathsf{N}(0,\sigma^2 I_n)}\,R(\theta, \hat{\theta}_{\mathsf{JS}}) \\ &= \iint \Big(n - \frac{(n-2)^2}{\|x\|_2^2}\Big) f(x \mid \theta, \sigma^2)\, p(\theta \mid \sigma^2)\, dx\, d\theta \\ &= \int \Big(n - \frac{(n-2)^2}{\|x\|_2^2}\Big) f(x \mid \sigma^2)\, dx \\ &= n - \frac{n-2}{1+\sigma^2} = \frac{n\sigma^2}{1+\sigma^2} + \frac{2}{1+\sigma^2} = R(\hat{\theta}_{\mathsf{B}}) + \frac{2}{1+\sigma^2}.\end{aligned}$$

The relative increase of risk is

$$\frac{R(\hat{\theta}_{\mathsf{JS}}) - R(\hat{\theta}_{\mathsf{B}})}{R(\hat{\theta}_{\mathsf{B}})} = \frac{2}{n\sigma^2},$$

which is small if σ^2 is fixed and n is large.

Positive-part Stein estimator Instead, we can consider the maximum likelihood estimator (MLE) of $\frac{1}{1+\sigma^2}$ based on the marginal density (2.1). Namely, we solve

$$\max_{\tau} \left(\frac{\tau}{2\pi}\right)^{n/2} e^{-\frac{\tau}{2}\|X\|_2^2}$$

which yields $\tilde{\tau}(X) = \min\{\frac{n}{\|X\|_2^2}, 1\}$. Therefore, the associated empirical Bayes estimator is

$$\hat{\theta}_{\mathsf{PS}}(X) = \Big(1 - \min\Big\{\frac{n}{\|X\|_2^2}, 1\Big\}\Big)X = \Big(1 - \frac{n}{\|X\|_2^2}\Big)^+ X,$$

which is called a positive-part Stein estimator.

2.3.3 General results for exponential families

Theorem 2.3. *Consider the exponential family with density*

$$f(x \mid \eta) = \exp\big(\eta^\top T(x) - A(\eta)\big)h(x),$$

where $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^m$. Suppose that $\eta \sim p(\eta)$ for a prior density p. Let the marginal density of X be $f(x) = \int f(x \mid \eta)\, p(\eta)\, d\eta$. Define a matrix $D \in \mathbb{R}^{n\times m}$ by $D_{i,j} = \partial T_j / \partial x_i$. Then

$$\mathbb{E}[D\eta \mid x] = \nabla \log f(x) - \nabla \log h(x).$$

In particular, if $T(x) = x$, then we have $D = I$ and

$$\mathbb{E}[\eta \mid x] = \nabla \log f(x) - \nabla \log h(x).$$

Moreover, under the squared loss $L(\eta, \hat{\eta}) = \|\eta - \hat{\eta}\|_2^2$, the risk achieved by the Bayes estimator $\mathbb{E}[\eta \mid x]$ is

$$R(\eta, \mathbb{E}[\eta \mid X]) = R(\eta, -\nabla \log h(X)) + \sum_{i=1}^m \mathbb{E}\left[2\frac{\partial^2}{\partial X_i^2}\log f(X) + \Big(\frac{\partial}{\partial X_i}\log f(X)\Big)^2\right].$$

Proof. Lengthy but straightforward computation, which uses Stein's lemma. See [LC06], Chapter 4, Theorem 3.2, Corollary 3.3, and Theorem 3.5. □

Theorem 2.4. *Consider $X \sim \mathcal{P}_\eta$ from the exponential family with density*

$$f(x \mid \eta) = \exp\big(\eta^\top x - A(\eta)\big)h(x),$$

where $x, \eta \in \mathbb{R}^m$. Suppose that the prior is $p(\eta \mid \gamma)$. Let $\hat{\gamma}$ be the MLE of γ based on the marginal $f(x \mid \gamma) = \int f(x \mid \eta, \gamma)\, p(\eta \mid \gamma)\, d\eta$. Then the empirical Bayes estimator under the squared loss is

$$\mathbb{E}[\eta \mid X, \hat{\gamma}] = \nabla \log f(X \mid \hat{\gamma}(X)) - \nabla \log h(X).$$

Proof. See Chapter 4, Theorem 6.3 of [LC06]. □

2.4 Minimax estimation

2.4.1 Definitions and examples

Consider $X \sim \mathcal{P}_\theta$ where $\theta \in \Theta$. For an estimator $\tilde{g}$ of g, the maximum risk is $\sup_{\theta\in\Theta} R(g(\theta), \tilde{g}(X))$. The minimax risk is the minimum of the maximum risk:

$$R^* := \inf_{\tilde{g}} \sup_{\theta\in\Theta} R(g(\theta), \tilde{g}(X)).$$

An estimator $\hat{g}(X)$ of $g(\theta)$ is called a minimax estimator if it achieves the minimax risk:

$$\sup_{\theta\in\Theta} R(g(\theta), \hat{g}(X)) = R^*.$$

Consider the minimum Bayes risk

$$R^*_\pi := \inf_{\tilde{g}} R_\pi(\tilde{g}) = \inf_{\tilde{g}} \mathbb{E}_{\theta\sim\pi} R(g(\theta), \tilde{g}(X)).$$

Proposition 2.5. *We have*

$$R^* = \inf_{\tilde{g}} \sup_{\theta\in\Theta} R(g(\theta), \tilde{g}(X)) = \inf_{\tilde{g}} \sup_{\pi} R_\pi(\tilde{g}) \geq \sup_{\pi} \inf_{\tilde{g}} R_\pi(\tilde{g}) = \sup_{\pi} R^*_\pi.$$

As an example, consider i.i.d. $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$ where σ^2 is known. We aim to estimate μ under the squared loss $L(\mu, \hat{\mu}) = (\mu - \hat{\mu})^2$.

- For $\hat{\mu}(X) = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, we have $R(\mu, \hat{\mu}) = \mathbb{E}(\bar{X} - \mu)^2 = \frac{\sigma^2}{n}$. Hence $R^* \leq \frac{\sigma^2}{n}$.
- For $\mu \sim \mathsf{N}(0, \tau^2)$, the Bayes estimator, the posterior mean, is $\tilde{\mu}(X) = \frac{n\tau^2}{\sigma^2+n\tau^2}\bar{X}$ with the Bayes risk $R^*_\pi = R_\pi(\tilde{\mu}) = \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}$. Since $\tau^2 \geq 0$ is arbitrary, Proposition 2.5 implies that $R^* \geq \frac{\sigma^2}{n}$.

Therefore, $R^* = \frac{\sigma^2}{n}$, and $\hat{\mu}(X) = \bar{X}$ is a minimax estimator.

Lemma 2.6. *Consider parameter spaces $\Theta \subset \Theta'$. Let $\hat{g}(X)$ be a minimax estimator of $g(\theta)$ over Θ. If*

$$\sup_{\theta\in\Theta} R(g(\theta), \hat{g}(X)) = \sup_{\theta\in\Theta'} R(g(\theta), \hat{g}(X)),$$

then $\hat{g}$ is also minimax over Θ'.

Proof. If there exists another estimator $\tilde{g}$ with a smaller maximum risk over Θ' than $\hat{g}$, then the same is true over Θ, contradicting that $\hat{g}$ is minimax over Θ. □

Let us consider variants of the above Gaussian example:

- Consider i.i.d. $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$ where μ and σ^2 are unknown. We aim to estimate μ under the squared loss. Assume that $\sigma^2 \leq M$ where $M > 0$, for otherwise the minimax risk is infinite.

  The maximum risk of the estimator $\bar{X}$ is

  $$\sup_{(\mu,\sigma^2):\sigma^2\leq M} \mathbb{E}(\bar{X} - \mu)^2 = \frac{M}{n}.$$

  Recall that $\bar{X}$ is minimax over $\{(\mu, \sigma^2) : \sigma^2 = M\}$ with minimax risk M/n. Hence it is minimax over $\{(\mu, \sigma^2) : \sigma^2 \leq M\}$ in view of Lemma 2.6.

  Since $\bar{X}$ does not depend on M, we say that $\bar{X}$ adapts to M and call $\bar{X}$ an adaptive estimator.

- We now consider the loss $L(\mu, \hat{\mu}) = \frac{(\mu-\hat{\mu})^2}{\sigma^2}$, and impose no upper bound on $\sigma^2 > 0$. The estimator $\bar{X}$ has maximum risk

$$\sup_{(\mu,\sigma^2)} \frac{(\mu - \bar{X})^2}{\sigma^2} = \frac{1}{n}.$$

  Since $\bar{X}$ is minimax over $\{(\mu, \sigma^2) : \sigma^2 = 1\}$ with the same minimax risk $1/n$, it is minimax over $\{(\mu, \sigma^2) : \sigma^2 > 0\}$ as well.

2.4.2 Some theoretical results

Theorem 2.7. *Let $\hat{g}$ be a Bayes estimator for the prior π. If*

$$\sup_{\theta\in\Theta} R(g(\theta), \hat{g}(X)) = R_\pi(\hat{g}),$$

then the following holds:

1. *$\hat{g}$ is minimax with minimax risk $R^* = R^*_\pi = R_\pi(\hat{g})$;*

2. *If $\hat{g}$ is the unique Bayes estimator for π, then it is the unique minimax estimator;*

3. *$R^*_\pi \geq R^*_{\pi'}$ for any other prior distribution π' on Θ.*

In particular, if $\pi(\Omega) = 1$ where $\Omega := \{\theta \in \Theta : R(g(\theta), \hat{g}(X)) = \sup_{\theta'\in\Theta} R(g(\theta'), \hat{g}(X))\}$, then $\hat{g}$ is minimax.

Proof. For any estimator $\tilde{g}$, we have

$$\sup_\theta R(g, \tilde{g}) \geq R_\pi(\tilde{g}) \geq R_\pi(\hat{g}) = \sup_\theta R(g, \hat{g}),$$

so $\hat{g}$ is minimax.

Replacing the second $\geq$ with $>$ gives uniqueness.

For any distribution π' on Θ, let $\hat{g}'$ be the Bayes estimator for π'. Then

$$R^*_{\pi'} = R_{\pi'}(\hat{g}') \leq R_{\pi'}(\hat{g}) \leq \sup_\theta R(g, \hat{g}) = R_\pi(\hat{g}) = R^*_\pi.$$

In particular, if $\pi(\Omega) = 1$, then $R_\pi(\hat{g}) = \sup_\theta R(g, \hat{g})$. □

Theorem 2.8. *Let π_n be a sequence of prior distributions on Θ such that the following limit exists:*

$$R := \lim_{n\to\infty} R^*_{\pi_n}.$$

If $\hat{g}$ is an estimator for which

$$\sup_{\theta\in\Theta} R(g(\theta), \hat{g}(X)) = R,$$

then we have:

1. *$\hat{g}$ is minimax with minimax risk $R^* = R$;*

2. *$\lim_{n\to\infty} R^*_{\pi_n} \geq R^*_\pi$ for any prior distribution π on Θ.*

Proof. For any other estimator $\tilde{g}$, it holds

$$\sup_\theta R(g, \tilde{g}) \geq R_{\pi_n}(\tilde{g}) \geq R^*_{\pi_n}.$$

Taking the limit as $n \to \infty$, we have

$$\sup_\theta R(g, \tilde{g}) \geq R = \sup_{\theta \in \Theta} R(g, \hat{g}).$$

The second statement follows from $R^*_\pi \leq R^* = R$. □

2.4.3 Efron–Morris estimator

Consider i.i.d. $X_1, \ldots, X_n \sim \mathsf{N}(\mu, 1)$ where $\mu \sim \mathsf{N}(0, \tau^2)$. We aim to estimate μ under the squared loss. The Bayes estimator is $c\bar{X}$ with $c = \frac{n\tau^2}{1+n\tau^2}$. Its risk is

$$\mathbb{E}(c\bar{X} - \mu)^2 = c^2\,\mathbb{E}(\bar{X} - \mu)^2 + (1-c)^2\mu^2 = \frac{n^2\tau^4 + \mu^2}{(1+n\tau^2)^2},$$

and the maximum risk is infinite over $\mu \in \mathbb{R}$.

As a compromise, consider

$$\hat{\mu} := \begin{cases} \bar{X} + M & \text{if } \bar{X} < -\frac{M}{1-c} \\ c\bar{X} & \text{if } \bar{X} \in [-\frac{M}{1-c}, \frac{M}{1-c}] \\ \bar{X} - M & \text{if } \bar{X} > \frac{M}{1-c} \end{cases}$$

for $c \in (0, 1)$. Its risk is bounded.

2.5 Admissibility

An estimator $\hat{g}$ is inadmissible if there exists an estimator $\tilde{g}$ that dominates $\hat{g}$, i.e.,

- $R(g(\theta), \tilde{g}(X)) \leq R(g(\theta), \hat{g}(X))$ for all $\theta \in \Theta$, and
- $R(g(\theta), \tilde{g}(X)) < R(g(\theta), \hat{g}(X))$ for some $\theta \in \Theta$.

Otherwise, the estimator is called admissible.

2.5.1 Admissible estimators

Theorem 2.9. *Any unique Bayes estimator is admissible.*

Proof. If $\hat{g}$ is the unique Bayes estimator of g with respect to prior π and is dominated by $\tilde{g}$, then

$$\mathbb{E}_{\theta\sim\pi}\, R(\theta, \tilde{g}) \leq \mathbb{E}_{\theta\sim\pi}\, R(\theta, \hat{g}),$$

contradicting uniqueness. □

Theorem 2.10. *Any unique minimax estimator is admissible.*

Proof. If a minimax estimator is inadmissible, then another estimator dominates it and thus is also minimax, contradicting uniqueness. □

Theorem 2.11. *If an estimator has constant risk and is admissible, then it is minimax.*

Proof. If an estimator with constant risk is not minimax, then another estimator has smaller maximum risk and thus uniformly smaller risk. □

Theorem 2.12. *Suppose that $L(g, \cdot)$ is a strictly convex loss, and $\hat{g}(X)$ is an admissible estimator of $g(\theta)$. If $\tilde{g}(X)$ is another estimator of $g(\theta)$ such that $R(g, \hat{g}) = R(g, \tilde{g})$ at all θ, then $\hat{g} = \tilde{g}$ with probability 1.*

Proof. Define $\bar{g} = \frac{1}{2}(\hat{g} + \tilde{g})$. Then

$$L(g, \bar{g}) < \frac{1}{2}\big(L(g, \hat{g}) + L(g, \tilde{g})\big)$$

wherever $\hat{g} \neq \tilde{g}$. If this happens with nonzero probability, then

$$R(g, \bar{g}) < \frac{1}{2}\big(R(g, \hat{g}) + R(g, \tilde{g})\big) = R(g, \hat{g}),$$

contradicting the admissibility of $\hat{g}$. □

In Gaussian mean estimation, is the minimax estimator $\bar{X}$ admissible?

Proposition 2.13. *Given i.i.d. $X_1, \dots, X_n \sim \mathsf{N}(\mu, \sigma^2)$ where σ^2 is known, $\bar{X}$ is an admissible estimator and is the unique minimax estimator of μ under the squared loss.*

Proof. Consider any estimator $\hat{\mu}$ such that

$$R(\mu, \hat{\mu}) \leq \frac{\sigma^2}{n} = R(\mu, \bar{X}).$$

By the bias-variance decomposition, have

$$R(\mu, \hat{\mu}) = \mathsf{Var}_\mu(\hat{\mu}) + b(u)^2,$$

where $b(u) = \mathbb{E}_\mu[\hat{\mu}] - \mu$. Then the Cramér–Rao bound gives

$$R(\mu, \hat{\mu}) \geq \frac{(1 + b'(\mu))^2}{I(\mu)} + b(\mu)^2 = \frac{\sigma^2(1 + b'(\mu))^2}{n} + b(\mu)^2 \tag{2.4}$$

where $I(\mu) = -\mathbb{E}[\frac{\partial^2}{\partial \mu^2} \log f(X \mid \mu)] = n/\sigma^2$. Hence we obtain

$$\frac{(1 + b'(\mu))^2}{n} + \frac{b(\mu)^2}{\sigma^2} \leq \frac{1}{\sigma^2} R(\mu, \hat{\mu}) \leq \frac{1}{n}. \tag{2.5}$$

We claim that $\hat{\mu}$ is unbiased, i.e., $b(\mu) \equiv 0$:

1. The bias $b(\mu)$ is clearly bounded.

2. We have $(1 + b'(\mu))^2 = 1 + 2b'(\mu) + b(\mu)^2 \leq 1$, so $b'(\mu) \leq 0$, i.e., $b(\mu)$ is nonincreasing.

3. If $b'(\mu) < -\varepsilon$ for a fixed $\varepsilon > 0$ as $\mu \to \pm\infty$, then $b(\mu)$ cannot be bounded. Hence there is a sequence $\mu_i \to \pm\infty$ such that $b'(\mu_i) \to 0$.

4. By (2.5), $b(\mu_i) \to 0$ as $\mu_i \to \pm\infty$. Since $b(\mu)$ is nonincreasing, $b(\mu) \equiv 0$.

   Hence we also have $b'(\mu) \equiv 0$, and (2.4) implies that

$$R(\mu, \hat{\mu}) \geq \frac{\sigma^2}{n}.$$

We conclude that $R(\mu, \hat{\mu}) = R(\mu, \bar{X})$, and thus $\bar{X}$ is admissible. Moreover, $\bar{X}$ is the only minimax estimator by the above two theorems. □

2.5.2 Inadmissible estimators

The estimator $\bar{X}$ is no long admissible in the truncated case.

Proposition 2.14. *Consider $g(\theta)$ in a fixed interval $[a, b]$. Suppose the loss function $L(g, \hat{g})$ is zero at $\hat{g} = g$ and strictly increasing as $\hat{g}$ moves away from g. If $\hat{g}(X)$ takes values outside $[a, b]$ with positive probability, then $\hat{g}$ is inadmissible.*

Proof. Define an estimator $\tilde{g}$ by $\tilde{g} = \hat{g}$ if $\hat{g} \in [a, b]$, $\tilde{g} = a$ if $\hat{g} < a$, and $\tilde{g} = b$ if $\hat{g} > b$. Then $\tilde{g}$ dominates $\hat{g}$. □

Consider i.i.d. $X_1, \ldots, X_n \sim \mathsf{N}(\mu, 1)$ where $\mu \geq a$ for a fixed $a \in \mathbb{R}$. Then the above proposition implies that $\bar{X}$ is not admissible. However, $\bar{X}$ is still minimax. To see this, suppose that $\bar{X}$ is not minimax. Let $\hat{\mu}$ be an estimator such that

$$R(\mu, \hat{\mu}) \leq \frac{1}{n} - \varepsilon$$

for all $\mu \geq a$ and a fixed $\varepsilon > 0$. Hence the Cramér–Rao lower bound for biased estimators shows that

$$\frac{(1 + b'(\mu))^2}{n} + b(\mu)^2 \leq \frac{1}{n} - \varepsilon,$$

where $b(\mu) = \mathbb{E}[\hat{\mu}] - \mu$. Consequently, $b(\mu)$ is bounded, and $b'(\mu) \leq \sqrt{1 - \varepsilon n} - 1 \leq -\varepsilon n/2$ for all $\mu \geq a$, giving a contradiction.

If, in addition, $\mu \in [a, b]$, then $\bar{X}$ is neither admissible nor minimax. One can show that $\max\{a, \min\{\bar{X}, b\}\}$ has a uniformly smaller risk.

2.6 Shrinkage estimators and Stein's effect

2.6.1 Gaussian estimation

Proposition 2.15. *Consider $X \sim \mathsf{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d$ is to be estimated and $\Sigma \in \mathbb{R}^{d \times d}$ is known. Define a class of estimators*

$$\hat{\mu}(X) := \left(1 - \frac{h(\|X\|_2^2)}{\|X\|_2^2}\right) X,$$

where h is a real-valued function. If h is nondecreasing and $0 < h(\cdot) \leq 2\,\mathsf{tr}(\Sigma) - 4\,\lambda_{\mathsf{max}}(\Sigma)$, then under the loss $L(\mu, \hat{\mu}) = \|\mu - \hat{\mu}\|_2^2$, the estimator $\hat{\mu}$ dominates X. In particular, X is inadmissible and $\hat{\mu}$ is minimax.

Proof. For the last statement, it can be shown that X is minimax in a way similar to the one-dimensional case. Hence, $\hat{\mu}$ is also minimax. It remains to prove that $\hat{\mu}$ dominates X.

First, note that the risk of X is

$$\begin{aligned} R(\mu, X) &= \mathbb{E}\,\|\mu - X\|_2^2 = \mathbb{E}[(X-\mu)^\top (X-\mu)] = \mathbb{E}\left[\mathsf{tr}\left((X-\mu)^\top (X-\mu)\right)\right] \\ &= \mathbb{E}\left[\mathsf{tr}\left((X-\mu)(X-\mu)^\top\right)\right] = \mathsf{tr}\left(\mathbb{E}\left[(X-\mu)(X-\mu)^\top\right]\right) = \mathsf{tr}(\Sigma). \end{aligned}$$

The risk of $\hat{\mu}$ is

$$\begin{aligned} R(\mu, \hat{\mu}) = \mathbb{E}\,\|\mu - \hat{\mu}\|_2^2 &= \mathbb{E}[(\hat{\mu}-\mu)^\top (\hat{\mu}-\mu)] \\ &= \mathbb{E}[(X-\mu)^\top (X-\mu)] - 2\,\mathbb{E}\Big[\frac{h(\|X\|_2^2)}{\|X\|_2^2} X^\top (X-\mu)\Big] + \mathbb{E}\Big[\frac{h(\|X\|_2^2)^2}{\|X\|_2^2}\Big]. \end{aligned}$$

Let us write $Y = \Sigma^{-1/2} X \sim \mathsf{N}(\eta, I_d)$ where $\eta := \Sigma^{-1/2}\mu$. Then $\|X\|_2^2 = X^\top X = Y^\top \Sigma Y$, so

$$\mathbb{E}\Big[\frac{h(\|X\|_2^2)}{\|X\|_2^2} X^\top (X-\mu)\Big] = \mathbb{E}\Big[\frac{h(Y^\top \Sigma Y)}{Y^\top \Sigma Y} Y^\top \Sigma (Y-\eta)\Big] = \sum_{i=1}^d \mathbb{E}\Big[\frac{h(Y^\top \Sigma Y)}{Y^\top \Sigma Y} \sum_{j=1}^d Y_j \Sigma_{j,i} (Y_i - \eta_i)\Big].$$

Conditioning on $\{Y_j\}_{j \neq i}$ and applying Stein's lemma $\mathbb{E}[g(Y_i)(Y_i - \eta_i)] = \mathbb{E}[g'(Y_i)]$ for $Y_i \sim \mathsf{N}(\eta_i, 1)$, we obtain that each summand above is equal to

$$\begin{aligned} &\mathbb{E}\Big[\frac{\partial}{\partial Y_i}\Big(\frac{h(Y^\top \Sigma Y)}{Y^\top \Sigma Y} \sum_{j=1}^d Y_j \Sigma_{j,i}\Big)\Big] \\ &= \mathbb{E}\Big[\frac{h(Y^\top \Sigma Y)}{Y^\top \Sigma Y} \Sigma_{i,i} + \frac{2[h'(Y^\top \Sigma Y) Y^\top \Sigma Y - h(Y^\top \Sigma Y)] \sum_{k=1}^d \Sigma_{i,k} Y_k}{(Y^\top \Sigma Y)^2} \sum_{j=1}^d Y_j \Sigma_{j,i}\Big], \end{aligned}$$

where we used the fact that

$$\frac{\partial}{\partial Y_i}(Y^\top \Sigma Y) = \frac{\partial}{\partial Y_i}\Big(\sum_{j,k=1}^d Y_j \Sigma_{j,k} Y_k\Big) = 2\sum_{k=1}^d \Sigma_{i,k} Y_k.$$

Summing over i yields

$$\mathbb{E}\Big[\frac{h(\|X\|_2^2)}{\|X\|_2^2} X^\top (X-\mu)\Big] = \mathsf{tr}(\Sigma)\,\mathbb{E}\Big[\frac{h(\|X\|_2^2)}{\|X\|_2^2}\Big] + 2\,\mathbb{E}\Big[\frac{h'(\|X\|_2^2)\|X\|_2^2 - h(\|X\|_2^2)}{\|X\|_2^4} X^\top \Sigma X\Big].$$

Combining everything, we conclude that

$$R(\mu, \hat{\mu}) = \mathsf{tr}(\Sigma) + \mathbb{E}\Big[\frac{h(\|X\|_2^2)}{\|X\|_2^2}\Big(h(\|X\|_2^2) - 2\,\mathsf{tr}(\Sigma) + 4\frac{X^\top \Sigma X}{\|X\|_2^2}\Big)\Big] - 4\,\mathbb{E}\Big[\frac{h'(\|X\|_2^2)}{\|X\|_2^2} X^\top \Sigma X\Big],$$

which is smaller than $\mathsf{tr}(\Sigma)$ by the assumptions on h. □

Note that for $d \geq 3$ and $\Sigma = I_d$, there exists a function h satisfying the assumptions, making X inadmissible. This counterintuitive result is known as Stein's example or Stein's effect.

2.6.2 Poisson estimation

Lemma 2.16. *For a random variable X and functions f and g, suppose that $\mathbb{E}[f(X)]$, $\mathbb{E}[g(X)]$ and $\mathbb{E}[f(X)g(X)]$ all exist. If f and g are both nondecreasing, then*

$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)] \cdot \mathbb{E}[g(X)].$$

Moreover, if f and g are strictly increasing and X is not constant, then the above inequality is strict.

Proof. Let Y be an independent copy of X. Then

$$f(X)g(X) + f(Y)g(Y) - f(X)g(Y) - f(Y)g(X) = \big(f(X) - f(Y)\big)\big(g(X) - g(Y)\big) \geq 0.$$

Taking the expectation yields the desired inequality. □

Proposition 2.17. *Consider independent Poisson random variables $X_i \sim \mathsf{Poi}(\lambda_i)$ for $i \in [d]$ where $d \geq 2$. Let $\lambda = (\lambda_1, \ldots, \lambda_d) \in (0, \infty)^d$. Define the class of estimators*

$$\hat{\lambda}(X) := \left(1 - \frac{h(\sum_{i=1}^d X_i)}{\sum_{i=1}^d X_i + b}\right) X$$

where h is a real-valued nondecreasing function and $b > 0$. Consider the loss

$$L(\lambda, \hat{\lambda}) := \sum_{i=1}^d \frac{(\lambda_i - \hat{\lambda}_i)^2}{\lambda_i}.$$

If h is nondecreasing, $0 < h(\cdot) \leq 2(d-1)$, and $b \geq d-1$, then the estimator $\hat{\lambda}$ dominates X.

Proof. Let $Z = \sum_i X_i$. Then we have

$$\begin{aligned} R(\lambda, \hat{\lambda}) &= \mathbb{E}\Big[\sum_{i=1}^d \frac{1}{\lambda_i}\Big(X_i - \frac{h(Z)}{Z+b}X_i - \lambda_i\Big)^2\Big] \\ &= d - 2\,\mathbb{E}\Big[\frac{h(Z)}{Z+b}\sum_{i=1}^d \frac{1}{\lambda_i} X_i(X_i - \lambda_i)\Big] + \mathbb{E}\Big[\frac{h(Z)^2}{(Z+b)^2}\sum_{i=1}^d \frac{1}{\lambda_i} X_i^2\Big]. \end{aligned}$$

It is known that the conditional distribution $X_i \mid Z$ is multinomial with $\mathbb{E}[X_i \mid Z] = Z\lambda_i/\Lambda$ and $\mathsf{Var}(X_i \mid Z) = Z(\lambda_i/\Lambda)(1 - \lambda_i/\Lambda)$ where $\Lambda := \sum_{i=1}^d \lambda_i$. Therefore,

$$\begin{aligned} \mathbb{E}\Big[\sum_{i=1}^d \frac{1}{\lambda_i} X_i^2 \mid Z\Big] &= \sum_{i=1}^d Z\frac{1}{\Lambda}\Big(1 - \frac{\lambda_i}{\Lambda}\Big) + Z^2\frac{\lambda_i}{\Lambda^2} = \frac{Z}{\Lambda}(d - 1 + Z), \\ \mathbb{E}\Big[\sum_{i=1}^d \frac{1}{\lambda_i} X_i(X_i - \lambda_i) \mid Z\Big] &= \frac{Z}{\Lambda}(d - 1 + Z) - Z = \frac{Z}{\Lambda}(d - 1 + Z - \Lambda). \end{aligned}$$

We obtain

$$\begin{aligned} R(\lambda, \hat{\lambda}) &= d + \mathbb{E}\Big[\frac{h(Z)Z}{(Z+b)\Lambda}\Big(\frac{h(Z)}{Z+b}(d - 1 + Z) - 2(d-1) + 2(\Lambda - Z)\Big)\Big] \\ &\leq d + 2\,\mathbb{E}\Big[\frac{h(Z)Z}{(Z+b)\Lambda}(\Lambda - Z)\Big] \end{aligned}$$

by the assumptions $b \geq d-1$ and $h(\cdot) \leq 2(d-1)$. By Lemma 2.16,

$$\mathbb{E}\left[\frac{h(Z)Z}{(Z+b)\Lambda}(\Lambda - Z)\right] < \mathbb{E}\left[\frac{h(Z)Z}{(Z+b)\Lambda}\right] \cdot \mathbb{E}[\Lambda - Z] = 0,$$

which completes the proof. □

Chapter 3

Asymptotic estimation

In this chapter, we are interested in the scenario where the sample size n goes to infinity. Sections 3.1 to 3.5 establish the general theory. Topics of Sections 3.6 and 3.7 are not directly about asymptotic estimation but are related.

3.1 Convergence of random variables

3.1.1 Convergence in probability

Definition 3.1. *A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges to a random variable X in probability, denoted by $X_n \xrightarrow{p} X$, if for every $\varepsilon > 0$,*

$$\mathbb{P}\{|X_n - X| \geq \varepsilon\} \to 0 \quad \text{as } n \to \infty.$$

Definition 3.2. *Given i.i.d. observations $X_1, \ldots, X_n \sim \mathbb{P}_\theta$ where $\theta \in \Theta$, let us consider the estimator $\hat{g}_n = \hat{g}_n(X_1, \ldots, X_n)$ of $g(\theta)$. A sequence of estimators $\{\hat{g}_n\}_{n=1}^{\infty}$ is consistent if for every $\theta \in \Theta$, $\hat{g}_n \xrightarrow{p} g(\theta)$ with respect to $\mathbb{P}_\theta$.*

Theorem 3.3. *Let $\{\hat{g}_n\}_{n=1}^{\infty}$ be a sequence of estimators of $g(\theta) \in \mathbb{R}$, and consider the mean squared error (MSE) $\mathbb{E}(\hat{g}_n - g(\theta))^2$.*

- *If $\mathbb{E}(\hat{g}_n - g(\theta))^2 \to 0$ as $n \to \infty$ for all $\theta \in \Theta$, then $\hat{g}_n$ is consistent for estimating $g(\theta)$.*
- *As a result, if the bias and variance of $\hat{g}_n$ both converge to zero for all $\theta \in \Theta$, then $\hat{g}_n$ is a consistent estimator. In particular, any unbiased estimator with variance converging to zero is consistent.*

Proof. This is a result of Chebychev's inequality $\mathbb{P}_\theta\{|\hat{g}_n - g(\theta)| \geq \varepsilon\} \leq \frac{1}{\varepsilon^2}\,\mathbb{E}(\hat{g}_n - g(\theta))^2$. □

Some remarks about convergence in probability and consistency:

- Convergence in probability is preserved under sum, product, and continuous mapping.
- (Weak law of large numbers) Let $X_1, \ldots, X_n$ be i.i.d. with finite mean μ and variance σ^2. Since $\bar{X}$ has variance σ^2/n, it is consistent for estimating μ by the above theorem, i.e., $\bar{X} \xrightarrow{p} \mu$.

- In the above setting, consider the unbiased estimator of the variance,

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

  To see that this estimator is consistent, it suffices to note that

$$S_n^2 = \frac{n}{n-1}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\bar{X}-\mu)^2\Big),$$

  which converges to σ^2 in probability.

  Asymptotically, "optimal" estimators are typically not unique.

Theorem 3.4. *If $X_1, \ldots, X_n$ are i.i.d. with expectation μ, and g is a function that is continuous at μ, then $g(\bar{X}) \xrightarrow{p} g(\mu)$. In particular, if g is continuous, then the plug-in estimator $g(\bar{X})$ is consistent for estimating $g(\mu)$.*

Proof. Since $\bar{X} \xrightarrow{p} \mu$ as above, this follows from the continuous mapping theorem. □

- A sequence of estimators $\hat{g}_n$ of $g(\theta)$ is asymptotically unbiased (or unbiased in the limit) if $\mathbb{E}[\hat{g}_n] \to g(\theta)$.

- Instead of consistency, sometimes we are interested in finer results—rates of convergence. Namely, we may aim to establish

$$\mathbb{P}\{|\hat{g}_n - g(\theta)| \le r_n\} \ge 1 - \delta_n,$$

  where r_n and δ_n both go to zero as $n \to \infty$. This is clearly stronger than convergence in probability, and the quantity r_n is an upper bound on the rate of convergence.

3.1.2 Convergence in distribution

Definition 3.5. *Let $\{X_n\}_{i=1}^{\infty}$ be a sequence of random variables with CDFs $F_n(t) = \mathbb{P}\{X_n \le t\}$. Suppose that there exists a random variable X with CDF $F(t)$ such that $F_n(t) \to F(t)$ for all t at which F is continuous. Then we say that X_n converges to X in distribution or in law, denoted by $X_n \xrightarrow{d} X$, and that F_n converges to F weakly.*

- Convergence in distribution is preserved under continuous mapping, but not necessarily under sum or product.

- We have $X_n \xrightarrow{d} X$ if and only if $\mathbb{E}\, f(X_n) \to \mathbb{E}\, f(X)$ for every bounded continuous real-valued function f.

- (Central limit theorem) Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Then $\sqrt{n}(\bar{X}-\mu)/\sigma$ converges in distribution to the standard Gaussian $\mathsf{N}(0,1)$.

  Some properties about the two types of convergence:

- If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.

- If $X_n \xrightarrow{d} x$ for a constant x, then $X_n \xrightarrow{p} x$.
- If $X_n \xrightarrow{d} X$, $A_n \xrightarrow{d} a$ for a constant a, and $B_n \xrightarrow{d} b$ for a constant b, then $A_n + B_n X_n \xrightarrow{d} a + bX$.
- If $X_n \xrightarrow{d} X$, $y_n \to y$ where $\{y_n\}$ is a sequence of real numbers, and X has CDF $F(t)$ which is continuous at $t = y$, then we have $\mathbb{P}\{X_n \le y_n\} \to \mathbb{P}\{X \le y\} = F(y)$.

Theorem 3.6 (Delta method)**.** *Suppose that a real-valued function g on Θ has a nonzero derivative $g'(\theta)$ at θ. If $\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathsf{N}(0, \sigma^2)$, then*

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

If $g'(\theta) = 0$ and $g''(\theta) \neq 0$, then

$$n(g(T_n) - g(\theta)) \xrightarrow{d} \frac{1}{2}\sigma^2 g''(\theta)\chi_1^2,$$

where χ_1^2 is the chi-squared distribution with 1 degree of freedom.

Proof. Consider the first-order Taylor expansion of $g(T_n)$ around $g(\theta)$:

$$g(T_n) = g(\theta) + g'(\theta)(T_n - \theta) + R_n(T_n - \theta),$$

where $R_n \to 0$ as $T_n \to \theta$. The result then follows from the above properties of convergence.

If $g'(\theta) = 0$, consider the second-order Taylor expansion:

$$g(T_n) = g(\theta) + \frac{1}{2}g''(\theta)(T_n - \theta)^2 + \frac{1}{2}R_n(T_n - \theta)^2,$$

where $R_n \to 0$ as $T_n \to \theta$. Note that $n(T_n - \theta)^2 \xrightarrow{d} \sigma^2\chi_1^2$, so the same reasoning finishes the proof. □

Example: Bernoulli variance estimation Let $X_1, \ldots, X_n$ be i.i.d. $\mathsf{Ber}(p)$ random variables. Then the central limit theorem implies $\sqrt{n}(\bar{X} - p) \xrightarrow{d} \mathsf{N}(0, p(1-p))$.

Consider estimating $g(p) = p(1-p)$ by $\bar{X}(1-\bar{X})$. Since $g'(p) = 1 - 2p \neq 0$ for $p \neq 1/2$, it follows from the Delta method that

$$\sqrt{n}\big(\bar{X}(1-\bar{X}) - p(1-p)\big) \xrightarrow{d} \mathsf{N}\left(0, (1-2p)^2 p(1-p)\right).$$

For $p = 1/2$, we have $g'(1/2) = 0$ and $g''(1/2) = -2$. Hence the Delta method implies

$$n\big(\bar{X}(1-\bar{X}) - 1/4\big) \to -\frac{1}{4}\chi_1^2.$$

3.2 Asymptotic efficiency

Definition 3.7. *Consider i.i.d. $X_1, \dots, X_n \sim \mathcal{P}_\theta$ and an estimator $\hat{g}_n(X_1, \dots, X_n)$ of $g(\theta) \in \mathbb{R}$. We say that $\hat{g}_n$ is asymptotically normal if*

$$\sqrt{n}\big(\hat{g}_n - g(\theta)\big) \xrightarrow{d} \mathsf{N}\big(0, v(\theta)\big) \quad \text{for } v(\theta) > 0.$$

The quantity $v(\theta)$ is called the asymptotic variance of $\hat{g}_n$.

Definition 3.8. *A sequence of estimators $\{\hat{g}_n\}_{n=1}^{\infty}$ is called asymptotically efficient if*

$$\sqrt{n}\big(\hat{g}_n - g(\theta)\big) \xrightarrow{d} \mathsf{N}\Big(0, \frac{(g'(\theta))^2}{I(\theta)}\Big),$$

where $I(\theta)$ is the Fisher information each X_i contains about θ.

If $\hat{g}_n$ is unbiased, the Cramér–Rao bound says that

$$\mathsf{Var}_\theta(\hat{g}_n) \geq \frac{(g'(\theta))^2}{nI(\theta)}.$$

When do we have

$$v(\theta) \geq \frac{(g'(\theta))^2}{I(\theta)}?$$

A sufficient condition If $\hat{g}_n$ is unbiased and $\mathsf{Var}(\sqrt{n}\hat{g}_n) \to v(\theta)$, then

$$v(\theta) = \lim_{n\to\infty} \mathsf{Var}\left(\sqrt{n}\hat{g}_n\right) \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

Counterexample: superefficient estimator For i.i.d. $X_1, \dots, X_n \sim \mathsf{N}(\theta, 1)$ and $g(\theta) = \theta$, we have seen that $I(\theta) = 1$ (where $I(\theta)$ denotes the Fisher information of a single observation). Is it always true that $v(\theta) \geq 1$? No.

Consider the sequence of estimators

$$\hat{\theta}_n = \begin{cases} \bar{X} & \text{if } |\bar{X}| \geq 1/n^{1/4}, \\ a\bar{X} & \text{if } |\bar{X}| < 1/n^{1/4}, \end{cases}$$

where $a \in (0, 1)$. Therefore, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) = \begin{cases} \sqrt{n}(\bar{X} - \theta) & \text{if } |\sqrt{n}\bar{X}| \geq n^{1/4}, \\ a\sqrt{n}\bar{X} - \sqrt{n}\theta & \text{if } |\sqrt{n}\bar{X}| < n^{1/4}. \end{cases}$$

Let $Z_n = \sqrt{n}(\bar{X} - \theta) \sim \mathsf{N}(0, 1)$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta) = \begin{cases} Z_n & \text{if } |Z_n + \sqrt{n}\theta| \geq n^{1/4}, \\ a(Z_n + \sqrt{n}\theta) - \sqrt{n}\theta & \text{if } |Z_n + \sqrt{n}\theta| < n^{1/4}. \end{cases}$$

Consequently, if $\theta \neq 0$, then we have

$$\mathbb{P}\{\sqrt{n}(\hat{\theta}_n - \theta) \leq c\} - \mathbb{P}\{Z_n \leq c\} \to 0.$$

That is, $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathsf{N}(0, 1)$. On the other hand, if $\theta = 0$, then

$$\mathbb{P}\{\sqrt{n}(\hat{\theta}_n - \theta) \leq c\} - \mathbb{P}\{aZ_n \leq c\} \to 0.$$

That is, $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathsf{N}(0, a)$. However, $v(0) = a < 1$.

General result Under some reasonable general conditions, we have $v(\theta) \geq \frac{(g'(\theta))^2}{I(\theta)}$ except on a set of measure zero. See Chapter 6, Theorem 2.6 of [LC06].

3.3 Asymptotic properties of maximum likelihood estimation

Throughout the section, we consider i.i.d. observations $X_1, \ldots, X_n \sim \mathcal{P}_{\theta^*}$, where the distributions $\{\mathcal{P}_\theta\}_{\theta \in \Theta}$ are distinct and have common support. Suppose that θ^* is in the interior of Θ.

Recall that the likelihood (function) is $\mathcal{L}(\theta \mid x) = \prod_{i=1}^n f(x_i \mid \theta)$, and thus the log-likelihood is $\ell(\theta \mid x) = \log \mathcal{L}(\theta \mid x) = \sum_{i=1}^n \log f(x_i \mid \theta)$. Moreover, the MLE of θ is defined as $\hat{\theta} := \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(\theta \mid x)$. The MLE of $g(\theta)$ is defined to be $g(\hat{\theta})$. We call $\frac{\partial}{\partial \theta} \ell(\theta \mid x) = 0$ the likelihood equation, solving which yields the MLE (if it is unique).

3.3.1 Asymptotic consistency

Theorem 3.9. *We have that for any fixed $\theta \neq \theta^*$,*

$$\mathbb{P}_{\theta^*} \left\{ \mathcal{L}(\theta^* \mid X) > \mathcal{L}(\theta \mid X) \right\} \to 1 \quad \text{as } n \to \infty.$$

Proof. By the weak law of large numbers, we have

$$\frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i \mid \theta)}{f(X_i \mid \theta^*)} \xrightarrow{p} \mathbb{E}_{\theta^*} \Big[\log \frac{f(X \mid \theta)}{f(X \mid \theta^*)} \Big].$$

In addition, since $\log(\cdot)$ is strictly concave, Jensen's inequality implies

$$\mathbb{E}_{\theta^*} \Big[\log \frac{f(X \mid \theta)}{f(X \mid \theta^*)} \Big] < \log \mathbb{E}_{\theta^*} \Big[\frac{f(X \mid \theta)}{f(X \mid \theta^*)} \Big] = 0.$$

Therefore, it holds that

$$\lim_{n \to \infty} \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i \mid \theta)}{f(X_i \mid \theta^*)} < 0 \right\} \to 1,$$

which is equivalent to what we need to prove. □

Finite parameter space Let us consider a finite parameter space Θ. A sequence of estimators $\hat{\theta}_n$ is consistent if and only if

$$\mathbb{P}_{\theta *} \{ \hat{\theta}_n = \theta^* \} \to 1 \quad \text{for any } \theta^* \in \Theta.$$

We immediately obtain the following result.

Corollary 3.10. *If Θ is finite, then the MLE exists, is unique with probability going to 1, and is consistent.*

Real parameter space We consider an open set of parameters $\Theta \subset \mathbb{R}$ and use the shorthand $\ell(\theta \mid x) = \log \mathcal{L}(\theta \mid x)$.

Theorem 3.11. *Suppose that $f(x \mid \theta)$ is differentiable with respect to $\theta \in \Theta \subset \mathbb{R}$ for all x. Then with probability going to 1, the likelihood equation*

$$\ell'(\theta \mid X) = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(X_i \mid \theta)}{f(X_i \mid \theta)} = 0$$

has a root $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$, and that root satisfies $\hat{\theta}_n \xrightarrow{p} \theta^*$.

Proof. For $(\theta^* - \varepsilon, \theta^* + \varepsilon) \subset \Theta$, let

$$S_n := \big\{\ell(\theta^* \mid X) > \ell(\theta^* - \varepsilon \mid X) \text{ and } \ell(\theta^* \mid X) > \ell(\theta^* + \varepsilon \mid X)\big\}.$$

By Theorem 3.9, $\mathbb{P}_{\theta^*}\{S_n\} \to 1$. For any $x \in S_n$, there exists $\hat{\theta}_n \in (\theta^* - \varepsilon, \theta^* + \varepsilon)$ at which $\ell'(\hat{\theta}_n \mid x) = 0$. Hence we have

$$\mathbb{P}_{\theta^*}\{|\hat{\theta}_n - \theta^*| < \varepsilon\} \to 1.$$

Note that we can choose $\hat{\theta}_n$ to be the root closest to θ^* so that it does not depend on ε. □

However, Theorem 3.11 does not provide a practical way of choosing $\hat{\theta}_n$ in general, because θ^* is unknown and we cannot choose the root closest to θ^*.

Corollary 3.12. *If the likelihood equation has a unique root for all x and n, then $\hat{\theta}_n$ is a consistent estimator of θ. If, in addition, $\Theta = (a, b)$ where $-\infty \le a < b \le \infty$, then $\hat{\theta}_n$ is the MLE with probability going to 1.*

Proof. The first statement is immediate. To prove the second, suppose that $\hat{\theta}_n$ is not the MLE with probability bounded away from zero. No other interior point can be the MLE without satisfying the likelihood equation. On the other hand, if the likelihood converges to a supremum as $\theta \to a$ or b, this contradicts Theorem 3.9. □

3.3.2 Asymptotic efficiency

We in fact have asymptotic efficiency for the sequence $\hat{\theta}_n$ in Theorem 3.11.

Theorem 3.13. *Suppose that for all x and all $\theta \in (\theta^* - \varepsilon, \theta^* + \varepsilon)$,*

$$\Big|\frac{\partial^3}{\partial\theta^3} \log f(x \mid \theta)\Big| \le M(x) \quad \text{and} \quad \mathbb{E}_{\theta^*}[M(X)] < \infty.$$

Then the sequence $\hat{\theta}_n$ in Theorem 3.11 satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathsf{N}\Big(0, \frac{1}{I(\theta^*)}\Big),$$

i.e., it is asymptotically efficient.

Proof. For any fixed x, the Taylor expansion $\ell'(\hat{\theta}_n) = \ell'(\hat{\theta}_n \mid x)$ about θ^* gives

$$0 = \ell'(\hat{\theta}_n) = \ell'(\theta^*) + (\hat{\theta}_n - \theta^*)\ell''(\theta^*) + \frac{1}{2}(\hat{\theta}_n - \theta^*)^2\ell'''(\beta),$$

where β lies between θ^* and $\hat{\theta}_n$. Hence we have

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = \frac{\frac{1}{\sqrt{n}}\ell'(\theta^*)}{-\frac{1}{n}\ell''(\theta^*) - \frac{1}{2n}(\hat{\theta}_n - \theta^*)\ell'''(\beta)}.$$

Note that

$$\frac{1}{\sqrt{n}}\ell'(\theta^*) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{\partial}{\partial\theta}f(X_i \mid \theta^*)}{f(X_i \mid \theta^*)} \xrightarrow{d} \mathsf{N}\left(0, I(\theta^*)\right)$$

by the central limit theorem. Moreover,

$$\begin{aligned}-\frac{1}{n}\ell''(\theta^*) &= \frac{1}{n}\sum_{i=1}^{n}\frac{(\frac{\partial}{\partial\theta}f(X_i \mid \theta^*))^2 - f(X_i \mid \theta^*)\frac{\partial^2}{\partial\theta^2}f(X_i \mid \theta^*)}{f(X_i \mid \theta^*)^2} \\ &\xrightarrow{p} \mathbb{E}_{\theta^*}\frac{(\frac{\partial}{\partial\theta}f(X_i \mid \theta^*))^2}{f(X_i \mid \theta^*)^2} - \mathbb{E}_{\theta^*}\frac{\frac{\partial^2}{\partial\theta^2}f(X_i \mid \theta^*)}{f(X_i \mid \theta^*)} = I(\theta^*),\end{aligned}$$

by the law of large numbers. In addition, by the bound on the third derivative,

$$\left|\frac{1}{n}\ell'''(\beta)\right| \leq \frac{1}{n}\sum_{i=1}^{n}M(X_i) \xrightarrow{p} \mathbb{E}_{\theta^*}[M(X)].$$

Combining the above pieces completes the proof. □

Corollary 3.14. *If $\Theta = (a, b)$ and the likelihood equation has a unique root for all x and n, then the MLE is asymptotically efficient.*

Exponential family Consider the exponential family $f(x_i \mid \eta) = \exp\left(\eta T(x_i) - A(\eta)\right)$. The likelihood equation is

$$\frac{1}{n}\sum_{i=1}^{n}T(X_i) = A'(\eta) = \mathbb{E}_\eta[T(X_i)].$$

Moreover, $A''(\eta) = \mathsf{Var}_\eta\left(T(X_i)\right) = I(\eta) > 0$, so the RHS of the above equation is increasing in η. Hence the likelihood equation has exactly one solution $\hat{\eta}_n$, which satisfies

$$\sqrt{n}(\hat{\eta}_n - \eta) \xrightarrow{d} \mathsf{N}\left(0, 1/\mathsf{Var}(T)\right).$$

3.4 Examples of maximum likelihood estimation

In this section, we discuss some examples of maximum likelihood estimation, complementing the theory developed in Section 3.3.

3.4.1 Some examples

Let $X_1, \dots, X_n$ be i.i.d. observations from some parametric model. We consider finding the MLE by solving the log-likelihood equation. As we will see, this is not always possible.

Weibull distribution (MLE is the unique solution) The Weibull distribution on $(0, \infty)$ parametrized by $\lambda, k > 0$ is used in, for example, survival analysis. Its density is $f(x \mid \lambda, k) = \frac{k}{\lambda}\big(\frac{x}{\lambda}\big)^{k-1} e^{-(x/\lambda)^k}$ for $x > 0$. The log-likelihood is

$$\ell(\lambda, k \mid x) = \sum_{i=1}^{n} \bigg[\log \frac{k}{\lambda} + (k-1) \log \frac{x_i}{\lambda} - \Big(\frac{x_i}{\lambda}\Big)^k \bigg].$$

Therefore, the likelihood equation, or more precisely, the system of likelihood equations, is

$$\begin{aligned} \frac{\partial}{\partial \lambda} \ell(\lambda, k \mid x) &= \sum_{i=1}^{n} \Big(-\frac{k}{\lambda} + \frac{k x_i^k}{\lambda^{k+1}} \Big) = 0, \\ \frac{\partial}{\partial k} \ell(\lambda, k \mid x) &= \sum_{i=1}^{n} \Big(\frac{1}{k} + \log \frac{x_i}{\lambda} - \frac{x_i^k}{\lambda^k} \log \frac{x_i}{\lambda} \Big) = 0. \end{aligned}$$

We obtain that the MLE of λ is $\hat{\lambda}_n = \big(\sum_{i=1}^n x_i^k\big)^{1/k}$. To see the uniqueness of the MLE of k, note that with probability one, we have $x_i \neq \lambda$ for all $i \in [n]$. Then

- $\frac{\partial}{\partial k} \ell(\lambda, k \mid x) \to \infty$ as $k \to 0$,
- $\frac{\partial}{\partial k} \ell(\lambda, k \mid x) < 0$ as $k \to \infty$, and
- $\frac{\partial^2}{\partial^2 k} \ell(\lambda, k \mid x) = -\frac{1}{k^2} - \frac{x_i^k}{\lambda^k} \big(\log \frac{x_i}{\lambda} \big)^2 < 0$.

We see that the second likelihood equation also has a unique solution $\hat{k}_n$, which gives the MLE of k. The asymptotic efficiency of $\hat{\lambda}_n$ and $\hat{k}_n$ is guaranteed by the general theory. (However, we need a multivariate version of the theory proved above.)

Mixture of Gaussian distributions (MLE does not exist) Consider the mixture of two Gaussians $p\,\mathsf{N}(\mu, \sigma^2) + (1-p)\,\mathsf{N}(\eta, \tau^2)$ where $p \in (0, 1)$. The likelihood is

$$\mathcal{L}(\mu, \sigma^2, \eta, \tau^2 \mid x) = \prod_{i=1}^{n} \bigg[\frac{p}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x_i - \mu)^2}{2\sigma^2} \Big) + \frac{1-p}{\sqrt{2\pi}\tau} \exp\Big(-\frac{(x_i - \eta)^2}{2\tau^2} \Big) \bigg].$$

If we expand the product, one term will be

$$\frac{p(1-p)^{n-1}}{(2\pi)^{n/2} \sigma \tau^{n-1}} \exp\Big(-\frac{(x_1 - \mu)^2}{2\sigma^2} - \sum_{i=2}^{n} \frac{(x_i - \eta)^2}{2\tau^2} \Big).$$

When $\mu = x_1$ and $\sigma \to 0$, this term goes to infinity. Therefore, the likelihood is unbounded, and the MLE does not exist.

Uniform distribution (MLE is not asymptotically normal) Consider the uniform distributions $\mathsf{Unif}(0,\theta)$ parametrized by $\theta > 0$, which do not have a common support. The MLE of θ is $\hat{\theta}_n = X_{(n)}$ and the MVUE is $\tilde{\theta}_n = \frac{n+1}{n} X_{(n)}$. In addition, we have

$$n(\theta - \hat{\theta}_n) \xrightarrow{d} \mathsf{Exp}(0,\theta) \quad \text{and} \quad n(\theta - \tilde{\theta}_n) \xrightarrow{d} \mathsf{Exp}(-\theta,\theta),$$

where $\mathsf{Exp}(a,b)$ denotes the exponential distribution with density $\frac{1}{b} e^{-(x-a)/b} \mathbb{1}_{[a,\infty)}(x)$. In addition, one can show that

$$\mathbb{E}\left(n(\hat{\theta}_n - \theta)\right)^2 = 2\theta^2 \frac{n^2}{n^2+3n+2} \to 2\theta^2 \quad \text{and} \quad \mathbb{E}\left(n(\tilde{\theta}_n - \theta)\right)^2 = \theta^2 \frac{n^2}{n^2+2n} \to \theta^2.$$

3.4.2 Linear regression

We now move away from the i.i.d. case. Consider the linear regression model

$$Y = X\beta^* + \varepsilon,$$

where Y is the vector of observations in $\mathbb{R}^n$, X is the design matrix in $\mathbb{R}^{n \times d}$, β^* is the parameter vector in $\mathbb{R}^d$ to be estimated, and ε is the random vector of errors in $\mathbb{R}^n$. There are two types of assumptions on the design matrix X:

- Fixed design: X is a deterministic matrix.
- Random design: X is a random matrix. For example, the n rows of X are i.i.d. random vectors in $\mathbb{R}^d$.

In this section, we assume that $\varepsilon \sim \mathsf{N}(0, \sigma^2 I_n)$, so, equivalently, $Y \sim \mathsf{N}(X\beta^*, \sigma^2 I_n)$ with density

$$f(Y \mid \beta^*) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{\|Y - X\beta^*\|_2^2}{2\sigma^2}\Big)$$

in the case of a fixed design. For a random design, the above is still true conditional on X. The log-likelihood of β is therefore

$$\ell(\beta \mid Y) = -\frac{\|Y - X\beta\|_2^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2).$$

As a result, the maximum likelihood estimator $\hat{\beta}$ is the least squares estimator (LSE)

$$\hat{\beta} := \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \|Y - X\beta\|_2^2.$$

Low-dimensional case Let us first consider the case where $n \geq d$ and X is of rank d. The likelihood equation is $-2X^\top(Y - X\beta) = 0$, which we can solve to obtain the LSE

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Consider a random design where the n rows $\{X_i^\top\}_{i=1}^n$ of X are i.i.d. from some nice distribution with covariance $\Sigma = \mathbb{E}[X_i X_i^\top]$. Then the n entries of Y are i.i.d., and the asymptotic efficiency of $\hat{\beta}$ is guaranteed by the general multivariate theory. While it is not easy to verify this statement

directly, let us show that $\Delta_n := \sqrt{n}(\hat{\beta} - \beta^*)$ is Gaussian conditional on X and has the correct asymptotic covariance.

We fix n and condition on X in the sequel. First, the conditional Gaussianity of Δ_n is clearly true. Second, $\hat{\beta}$ is unbiased:

$$\mathbb{E}[\hat{\beta} \mid X] = (X^\top X)^{-1} X^\top X \beta^* = \beta^*,$$

so the Gaussian vector Δ_n has mean zero. Third, the covariance matrix of Δ_n can be computed:

$$\begin{aligned}
\mathbb{E}[\Delta_n \Delta_n^\top \mid X] &= n\, \mathbb{E}[(\hat{\beta} - \beta^*)(\hat{\beta} - \beta^*)^\top \mid X] \\
&= n\, \mathbb{E}[(X^\top X)^{-1} X^\top \varepsilon \varepsilon^\top X (X^\top X)^{-1} \mid X] \\
&= n (X^\top X)^{-1} X^\top (\sigma^2 I_n) X (X^\top X)^{-1} = \sigma^2 \Big(\frac{1}{n} X^\top X \Big)^{-1}.
\end{aligned}$$

To see that this gives the correct covariance asymptotically, note that $\frac{1}{n} X^\top X \xrightarrow{p} \Sigma$, so we need to check that $\sigma^2 \Sigma^{-1}$ is the inverse Fisher information matrix. Indeed, the Fisher information that each Y_i contains about β^* conditional on X_i is

$$I(\beta \mid X_i) = -\,\mathbb{E}[\nabla_\beta^2 \log f(Y_i \mid \beta^*, X_i) \mid X_i] = -\,\mathbb{E}\left[\nabla_\beta \Big(\frac{1}{\sigma^2} X_i (Y_i - X_i^\top \beta) \Big) \;\Big|\; X_i \right] = \frac{1}{\sigma^2} X_i X_i^\top.$$

Its expectation is $I(\beta) = \mathbb{E}[\frac{1}{\sigma^2} X_i X_i^\top] = \frac{1}{\sigma^2} \Sigma$, as expected.

High-dimensional case We now consider the case where the columns of X are not linearly independent, which is always true if $n < d$. Then the LSE $\hat{\beta}$ may not be unique. In this case, the problem with the formula $(X^\top X)^{-1} X^\top Y$ is that $X^\top X$ is not invertible. However, it suffices to replace $(X^\top X)^{-1}$ by the Moore–Penrose pseudoinverse $(X^\top X)^\dagger$, defined using the singular value decomposition (SVD). More precisely, define

$$\hat{\beta} = (X^\top X)^\dagger X^\top Y.$$

To see that $\hat{\beta}$ solves the likelihood equation $\nabla_\beta \|Y - X\beta\|_2^2 = -2(X^\top Y - X^\top X \beta) = 0$, it suffices to use basic properties of the pseudoinverse to check that

$$X^\top X \hat{\beta} = X^\top X (X^\top X)^\dagger X^\top Y = X^\top X X^\dagger Y = X^\top Y.$$

As $\hat{\beta}$ is not the unique solution of the likelihood equation, the general theory does not apply. In fact, since $n < d$, the number of dimensions has to grow as $n \to \infty$. Therefore, it is not even clear how we can talk about any asymptotic property.

3.5 Bernstein–von Mises theorem

While we have taken the frequentist point of view when discussing asymptotic estimation, it is also possible to analyze Bayesian procedures and talk about their asymptotics.

Gaussian mean estimation Recall the example of Bayesian Gaussian mean estimation: We observe i.i.d. $X_1, \dots, X_n \sim \mathsf{N}(\theta^*, 1)$ where $\theta^* \sim \mathsf{N}(0, \tau^2)$. Then the posterior mean is $\tilde{\theta}_n = \frac{n\tau^2}{1+n\tau^2}\bar{X}$. This is very close to the MLE $\hat{\theta}_n = \bar{X}$ when n is large, and $\tilde{\theta}_n \xrightarrow{p} \theta^*$ as $n \to \infty$. Moreover,

$$\sqrt{n}(\tilde{\theta}_n - \theta^*) = \sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathsf{N}(0, 1),$$

because $\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) = \sqrt{n}\frac{-1}{1+n\tau^2}\bar{X} \xrightarrow{p} 0$ and $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathsf{N}(0, 1)$ by the asymptotic efficiency of the MLE. Therefore, the Bayes estimator also enjoys the asymptotic efficiency just like the MLE.

We now introduce the Bernstein–von Mises theorem, which states that, as $n \to \infty$, the posterior distribution behaves like a Gaussian distribution centered at an efficient estimator (such as the MLE). The rigorous statement involves a set of regularity assumptions which we omit, and we only provide a very brief sketch of some ideas of the proof. See [vdV00] for the full statement and proof.

For two probability distributions $\mathcal{P}$ and $\mathcal{Q}$ with densities $f(x)$ and $g(x)$ respectively, define the total variation distance between them as

$$\mathsf{TV}(\mathcal{P}, \mathcal{Q}) := \frac{1}{2}\int \big|f(x) - g(x)\big|\, d\mu(x) = \frac{1}{2}\max_{|h(x)|\le 1}\int h(x)\big(f(x) - g(x)\big)\, d\mu(x).$$

Theorem 3.15 (Bernstein–von Mises; informal)**.** *Consider i.i.d. $X_1, \dots, X_n$ from a "nice" parametric model $\mathcal{P}_{\theta^*}$ where $\theta^* \in \Theta$. Let π be a prior on Θ with density $p(\theta) > 0$. Let π_n denote the posterior with density $p(\theta \mid X)$ where $X = (X_1, \dots, X_n)$. Moreover, let ϕ_n denote the distribution $\mathsf{N}\big(\hat{\theta}_n, \frac{1}{nI(\theta^*)}\big)$ where $\hat{\theta}_n$ is an asymptotically efficient estimator. Then we have that*

$$\mathsf{TV}(\pi_n, \phi_n) \xrightarrow{p} 0 \qquad \text{as } n \to \infty$$

with respect to the probability distribution $\mathcal{P}_{\theta^*}$.

Sketch of ideas. We are interested in the posterior around $\hat{\theta}_n$. Recall that the posterior is

$$p(\theta \mid x) = \frac{f(x \mid \theta)\cdot p(\theta)}{f(x)} = \frac{L(\theta \mid x)\cdot p(\theta)}{f(x)}$$

where L is the likelihood. Consider a change of variable $\phi = \sqrt{n}(\theta - \hat{\theta}_n)$, and let q denote the density of ϕ. Then our goal is to show that the posterior $q(\phi \mid x)$ is close to $\mathsf{N}\big(0, \frac{1}{I(\theta^*)}\big)$. Note that we have $\theta = \hat{\theta}_n + \phi/\sqrt{n}$ and

$$q(\phi \mid x) = \frac{1}{\sqrt{n}}p\Big(\hat{\theta}_n + \frac{\phi}{\sqrt{n}} \mid x\Big).$$

Therefore, combining the above two equations yields

$$\log q(\phi \mid x) = \ell\Big(\hat{\theta}_n + \frac{\phi}{\sqrt{n}}\Big) + \log p\Big(\hat{\theta}_n + \frac{\phi}{\sqrt{n}}\Big) + C(x),$$

where $\ell(\theta) = \log L(\theta \mid x)$ denotes the log-likelihood, and $C(x)$ is a quantity that only depends on x and whose particular value is not important. Taylor expansion yields

$$\log q(\phi \mid x) \approx \ell(\hat{\theta}_n) + \ell'(\hat{\theta}_n)\frac{\phi}{\sqrt{n}} + \frac{1}{2}\ell''(\hat{\theta}_n)\frac{\phi^2}{n} + \log p\Big(\hat{\theta}_n + \frac{\phi}{\sqrt{n}}\Big) + C(x).$$

Suppose that $\hat{\theta}_n$ is the MLE. Then, $\ell(\hat{\theta}_n)$ depends only on x, and $\ell'(\hat{\theta}_n) = 0$. Moreover, $\mathbb{E}[\ell''(\theta)] = -nI(\theta)$, so we have $\ell''(\hat{\theta}_n) \approx \ell''(\theta^*) \approx -nI(\theta^*)$. Finally, we approximate the term $\log p(\hat{\theta}_n + \frac{\phi}{\sqrt{n}})$ by $\log p(\theta^*)$. In summary,

$$\log q(\phi \mid x) \approx -\frac{1}{2} I(\theta^*)\phi^2 + C_2(x)$$

for a quantity $C_2(x)$ that only depends on x. Therefore, $q(\phi \mid x)$ is proportional to $\exp\left(\frac{\phi^2}{2/I(\theta^*)}\right)$. In other words, the posterior distribution of ϕ is approximately $\mathsf{N}\left(0, \frac{1}{I(\theta^*)}\right)$. □

3.6 Bootstrap methods

In this section, we introduce bootstrapping, a method based on sampling with replacement, and then study its asymptotic properties.

3.6.1 Jackknife estimator and bias reduction

Given i.i.d. observations $X_1, \ldots, X_n \sim \mathcal{P}_\theta$, let $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ be an estimator of θ with bias denoted by $b(\theta) = \mathbb{E}_\theta[\hat{\theta}_n] - \theta$. Can we estimate and reduce the bias?

Let $\hat{\theta}_{(-i)} = \hat{\theta}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ be the estimator of θ with the ith observation removed. The jackknife estimator of the bias $b(\theta)$ is defined to be

$$\hat{b}_n := (n-1)\Big(\frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(-i)} - \hat{\theta}_n\Big).$$

The jackknife bias-corrected estimator of θ is defined to be

$$\tilde{\theta}_n := \hat{\theta}_n - \hat{b}_n = n\hat{\theta}_n - \frac{n-1}{n}\sum_{i=1}^{n} \hat{\theta}_{(-i)}.$$

It is not hard to see that if

$$b(\theta) = \frac{A}{n} + \frac{B}{n^2} + O\Big(\frac{1}{n^3}\Big),$$

then the bias of the jackknife bias-corrected estimator is of a smaller order

$$\mathbb{E}[\tilde{\theta}_n] - \theta = O\Big(\frac{1}{n^2}\Big).$$

Bias reduction, however, does not necessarily yield a smaller risk.

3.6.2 Mean estimation and asymptotics

Let us use the simple problem of mean estimation to explain why bootstrapping is useful for studying properties of an estimator.

Fix i.i.d. observations $X_1, \ldots, X_n$ from a distribution with population mean $\theta \in \mathbb{R}$ and variance $\sigma^2 < \infty$. Suppose that we would like to study the sample mean

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Let $\mathcal{U}_n$ be the distribution of resampling from the observations uniformly, i.e., $\mathcal{U}_n = \mathsf{Unif}(\{X_i\}_{i=1}^n)$. Consider i.i.d. $Y_{n,i} \sim \mathcal{U}_n$ for $i = 1, \ldots, n$. We call

$$\bar{Y}_n := \frac{1}{n} \sum_{i=1}^{n} Y_{n,i}$$

the bootstrap sample mean. Conditional on $X_1, \ldots, X_n$, it is clear that each $Y_{n,i}$ has mean $\bar{X}_n$. Therefore, the bootstrap sample mean $\bar{Y}_n$ is an estimator of the sample mean $\bar{X}_n$. To understand the behavior of $\bar{X}_n$, the main idea of bootstrapping is that the distribution of $\bar{X}_n - \theta$ can be approximated by the distribution of $\bar{Y}_n - \bar{X}_n$. In the case of mean estimation, this claim is justified by the following theorem.

Theorem 3.16. *Let Φ denote the CDF of $\mathsf{N}(0,1)$. We have that, as $n \to \infty$,*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left\{ \sqrt{n} \left(\bar{Y}_n - \bar{X}_n \right) \le t \mid X_1, \ldots, X_n \right\} - \Phi(t/\sigma) \right| \longrightarrow 0$$

almost surely with respect to the randomness of $X_1, \ldots, X_n$.

The central limit theorem says that $\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{d} \mathsf{N}(0, \sigma^2)$, and the convergence is in fact uniform in the sense that

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left\{ \sqrt{n} \left(\bar{X}_n - \theta \right) \le t \right\} - \Phi(t/\sigma) \right| \longrightarrow 0$$

Comparing the above two displays, we see that the behavior of $\bar{X}_n - \theta$ is indeed similar to that of $\bar{Y}_n - \bar{X}_n$ when n is large.

In this setting, we understand the distribution of $\bar{X}_n - \theta$ very well in view of the central limit theorem, so there is no need to study $\bar{Y}_n - \bar{X}_n$. However, for more complicated models, if we have no idea of the behavior of

$$\hat{\theta}_n - \theta, \quad \text{where } \hat{\theta}_n := \hat{\theta}(X_1, \ldots, X_n),$$

bootstrapping becomes useful because we can generate bootstrap samples $\{Y_{n,1}, \ldots, Y_{n,n}\}$ and study the distribution of

$$\tilde{\theta}_n - \hat{\theta}_n, \quad \text{where } \tilde{\theta}_n := \hat{\theta}(Y_{n,1}, \ldots, Y_{n,n}).$$

There are typically two ways to generate the bootstrap samples $\{Y_{n,1}, \ldots, Y_{n,n}\}$:

- Nonparametric bootstrapping: We sample i.i.d. $Y_{n,1}, \ldots, Y_{n,n} \sim \mathcal{U}_n$ as above.
- Parametric bootstrapping: If we know that $X_1, \ldots, X_n$ are from a parametric model $\mathcal{P}_\theta$, instead of resampling $Y_{n,1}, \ldots, Y_{n,n}$ from $\{X_1, \ldots, X_n\}$, we can sample $Y_{n,1}, \ldots, Y_{n,n}$ from the model $\mathcal{P}_{\hat{\theta}_n}$, where $\hat{\theta}_n$ is the estimator in consideration. Then we can compute the bootstrap estimator $\tilde{\theta}_n$ in the same way.

For the proof of Theorem 3.16, we need the following lemma.

Lemma 3.17. *If X_n has CDF F_n and X has CDF F which is continuous, then*

$$X_n \xrightarrow{d} X \quad \Longrightarrow \quad \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \longrightarrow 0.$$

Proof. Since F is monotone and continuous, there exists x_i such that $F(x_i) = i/k$ for each $i = 1, \dots, k$ and $-\infty = x_0 < x_1 < \cdots < x_k = \infty$. For all $x \in [x_{i-1}, x_i]$, we have

$$F_n(x) - F(x) \le F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + 1/k,$$
$$F_n(x) - F(x) \ge F_n(x_{i-1}) - F(x_i) = F_n(x_{i-1}) - F(x_{i-1}) - 1/k.$$

Given any $\varepsilon > 0$, take k sufficiently large such that $1/k \le \varepsilon/2$. Then take n sufficiently large depending on ε and k such that

$$|F_n(x_i) - F(x_i)| \le \varepsilon/2 \quad \text{for all } i = 0, 1, \dots, k.$$

It follows that

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \le \varepsilon/2 + 1/k \le \varepsilon.$$

□

Proof sketch of Theorem 3.16. Let $X = (X_1, \dots, X_n)$. We have

$$\mathsf{Var}(Y_{n,i} \mid X) = \mathbb{E}[Y_{n,i}^2 \mid X] - (\mathbb{E}[Y_{n,i} \mid X])^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 \xrightarrow{p} \sigma^2$$

with respect to the randomness of X. By a version of the central limit theorem for the "triangular array" $\{Y_{n,i}\}$, we obtain

$$\sqrt{n}(\bar{Y}_n - \bar{X}_n) \xrightarrow{d} \mathsf{N}(0, \sigma^2).$$

Lemma 3.17 then implies that the desired result. □

3.7 Sampling methods

For a random variable $X \sim \mathcal{P}$ and a function g, it is a common task to compute the expectation $\mathbb{E}[g(X)]$. If this cannot be done analytically, then we can use numerical approximation. However, sometimes it is not even clear how we can sample $X \sim \mathcal{P}$ in practice.

3.7.1 Sampling with quantile function

Not every distribution is built in any program. Suppose that we would like to sample from a distribution $\mathcal{P}$. Let F be the CDF of $\mathcal{P}$. Then the quantile function of $\mathcal{P}$ is

$$Q(u) := \inf\{t : u \le F(t)\},$$

which is simply the inverse of F if F is invertible. If we know Q and can sample from $\mathsf{Unif}(0, 1)$, then we can sample from $\mathcal{P}$.

Proposition 3.18. *For any distribution $\mathcal{P}$ with quantile function Q, if $U \sim \mathsf{Unif}(0, 1)$, then $X = Q(U)$ has distribution $\mathcal{P}$.*

Proof. It suffices to note that

$$\mathbb{P}\{X \le t\} = \mathbb{P}\{Q(U) \le t\} = \mathbb{P}\{U \le F(t)\} = F(t).$$

□

For i.i.d. $X, X_1, \ldots, X_n \sim \mathcal{P}$, by the law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{p} \mathbb{E}[g(X)].$$

Therefore, we can approximate expectations via sampling.

3.7.2 Importance sampling

Sometimes the quantile function of a distribution $\mathcal{P}$ can be hard to compute, so we need another method to sample from $\mathcal{P}$. Let f be the density of $\mathcal{P}$. Let h be the density of a distribution $\mathcal{Q}$ whose support includes that of f. Suppose that we can sample from the distribution $\mathcal{Q}$. Let $X \sim \mathcal{P}$ and $Y \sim \mathcal{Q}$. We have

$$\mathbb{E}_{\mathcal{Q}}\left[\frac{g(Y)}{h(Y)}f(Y)\right] = \int \frac{g(y)}{h(y)}f(y)h(y)\,dx = \int g(x)f(x)\,dx = \mathbb{E}_{\mathcal{P}}[g(X)].$$

Therefore, to approximate $\mathbb{E}_{\mathcal{P}}[g(X)]$, we can sample i.i.d. $Y_1, \ldots, Y_n \sim \mathcal{Q}$ and use the fact

$$\frac{1}{n}\sum_{i=1}^{n} \frac{g(Y_i)}{h(Y_i)}f(Y_i) \xrightarrow{p} \mathbb{E}_{\mathcal{P}}[g(X)].$$

3.7.3 Metropolis–Hastings algorithm

We introduce a very simple example in the class of Metropolis–Hastings algorithms. Consider a setup similar to the previous subsection. Suppose that the densities f and h satisfy $f \le Mh$ for a constant $M > 0$. To sample $X \sim \mathcal{P}$, we can run:

1. Generate $Y \sim \mathcal{Q}$ and $U \sim \mathsf{Unif}(0,1)$.

2. If $U \le \frac{f(Y)}{Mh(Y)}$, take $X = Y$; otherwise, return to Step 1.

To see that $X \sim \mathcal{P}$, note that

$$\begin{aligned}\mathbb{P}\{X \le t\} = \mathbb{P}\left\{Y \le t \;\middle|\; U \le \frac{f(Y)}{Mh(Y)}\right\} &= \frac{\mathbb{P}\{Y \le t,\, U \le \frac{f(Y)}{Mh(Y)}\}}{\mathbb{P}\left\{U \le \frac{f(Y)}{Mh(Y)}\right\}} \\ &= \frac{\int_{-\infty}^{t} \mathbb{P}\left\{U \le \frac{f(y)}{Mh(y)}\right\} \cdot h(y)\,dy}{\int \mathbb{P}\left\{U \le \frac{f(y)}{Mh(y)}\right\} \cdot h(y)\,dy} = \frac{\frac{1}{M}\int_{-\infty}^{t} f(y)\,dy}{\frac{1}{M}\int f(y)\,dy} = F(t)\end{aligned}$$

which is the CDF of $\mathcal{P}$ at t.

3.7.4 Gibbs sampler

When dealing with a joint distribution, one may use the Gibbs sampler. Let us consider a simple example in the bivariate setting. Suppose that we would like to sample (X, Y) with density $f_{X,Y}$. Suppose that it is hard to sample from the joint distribution, but easy from the conditional distribution when one of the coordinate is fixed. Starting from a fixed value X_0, for $t \ge 1$, we do:

1. $Y_t \sim f_{Y|X}(\cdot \mid X = X_{t-1})$;

2. $X_t \sim f_{X|Y}(\cdot \mid Y = Y_t)$.

This algorithm generates a sequence $(X_t, Y_t)_{t=1}^n$ which is a Markov chain with invariant distribution $f_{X,Y}$. By the ergodic theorem, for a bivariate function g, we have

$$\frac{1}{n}\sum_{t=1}^{n} g(X_t, Y_t) \xrightarrow{p} \mathbb{E}[g(X, Y)].$$

This can be generalized to the multivariate case and is often useful in Bayesian estimation, for example, when computing posterior means.

Chapter 4

Finite-sample analysis

In this chapter, we focus on the minimax point of view and finite-sample analysis. We frequently prove results of the form

$$s_n \lesssim \inf_{\hat{g}_n} \sup_{\theta \in \Theta} R\big(g(\theta), \hat{g}_n\big) \lesssim r_n, \tag{4.1}$$

where "$\lesssim$" means "$\leq$" up to a constant factor (independent of n), and s_n and r_n are sequences that converge to zero as $n \to \infty$. The sequences s_n and r_n are referred to as rates of estimation or rates of convergence. Hopefully, we have $s_n = r_n$ in which case we obtain matching upper and lower bounds on the minimax risk.

A simple example we have seen is that, for $X_1, \ldots, X_n \sim \mathsf{N}(\mu, 1)$, we have

$$\inf_{\hat{\mu}_n} \sup_{\mu \in \mathbb{R}} \mathbb{E}(\mu - \hat{\mu})^2 = 1/n.$$

For more complex problems, we typically cannot obtain such a precise result.

Note that a result like (4.1) falls in the finite-sample category as it holds for any fixed n. Yet it has an asymptotic flavor because we are interested in large n and ignore constant factors. An upper bound like that in (4.1) strengthens asymptotic consistency since it gives an explicit rate of convergence over the entire parameter space.

4.1 Rates of estimation for linear regression

Recall the linear regression model

$$Y = X\beta^* + \varepsilon \tag{4.2}$$

with a fixed design matrix $X \in \mathbb{R}^{n \times d}$ and $\varepsilon \sim \mathsf{N}(0, \sigma^2 I_n)$. Let $\hat{\beta}$ be an estimator of β^*. In the sequel, we consider the following loss functions: (1) mean squared error $\frac{1}{d}\|\hat{\beta} - \beta^*\|_2^2$, and (2) mean squared prediction error $\frac{1}{n}\|X\hat{\beta} - X\beta^*\|_2^2$. There are usually two types of rates of estimation for problems studied here: (1) slow rate $1/\sqrt{n}$, and (2) fast rate $1/n$.

4.1.1 Fast rate for low-dimensional linear regression

Let us start with proving a fast rate of estimation in the low-dimensional case.

Theorem 4.1. *For the linear regression model* (4.2)*, the LSE* $\hat{\beta} = (X^\top X)^\dagger X^\top Y$ *satisfies*

$$\frac{1}{n}\,\mathbb{E}\,\|X\hat{\beta} - X\beta^*\|_2^2 = \sigma^2\frac{r}{n},$$

where r *is the rank of* X*. In addition, if* X *is of rank* d*, then*

$$\frac{1}{d}\,\mathbb{E}\,\|\hat{\beta} - \beta^*\|_2^2 = \frac{\sigma^2}{d}\sum_{i=1}^{d}\frac{1}{\lambda_i},$$

where $\lambda_1, \ldots, \lambda_d$ *are the eigenvalues of* $X^\top X$*.*

Proof. By the definition of $\hat{\beta}$ and basic properties of the pseudoinverse, we have

$$\begin{aligned} X\hat{\beta} - X\beta^* &= X(X^\top X)^\dagger X^\top Y - X\beta^* \\ &= X(X^\top X)^\dagger X^\top X\beta^* - X\beta^* + X(X^\top X)^\dagger X^\top\varepsilon \\ &= X(X^\top X)^\dagger X^\top\varepsilon. \end{aligned}$$

It follows that

$$\begin{aligned} \mathbb{E}\,\|X\hat{\beta} - X\beta^*\|_2^2 &= \mathbb{E}\,\|X(X^\top X)^\dagger X^\top\varepsilon\|_2^2 \\ &= \mathsf{tr}\,\mathbb{E}\left[X(X^\top X)^\dagger X^\top\varepsilon\varepsilon^\top X(X^\top X)^\dagger X^\top\right] \\ &= \sigma^2\,\mathsf{tr}\left(X(X^\top X)^\dagger X^\top\right) = \sigma^2 r, \end{aligned}$$

where the last equality can be seen from the SVD of X.

In addition, if X is of rank d, then similarly,

$$\hat{\beta} - \beta^* = (X^\top X)^{-1}X^\top X\beta^* - \beta^* + (X^\top X)^{-1}X^\top\varepsilon = (X^\top X)^{-1}X^\top\varepsilon.$$

As a result, we obtain

$$\mathbb{E}\,\|\hat{\beta} - \beta^*\|_2^2 = \sigma^2\,\mathsf{tr}\left((X^\top X)^{-1}\right),$$

from which the desired result follows. □

4.1.2 Maximal inequalities

As a preparation for the next subsection, we now introduce sub-Gaussian random variables and maximal inequalities. We say that a random variable X is sub-Gaussian with variance proxy σ^2 and write $X \sim \mathsf{subG}(\sigma^2)$ if the MGF of X satisfies

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

for any $\lambda > 0$. In particular, if $X \sim \mathsf{N}(\mu, \sigma^2)$, then $X \sim \mathsf{subG}(\sigma^2)$. Moreover, if $X \sim \mathsf{subG}(\sigma^2)$ and $a \in \mathbb{R}$, then $aX \sim \mathsf{subG}(a^2\sigma^2)$.

Proposition 4.2. *For (not necessarily independent) zero-mean random variables* $X_i \sim \mathsf{subG}(\sigma_i^2)$ *where* $i \in [n]$*, we have*

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \le \max_{i\in[n]}\sigma_i \cdot \sqrt{2\log n}.$$

Proof. Let $\sigma = \max_{i\in[n]} \sigma_i$. For any $\lambda > 0$, we have

$$\begin{aligned}\mathbb{E} \max_{i\in[n]} X_i &= \frac{1}{\lambda} \mathbb{E} \log e^{\lambda \max_{i\in[n]} X_i} \le \frac{1}{\lambda} \log \mathbb{E}\, e^{\lambda \max_{i\in[n]} X_i} \\ &\le \frac{1}{\lambda} \log \sum_{i\in[n]} \mathbb{E}\, e^{\lambda X_i} \le \frac{1}{\lambda} \log \sum_{i\in[n]} e^{\sigma_i^2 \lambda^2/2} \le \frac{\log n}{\lambda} + \frac{\lambda \sigma^2}{2}.\end{aligned}$$

Taking $\lambda = \sigma^{-1}\sqrt{2\log n}$ yields the result. □

We say that a random vector X in $\mathbb{R}^n$ is sub-Gaussian with variance proxy σ^2 and write $X \sim \mathsf{subG}_n(\sigma^2)$ if $v^\top X \sim \mathsf{subG}(\sigma^2)$ for any fixed unit vector $v \in \mathbb{R}^n$. In particular, if $X \sim \mathsf{N}(\mu, \sigma^2 I_n)$, then $X \sim \mathsf{subG}_n(\sigma^2)$.

Proposition 4.3. *Let K be a convex polytope in $\mathbb{R}^n$ with vertices $v_1, \dots, v_d$, and consider a zero-mean random vector $X \sim \mathsf{subG}_n(\sigma^2)$. Then we have*

$$\mathbb{E}\left[\max_{v\in K} X^\top v\right] \le \max_{i\in[d]} \|v_i\|_2 \cdot \sigma\sqrt{2\log d}.$$

Proof. Let us define

$$\Delta := \left\{\alpha \in \mathbb{R}^d : \alpha_i \ge 0, \sum_{i=1}^d \alpha_i = 1\right\}.$$

For any vector $v \in K$, we can write $v = \sum_{i=1}^d \alpha_i v_i$ where $\alpha \in \Delta$. Hence

$$\max_{v\in K} X^\top v = \max_{\alpha\in\Delta} \sum_{i=1}^d \alpha_i X^\top v_i = \max_{i\in[d]} X^\top v_i.$$

By the sub-Gaussianity of X, we have $X^\top v_i \sim \mathsf{subG}(\sigma^2\|v_i\|_2^2)$, so the result follows from Proposition 4.2. □

4.1.3 Slow rate for high-dimensional linear regression

Theorem 4.1 gives the fast rate d/n for low-dimensional linear regression when X is of rank-d. However, this rate becomes vacuous in the high-dimensional case since d/n is not shrinking to zero as n grows. For consistent estimation in high dimensions, we have to assume that β^* is of low complexity in a certain sense.

Let $K \subset \mathbb{R}^d$ be a closed set. If we know $\beta^* \in K$ a priori, we may consider the constrained LSE

$$\hat{\beta}_K := \operatorname*{argmin}_{\beta\in K} \|Y - X\beta\|_2^2. \tag{4.3}$$

A prominent example of K is the ℓ_1-ball of radius $\kappa > 0$, i.e., K is equal to

$$\mathcal{B}_1(\kappa) = \left\{\beta \in \mathbb{R}^d : \|\beta\|_1 = \sum_{i=1}^d |\beta_i| \le \kappa\right\}.$$

In this case, the optimization problem (4.3) is a computationally feasible convex program.

Theorem 4.4. *For the linear regression model* (4.2) *where* ε *is a zero-mean* $\mathsf{subG}_n(\sigma^2)$ *random vector, suppose that we have* $\beta^* \in \mathcal{B}_1(\kappa)$ *and* $\max_{j\in[d]} \|X_j\|_2 \le \sqrt{n}$ *where* X_j *denotes the* j*th column of* X*. Then the constrained LSE* $\hat{\beta}_{\mathcal{B}_1(\kappa)}$ *satisfies that*

$$\frac{1}{n}\,\mathbb{E}\,\|X\hat{\beta}_{\mathcal{B}_1(\kappa)} - X\beta^*\|_2^2 \lesssim \sigma\kappa\sqrt{\frac{\log d}{n}}.$$

Proof. For simplicity, we write $\hat{\beta} = \hat{\beta}_{\mathcal{B}_1(\kappa)}$. By the definition of $\hat{\beta}$, we have

$$\|Y - X\hat{\beta}\|_2^2 \le \|Y - X\beta^*\|_2^2 = \|\varepsilon\|_2^2.$$

Expanding the LHS gives

$$\|Y - X\hat{\beta}\|_2^2 = \|X\beta^* + \varepsilon - X\hat{\beta}\|_2^2 = \|X(\hat{\beta} - \beta^*)\|_2^2 - 2\varepsilon^\top X(\hat{\beta} - \beta^*) + \|\varepsilon\|_2^2.$$

As a result, we obtain

$$\|X(\hat{\beta} - \beta^*)\|_2^2 \le 2\varepsilon^\top X(\hat{\beta} - \beta^*) \le 4 \max_{\beta\in\mathcal{B}_1(\kappa)} \varepsilon^\top X\beta \le 4 \max_{v\in\mathcal{D}} \varepsilon^\top v,$$

where we used that $\|\hat{\beta} - \beta^*\|_1 \le 2\kappa$, and $\mathcal{D} := \{X\beta : \beta \in \mathcal{B}_1(\kappa)\} \subset \mathbb{R}^n$.

To bound this supremum, note that $\mathcal{D}$ is a polytope in $\mathbb{R}^n$ with at most $2d$ vertices that are in the set $\{\kappa X_1, -\kappa X_1, \ldots, \kappa X_d, -\kappa X_d\}$. In addition, each vertex has ℓ_2-norm bounded by $\kappa\|X_i\|_2 \le \kappa\sqrt{n}$ by assumption. As a result, we obtain

$$\mathbb{E}\,\|X(\hat{\beta} - \beta^*)\|_2^2 \le 4\,\mathbb{E} \max_{v\in\mathcal{D}} \varepsilon^\top v \le \kappa\sqrt{n} \cdot \sigma\sqrt{2\log 2d}$$

by Proposition 4.3. □

In fact, a rate of estimation like that in Theorem 4.1 also holds for $\hat{\beta}_{\mathcal{B}_1(\kappa)}$, so that we have

$$\frac{1}{n}\,\mathbb{E}\,\|X\hat{\beta}_{\mathcal{B}_1(\kappa)} - X\beta^*\|_2^2 \lesssim \min\left(\sigma^2\frac{d}{n}, \sigma\kappa\sqrt{\frac{\log d}{n}}\right).$$

Let us consider the case $\sigma = 1$ for simplicity. In the low-dimensional case $d \ll n$, we observe a fast rate d/n. In the high-dimensional case $d \gg n$, we obtain a slow yet consistent rate $\kappa\sqrt{\frac{\log d}{n}}$. This is usually called the elbow effect. In fact, one can still achieve a fast rate that scales as $1/n$ in the high-dimensional setting; we discuss a special case in the next section.

4.2 High-dimensional linear regression

4.2.1 Setup and estimators

As we have seen, when the dimension d of the problem exceeds the sample size n, we can still do consistent estimation if $\|\beta^*\|_1 \le \kappa$ for a small κ. Another popular assumption is that β^* is k-sparse, i.e., β^* is in

$$\mathcal{B}_0(k) = \left\{\beta \in \mathbb{R}^d : \|\beta\|_0 = \sum_{i=1}^d \mathbb{1}\{\beta_i \neq 0\} \le k\right\}.$$

There are many potential estimators of β^*:

- For $\|\beta^*\|_0 \le k$, the constrained LSE is

$$\hat{\beta}_{\mathcal{B}_0(k)} := \operatorname*{argmin}_{\|\beta\|_0 \le k} \|Y - X\beta\|_2^2.$$

  Unlike the constraint $\|\beta\|_1 \le k$ which is convex, here $\|\beta\|_0 \le k$ is a discrete constraint with more than $\binom{n}{k}$ possible choices of β. Hence this optimization problem is computationally infeasible in the worst case.

- If β^* is k-sparse and each entry of β^* is in $[-1, 1]$, then $\|\beta^*\|_1 \le k$. Hence we can still consider

$$\hat{\beta}_{\mathcal{B}_1(k)} := \operatorname*{argmin}_{\|\beta\|_1 \le k} \|Y - X\beta\|_2^2$$

  in this case, which is a computationally efficient and statistically consistent estimator.

- The ℓ_1-constrained LSE requires the knowledge of the (typically unknown) sparsity k. Instead, the most popularly used estimator is the LASSO estimator

$$\hat{\beta}^{\mathsf{LASSO}} := \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \|Y - X\beta\|_2^2 + 2\lambda\|\beta\|_1,$$

  where $\lambda = C\sigma\sqrt{n \log d}$ for a constant $C > 0$. This is an ℓ_1-penalized estimator, in comparison to the ℓ_1-constrained LSE above. They enjoy similar rates of estimation. Note that λ does not depend on k, so LASSO adapts to the sparsity of β^*.

- A more recently proposed estimator is the SLOPE estimator

$$\hat{\beta}^{\mathsf{SLOPE}} := \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \|Y - X\beta\|_2^2 + 2\tau\|\beta\|_*,$$

  where $\tau = C\sigma n$ for a constant $C > 0$ and the norm $\|\beta\|_*$ is defined as follows. Let $\lambda_i = \sqrt{\log(2d/i)}$ for $i \in [d]$. Let $\beta_{(1)}, \dots, \beta_{(d)}$ be the order statistics of $\beta_1, \dots, \beta_d$ such that

$$|\beta_{(1)}| \ge |\beta_{(2)}| \ge \cdots \ge |\beta_{(d)}|.$$

  Then we define

$$\|\beta\|_* := \sum_{i=1}^{d} \lambda_i |\beta_{(i)}|.$$

  The SLOPE estimator achieves slightly better rate of convergence than the LASSO estimator. Although the norm $\|\beta\|_*$ is more involved than $\|\beta\|_1$, the SLOPE estimator can be efficiently computed.

4.2.2 Fast rate for sparse linear regression

We now prove that the estimator $\hat{\beta}_{\mathcal{B}_0(k)}$ achieves a fast rate of estimation when the entries of β^* take discrete values. This setting is very special and the estimator is computationally infeasible. However, the simple proof provides some intuition for why we can obtain the fast rate of estimation in certain cases.

Theorem 4.5. *For the linear regression model* (4.2) *where* ε *is a zero-mean* $\mathsf{subG}_n(\sigma^2)$ *random vector, suppose that we have* $\beta^* \in \mathcal{B}_0(k)$ *and* $\beta_i^* \in \{-1, 0, 1\}$ *for each* $i \in [d]$*. Define an estimator*

$$\hat{\beta} := \operatorname*{argmin}_{\beta \in \mathcal{B}_0(k) \cap \{-1,0,1\}^d} \|X\beta - Y\|_2^2.$$

For any $\delta \in (0, 1)$*, it holds with probability at least* $1 - \delta$ *that*

$$\frac{1}{n}\|X\hat{\beta} - X\beta^*\|_2^2 \le 16\sigma^2 \frac{k}{n} \log \frac{5ed}{2k\delta} \lesssim \sigma^2 \frac{k}{n} \log \frac{d}{k\delta}.$$

Proof. Using the same argument as in the proof of Theorem 4.4, we obtain

$$\|X(\hat{\beta} - \beta^*)\|_2^2 \le 2\varepsilon^\top X(\hat{\beta} - \beta^*) = 2\|X(\hat{\beta} - \beta^*)\|_2 \cdot \varepsilon^\top \frac{X(\hat{\beta} - \beta^*)}{\|X(\hat{\beta} - \beta^*)\|_2}.$$

Define a set

$$\mathcal{D} := \left\{ v \in \mathbb{R}^d : v = \frac{Xu}{\|Xu\|_2} \text{ for } u \in \{-2, -1, 0, 1, 2\}^d,\ \|u\|_0 \le 2k \right\}.$$

Then we have

$$\|X(\hat{\beta} - \beta^*)\|_2 \le 2 \sup_{v \in \mathcal{D}} \varepsilon^\top v.$$

Since each $v \in \mathcal{D}$ is a unit vector, we have $\varepsilon^\top v \sim \mathsf{subG}(\sigma^2)$. As a result of a homework problem,

$$\mathbb{P}\{\varepsilon^\top v > t\} \le \exp\Big(-\frac{t^2}{2\sigma^2}\Big)$$

for any $t > 0$. Moreover, the set $\mathcal{D}$ has cardinality at most $\binom{d}{2k} 5^{2k} \le (\frac{ed}{2k})^{2k} 5^{2k}$. By a union bound,

$$\mathbb{P}\left\{ \sup_{v \in \mathcal{D}} \varepsilon^\top v > t \right\} \le \Big(\frac{ed}{2k}\Big)^{2k} 5^{2k} \cdot \exp\Big(-\frac{t^2}{2\sigma^2}\Big).$$

For $\delta \in (0, 1)$, we take $t = 2\sigma\sqrt{k \log \frac{5ed}{2k\delta}}$ and obtain

$$\mathbb{P}\left\{ \sup_{v \in \mathcal{D}} \varepsilon^\top v > 2\sigma\sqrt{k \log \frac{5ed}{2k\delta}} \right\} \le \delta.$$

This combined with $\|X(\hat{\beta} - \beta^*)\|_2 \le 2 \sup_{v \in \mathcal{D}} \varepsilon^\top v$ finishes the proof. □

4.2.3 Fast rate for LASSO

The computationally efficient LASSO estimator in fact achieves the fast rate of estimation for sparse linear regression under reasonable assumptions on the design matrix $X \in \mathbb{R}^{n \times d}$. We introduce two related assumptions:

- We say that the design matrix $X \in \mathbb{R}^{n\times d}$ is δ-incoherent if the matrix

$$\frac{1}{n}X^\top X - I_d$$

is entrywise bounded in absolute value by $\delta > 0$. Note that if the rows $x_i^\top$ of X are sampled i.i.d. from a distribution with mean zero and covariance I_d, then the above matrix is equal to

$$\frac{1}{n}\sum_{i=1}^n x_i x_i^\top - I_d$$

and converges to 0 by the law of large numbers. Therefore, incoherence is a reasonable assumption. Furthermore, it can be shown that as along as $n \gtrsim \frac{\log d}{\delta^2}$, we can sample a random matrix $X \in \mathbb{R}^{n\times d}$ that is δ-incoherent with probability 0.99.

- For any $\beta \in \mathbb{R}^d$ and $S \subset [d]$, let β_S denote the vector in $\mathbb{R}^d$ with $(\beta_S)_i = \beta_i$ for $i \in S$ and $(\beta_S)_i = 0$ for $i \in S^c := [d] \setminus S$. Define a cone of vectors

$$\mathcal{C}_S := \big\{\beta \in \mathbb{R}^d : \|\beta_{S^c}\|_1 \le 3\|\beta_S\|_1\big\}.$$

If $|S| \le k$, then the cone $\mathcal{C}_S$ contains approximately k-sparse vectors β with support approximately contained in S. We say that X satisfies the restricted eigenvalue (RE) condition for k-sparse vectors if

$$\inf_{|S|\le k}\inf_{\beta\in\mathcal{C}_S}\frac{\|X\beta\|_2^2}{n\|\beta\|_2^2} \ge \frac{1}{2}. \tag{4.4}$$

Note that if the infimum is taken over all $\beta \in \mathbb{R}^d$, then the condition is saying that the smallest eigenvalue of $\frac{1}{n}X^\top X$ is lower bounded by $1/2$. This is why the above condition is referred to as the RE condition. Furthermore, it can be shown that as soon as $n \gtrsim k\log d$, we can sample a random matrix $X \in \mathbb{R}^{n\times d}$ that satisfies (4.4) with probability 0.99.

The following result shows that incoherence is stronger than the RE condition when the parameters are appropriately chosen.

Proposition 4.6. *Consider a subset $S \subset [d]$ with $|S| \le k$ and a $\frac{1}{32k}$-incoherent matrix $X \in \mathbb{R}^{n\times d}$. If β satisfies the cone condition $\|\beta_{S^c}\|_1 \le 3\|\beta_S\|_1$, then it holds*

$$\|\beta\|_2^2 \le \frac{2}{n}\|X\beta\|_2^2.$$

As a result, every $\frac{1}{32k}$-incoherent matrix $X \in \mathbb{R}^{n\times d}$ satisfies (4.4).

Proof. We have

$$\|X\beta\|_2^2 = \|X\beta_S + X\beta_{S^c}\|_2^2 = \|X\beta_S\|_2^2 + \|X\beta_{S^c}\|_2^2 + 2\beta_S^\top X^\top X\beta_{S^c}.$$

The three terms can be controlled as follows:

- $\frac{1}{n}\|X\beta_S\|_2^2 = \|\beta_S\|_2^2 + \beta_S^\top\big(\frac{1}{n}X^\top X - I_d\big)\beta_S \ge \|\beta_S\|_2^2 - \frac{1}{32k}\|\beta_S\|_1^2$;
- $\frac{1}{n}\|X\beta_{S^c}\|_2^2 = \|\beta_{S^c}\|_2^2 + \beta_{S^c}^\top\big(\frac{1}{n}X^\top X - I_d\big)\beta_{S^c} \ge \|\beta_{S^c}\|_2^2 - \frac{1}{32k}\|\beta_{S^c}\|_1^2 \ge \|\beta_{S^c}\|_2^2 - \frac{9}{32k}\|\beta_S\|_1^2$;

- $\frac{1}{n}\beta_S^\top X^\top X \beta_{S^c} = \beta_S^\top \beta_{S^c} + \beta_S^\top \big(\frac{1}{n}X^\top X - I_d\big)\beta_{S^c} \geq -\frac{1}{32k}\|\beta_S\|_1\|\beta_{S^c}\|_1 \geq -\frac{3}{32k}\|\beta_S\|_1^2$.

Combining the three terms yields

$$\frac{1}{n}\|X\beta\|_2^2 \geq \|\beta\|_2^2 - \frac{1}{2k}\|\beta_S\|_1^2 \geq \|\beta\|_2^2 - \frac{|S|}{2k}\|\beta_S\|_2^2 \geq \frac{1}{2}\|\beta\|_2^2,$$

where we used the Cauchy–Schwarz inequality $\|\beta_S\|_1^2 \leq |S| \cdot \|\beta_S\|_2^2$. □

Lemma 4.7. *For $X \in \mathbb{R}^{n\times d}$, suppose that $\max_{j\in[d]} \|X_j\|_2^2 \leq 2n$ where X_j denotes the jth column of X. Let $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Then we have $\|X^\top\varepsilon\|_\infty \leq 2\sigma\sqrt{n\log(2d/\delta)}$ with probability at least $1-\delta$.*

Proof. We have $X_j^\top\varepsilon \sim \mathsf{subG}(2n\sigma^2)$ and thus $\mathbb{P}\{|X_j^\top\varepsilon| > t\} \leq \exp(\frac{-t^2}{4n\sigma^2})$ by a homework problem. Therefore,

$$\mathbb{P}\{\|X^\top\varepsilon\|_\infty > t\} \leq \mathbb{P}\Big\{\max_{j\in[d]}|X_j^\top\varepsilon| > t\Big\} \leq 2d\exp\Big(\frac{-t^2}{4n\sigma^2}\Big).$$

Choosing $t = 2\sigma\sqrt{n\log(2d/\delta)}$ completes the proof. □

Theorem 4.8. *Consider the linear regression model $Y = X\beta^* + \varepsilon$ where $\|\beta^*\|_0 \leq k$ and $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Suppose that X satisfies the RE condition (4.4) and that $\max_{j\in[d]} \|X_j\|_2^2 \leq 2n$ (both of which hold if X is $\frac{1}{32k}$-incoherent). Define the LASSO estimator*

$$\hat\beta := \operatorname*{argmin}_{\beta\in\mathbb{R}^d} \|Y - X\beta\|_2^2 + 2\lambda\|\beta\|_1, \qquad \text{where } \lambda := 8\sigma\sqrt{n\log(2d/\delta)}.$$

For any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that

$$\|\hat\beta - \beta^*\|_2^2 \leq \frac{2}{n}\|X\hat\beta - X\beta^*\|_2^2 \lesssim \sigma^2\frac{k}{n}\log\frac{d}{\delta}.$$

Proof. By the definition of the LASSO estimator, we have

$$\|Y - X\hat\beta\|_2^2 + 2\lambda\|\hat\beta\|_1 \leq \|Y - X\beta^*\|_2^2 + 2\lambda\|\beta^*\|_1.$$

Expanding $\|Y - X\hat\beta\|_2^2 = \|\varepsilon + X\beta^* - X\hat\beta\|_2^2$ and adding $\lambda\|\hat\beta - \beta^*\|_1$ on both sides, we obtain

$$\|X\beta^* - X\hat\beta\|_2^2 + \lambda\|\hat\beta - \beta^*\|_1 \leq 2\varepsilon^\top X(\hat\beta - \beta^*) + \lambda\|\hat\beta - \beta^*\|_1 + 2\lambda\|\beta^*\|_1 - 2\lambda\|\hat\beta\|_1.$$

By Hölder's inequality and the above lemma, it holds with probability at least $1-\delta$ that

$$\varepsilon^\top X(\hat\beta - \beta^*) \leq \|X^\top\varepsilon\|_\infty\|\hat\beta - \beta^*\|_1 \leq 4\sigma\sqrt{n\log(2d/\delta)}\,\|\hat\beta - \beta^*\|_1 = \frac{\lambda}{2}\,\|\hat\beta - \beta^*\|_1.$$

As a result, with S chosen to be the support of β^*, we get

$$\begin{aligned}\|X\beta^* - X\hat\beta\|_2^2 + \lambda\|\hat\beta - \beta^*\|_1 &\leq 2\lambda\|\hat\beta - \beta^*\|_1 + 2\lambda\|\beta^*\|_1 - 2\lambda\|\hat\beta\|_1 \\ &= 2\lambda\|\hat\beta_S - \beta_S^*\|_1 + 2\lambda\|\beta_S^*\|_1 - 2\lambda\|\hat\beta_S\|_1 \leq 4\lambda\|\hat\beta_S - \beta_S^*\|_1. \qquad (4.5)\end{aligned}$$

This in particular implies that

$$\|\hat\beta_{S^c} - \beta_{S^c}^*\|_1 \leq 3\|\hat\beta_S - \beta_S^*\|_1,$$

that is, $\hat{\beta} - \beta^*$ satisfies the cone condition. Then it follows from the Cauchy–Schwarz inequality and (4.4) that

$$\|\hat{\beta}_S - \beta^*_S\|_1 \le \sqrt{k}\,\|\hat{\beta}_S - \beta^*_S\|_2 \le \sqrt{k}\,\|\hat{\beta} - \beta^*\|_2 \le \sqrt{\frac{2k}{n}}\,\|X\hat{\beta} - X\beta^*\|_2.$$

Plugging this bound back into (4.5), we obtain

$$\|X\beta^* - X\hat{\beta}\|_2^2 \le 4\lambda\sqrt{\frac{2k}{n}}\,\|X\hat{\beta} - X\beta^*\|_2 \qquad \Longrightarrow \qquad \|X\beta^* - X\hat{\beta}\|_2^2 \le 32\lambda^2\frac{k}{n}.$$

This completes the proof in view of the definition of λ and that $\|\hat{\beta} - \beta^*\|_2^2 \le \frac{2}{n}\|X\hat{\beta} - X\beta^*\|_2^2$. □

4.3 Generalized linear regression

4.3.1 Setup and models

A generalized linear model can be defined using an exponential family. Fix a parameter vector $\beta^* \in \mathbb{R}^d$. For each $i = 1, \dots, n$, suppose that we observe a design point $x_i \in \mathbb{R}^d$ and a random outcome $Y_i \sim \mathcal{P}_{x_i^\top \beta^*}$ with density of the form

$$f(y_i \mid x_i^\top \beta^*) = \exp\Big(\frac{y_i \cdot x_i^\top \beta^* - A(x_i^\top \beta^*)}{\sigma^2}\Big) \cdot h(y_i, \sigma)$$

for functions $A(\cdot)$, $h(\cdot)$, and noise parameter $\sigma > 0$. Suppose the observations are independent so that the log-likelihood is

$$\ell(\beta \mid Y) = \sum_{i=1}^n \Big(\frac{Y_i \cdot x_i^\top \beta - A(x_i^\top \beta)}{\sigma^2} + \log h(Y_i, \sigma)\Big).$$

Usually the function A is convex, so we can efficiently solve for the MLE.

Let us see two examples of the above general model:

Gaussian linear regression In linear regression (4.2) with Gaussian noise $\varepsilon \sim \mathsf{N}(0, \sigma^2 I_n)$, let Y_i be the ith entry of Y, and let $x_i^\top$ be the ith row of X. Then we have

$$f(y_i \mid x_i^\top \beta^*) = \exp\Big(\frac{y_i \cdot x_i^\top \beta^* - (x_i^\top \beta^*)^2/2}{\sigma^2}\Big) \cdot \exp\Big(\frac{y_i^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}\Big).$$

Hence, this is a special of the general model.

Logistic regression Consider independent binary observations $Y_1, \dots, Y_n$, where

$$Y_i \sim \mathsf{Ber}\Big(\frac{1}{1 + \exp(-x_i^\top \beta^*)}\Big). \tag{4.6}$$

Then we have

$$\begin{aligned} f(y_i \mid x_i^\top \beta^*) &= \Big(\frac{1}{1+\exp(-x_i^\top \beta^*)}\Big)^{y_i} \Big(\frac{\exp(-x_i^\top \beta^*)}{1+\exp(-x_i^\top \beta^*)}\Big)^{1-y_i} \\ &= \exp\left(y_i \log \frac{1}{1+\exp(-x_i^\top \beta^*)} + (1-y_i) \log \frac{\exp(-x_i^\top \beta^*)}{1+\exp(-x_i^\top \beta^*)}\right) \\ &= \exp\left(y_i \cdot x_i^\top \beta^* - \log\Big(1+\exp(x_i^\top \beta^*)\Big)\right), \end{aligned}$$

which is again a special case of the general model. What motivates the logistic regression model? The task of classification: β^* is a linear classifier that we aim to learn, each x_i is a vector of d features, and each Y_i represents an outcome of classification.

There is a different class of generalized linear models:

$$Y_i = F(x_i^\top \beta^*) + \varepsilon_i$$

for $i = 1, \dots, n$, where $F : \mathbb{R} \to \mathbb{R}$ is a known, increasing function, and ε_i is zero-mean noise. In the matrix form, we have

$$Y = F(X\beta^*) + \varepsilon, \tag{4.7}$$

where $X \in \mathbb{R}^{n \times d}$, $\beta^* \in \mathbb{R}^d$, $\varepsilon \in \mathbb{R}^n$, and F applies entrywise to $X\beta^*$.

Note that (4.7) reduces to linear regression if $F(t) = t$. Moreover, if F is the logistic function defined by $F(t) = \frac{1}{1+e^{-t}}$ and $\varepsilon_i = Y_i - F(x_i^\top \beta^*)$, then (4.7) reduces to logistic regression (4.6). Another example is the probit model, where F is taken to be the CDF of a standard Gaussian, and $Y_i \sim \mathsf{Ber}(x_i^\top \beta^*)$.

4.3.2 Maximum likelihood estimation for logistic regression

To study logistic regression (4.6), let us focus on the low-dimensional regime where $d \le n$ and X is of rank r. For Gaussian linear regression, Theorem 4.1 shows that the MLE (i.e., the LSE) achieves the rate of estimation r/n. We now show that this is also the case for logistic regression.

To be more precise, for a constant $B > 0$, consider the parameter space

$$\Theta := \{\beta \in \mathbb{R}^d : |x_i^\top \beta| \le B,\ i \in [n]\},$$

and suppose $\beta^* \in \Theta$. The function $F(t) = \frac{1}{1+e^{-t}}$ satisfies that $1 - F(t) = F(-t)$, so we have

$$f(y_i \mid x_i^\top \beta^*) = F(x_i^\top \beta^*)^{y_i} F(-x_i^\top \beta^*)^{1-y_i}.$$

Define $g(t) := \log F(t)$. Then the MLE is

$$\hat{\beta} := \operatorname*{argmax}_{\beta \in \Theta} \sum_{i=1}^n \Big[Y_i g(x_i^\top \beta) + (1-Y_i) g(-x_i^\top \beta)\Big].$$

Theorem 4.9. *Consider the logistic regression model* (4.6) *where* $\beta^* \in \Theta$. *Then the MLE* $\hat{\beta}$ *satisfies that*

$$\frac{1}{n} \mathbb{E} \|X\hat{\beta} - X\beta^*\|_2^2 \lesssim_B \frac{r}{n},$$

where the hidden constant depends on B *and* r *is the rank of* X.

Proof. One can check that $g(t)$ is κ-strongly concave in the sense that

$$g(t) \le g(t^*) + g'(t^*)(t - t^*) - \kappa(t - t^*)^2$$

for all t, t^* with $|t| \le B$, $|t^*| \le B$, for a constant $\kappa = \kappa(B) > 0$. Since $|x_i^\top \beta| \le B$, we see that

$$\begin{aligned}
\ell(\hat{\beta}) = \ell(\hat{\beta} \mid Y) &= \sum_{i=1}^n \Big[Y_i g(x_i^\top \hat{\beta}) + (1 - Y_i) g(-x_i^\top \hat{\beta}) \Big] \\
&\le \sum_{i=1}^n \Big[Y_i g(x_i^\top \beta^*) + (1 - Y_i) g(-x_i^\top \beta^*) \Big] \\
&\quad + \sum_{i=1}^n \Big[Y_i g'(x_i^\top \beta^*) - (1 - Y_i) g'(-x_i^\top \beta^*) \Big] (x_i^\top \hat{\beta} - x_i^\top \beta^*) \\
&\quad - \sum_{i=1}^n \kappa (x_i^\top \hat{\beta} - x_i^\top \beta^*)^2 \\
&= \ell(\beta^*) + \varepsilon^\top (X\hat{\beta} - X\beta^*) - \kappa \|X\hat{\beta} - X\beta^*\|_2^2,
\end{aligned}$$

where ε is the vector with entries $\varepsilon_i := Y_i g'(x_i^\top \beta^*) - (1 - Y_i) g'(-x_i^\top \beta^*)$. In addition, $\ell(\hat{\beta}) \ge \ell(\beta^*)$ by the definition of the MLE, so

$$\|X\hat{\beta} - X\beta^*\|_2^2 \le \frac{1}{\kappa} \varepsilon^\top (X\hat{\beta} - X\beta^*) = \frac{1}{\kappa} \varepsilon^\top X X^\dagger X (\hat{\beta} - \beta^*) \le \frac{1}{\kappa} \|\varepsilon^\top X X^\dagger\|_2 \|X(\hat{\beta} - \beta^*)\|_2$$

by the definition of $X^\dagger$ and the Cauchy-Schwarz inequality. Rearranging terms yields

$$\|X\hat{\beta} - X\beta^*\|_2^2 \le \frac{1}{\kappa^2} \|\varepsilon^\top X X^\dagger\|_2^2.$$

Consider the SVD $X = U\Sigma V^\top$. Then $XX^\dagger = U\Sigma\Sigma^\dagger U^\top$ and thus

$$\|\varepsilon^\top X X^\dagger\|_2^2 = \|\varepsilon^\top U \Sigma \Sigma^\dagger\|_2^2 = \sum_{j=1}^r (\varepsilon^\top u_j)^2,$$

where u_j is the jth column of U. Moreover, the fact that $g'(t) = 1 - F(t) = F(-t)$ implies

$$\varepsilon_i := Y_i \big(1 - F(x_i^\top \beta^*)\big) - (1 - Y_i) F(x_i^\top \beta^*) = Y_i - F(x_i^\top \beta^*).$$

In view of the model (4.6), we see that ε_i is simply the deviation of Y_i from its mean, so $\mathbb{E}[\varepsilon_i^2] \le 1/4$. It follows that

$$\mathbb{E}(\varepsilon^\top u_j)^2 = \sum_{i=1}^n \mathbb{E}\left(\varepsilon_i \cdot (u_j)_i\right)^2 \le \frac{1}{4} \|u\|_2^2 = \frac{1}{4}.$$

Combining everything yields

$$\frac{1}{n} \mathbb{E}\, \|X\hat{\beta} - X\beta^*\|_2^2 \le \frac{1}{\kappa^2 n} \sum_{j=1}^r \mathbb{E}(\varepsilon^\top u_j)^2 \le \frac{r}{4\kappa^2 n},$$

which completes the proof. □

4.4 Nonparametric regression

Let us consider an even more general regression model

$$Y_i = f(x_i) + \varepsilon_i, \quad i \in [n], \tag{4.8}$$

where $x_i \in \mathbb{R}^d$ are the design points, and ε_i are independent noise with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] \le \sigma^2$. The linear and generalized linear regression models assume $f(x_i) = x_i^\top \beta^*$ and $f(x_i) = F(x_i^\top \beta^*)$ respectively. Nonparametric regression, on the other hand, does not assume that there is an underlying parameter vector β^*. Instead, we impose a certain nonparametric assumption on the function f, such as smoothness, monotonicity, or convexity.

4.4.1 Model and estimators

Let us focus on the simple setting where $d = 1$, $x_i = i/n$, and $f : [0,1] \to \mathbb{R}$ is Hölder smooth in the following sense.

Definition 4.10. *Fix $\beta > 0$ and let $\ell := \lfloor \beta \rfloor$. The Hölder class $\Sigma(\beta)$ on $[0,1]$ is defined as the set of ℓ times differentiable functions $f : [0,1] \to \mathbb{R}$ whose ℓth derivative $f^{(\ell)}$ satisfies*

$$|f^{(\ell)}(x) - f^{(\ell)}(x')| \le L|x - x'|^{\beta - \ell}, \quad \text{for all } x, x' \in [0,1],$$

for some constant $L > 0$. We also use $\Sigma(\beta, L)$ to denote all functions f satisfying the above conditions for a fixed $L > 0$.

Note that for a larger β, the condition is stronger and thus $\Sigma(\beta)$ is smaller.

Kernels Before defining the estimator of interest, we first introduce a kernel function $K : \mathbb{R} \to \mathbb{R}$, such as:

- Rectangular kernel: $K(u) = \frac{1}{2}\mathbb{1}\{|u| \le 1\}$;
- Triangular kernel: $K(u) = (1 - |u|)\mathbb{1}\{|u| \le 1\}$;
- Parabolic kernel: $K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}\{|u| \le 1\}$;
- Quartic kernel: $K(u) = \frac{15}{16}(1 - u^2)^2\mathbb{1}\{|u| \le 1\}$;
- Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$.

In the sequel, we focus on kernels that satisfy

$$0 \le K \le 1, \quad \mathrm{supp}(K) \subset [-1,1], \quad K(u) = K(-u), \quad \int K = 1. \tag{4.9}$$

Note that the above kernels except the Gaussian kernel satisfy these conditions.

Nadaraya–Watson estimator Given a kernel K and a bandwidth $h > 0$, a prominent kernel estimator of the regression function f is the Nadaraya–Watson estimator $\hat{f}^{\mathsf{NW}}$, defined as

$$\hat{f}^{\mathsf{NW}}(x) := \frac{\sum_{i=1}^n Y_i K(\frac{x_i - x}{h})}{\sum_{i=1}^n K(\frac{x_i - x}{h})}$$

if $\sum_{i=1}^n K(\frac{x_i - x}{h}) \neq 0$, and $\hat{f}^{\mathsf{NW}}(x) := 0$ otherwise.

More generally, one can consider a linear nonparametric regression estimator $\hat{f}^{\mathsf{linear}}$ of the form

$$\hat{f}^{\mathsf{linear}}(x) = \sum_{i=1}^n Y_i W_i(x),$$

where $W_i(x) = W_{n,i}(x, x_1, \ldots, x_n)$ with $\sum_{i=1}^n W_i(x) = 1$.

Consider the Nadaraya–Watson estimator with the rectangular kernel: If $K(u) = \frac{1}{2}\mathbb{1}\{|u| \leq 1\}$, then $\hat{f}^{\mathsf{NW}}(x)$ is the average of Y_i for which $x_i \in [x - h, x + h]$.

- As $h \to \infty$, we have that $\hat{f}^{\mathsf{NW}}(x)$ tends to $\frac{1}{n}\sum_{i=1}^n Y_i$, which is a constant that does not depend on x. Then the bias is too large, and this situation is called over-smoothing.
- As $h \to 0$, we have $\hat{f}^{\mathsf{NW}}(x_i) = Y_i$ and $\hat{f}^{\mathsf{NW}}(x) = 0$ if $x \neq x_i$ for any $i \in [n]$. Then the variance is too large, and this situation is called under-smoothing.

We need to choose an appropriate bandwidth h to achieve an optimal bias-variance trade-off.

The Nadaraya–Watson estimator can also be interpreted as a weighted LSE:

$$\hat{f}^{\mathsf{NW}}(x) = \operatorname*{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^n (Y_i - \theta)^2 K\Big(\frac{x_i - x}{h}\Big),$$

where the kernel downweights x_i if x_i is far away from x.

Local polynomial estimator Following the intuition of weight least squares, we can design a more sophisticated estimator using the Taylor expansion of f, and with θ replaced by a polynomial of degree $\ell = \lfloor \beta \rfloor$. For $f \in \Sigma(\beta, L)$ where $\beta > 1$, for z close to x, we have

$$f(z) \approx f(x) + f'(x)(z - x) + \frac{f''(x)}{2}(z - x)^2 + \cdots + \frac{f^{(\ell)}(x)}{\ell!}(z - x)^\ell = \theta(x)^\top U\Big(\frac{z - x}{h}\Big),$$

where the vectors $\theta(x) = \theta_h(x)$ and $U(u)$ are defined by

$$\theta(x) = \big(f(x), f'(x)h, f''(x)h^2, \ldots, f^{(\ell)}h^\ell\big)^\top,$$
$$U(u) = \Big(1, u, \frac{u^2}{2}, \ldots, \frac{u^\ell}{\ell!}\Big)^\top.$$

Definition 4.11. *The local polynomial estimator of order ℓ (LP(ℓ) estimator) of $\theta(x)$ is the vector $\hat{\theta}(x)$ in $\mathbb{R}^{\ell+1}$ defined by*

$$\hat{\theta}(x) := \operatorname*{argmin}_{\theta \in \mathbb{R}^{\ell+1}} \sum_{i=1}^n \Big[Y_i - \theta^\top U\Big(\frac{x_i - x}{h}\Big)\Big]^2 K\Big(\frac{x_i - x}{h}\Big).$$

Moreover, the LP(ℓ) estimator of $f(x)$ is defined by

$$\hat{f}(x) := \hat{U}(0)^\top \theta(x) = \hat{\theta}(x)_1.$$

Note that $\hat{f}^{\mathsf{NW}}$ is simply the LP(0) estimator.

We can rewrite $\hat{\theta}(x)$ as

$$\hat{\theta}(x) = \operatorname*{argmin}_{\theta \in \mathbb{R}^{\ell+1}} -2\theta^\top a(x) + \theta^\top B(x)\theta,$$

where the vector $a(x)$ and the matrix $B(x)$ are defined by

$$a(x) = \frac{1}{nh}\sum_{i=1}^n Y_i U\Big(\frac{x_i - x}{h}\Big) K\Big(\frac{x_i - x}{h}\Big),$$
$$B(x) = \frac{1}{nh}\sum_{i=1}^n U\Big(\frac{x_i - x}{h}\Big) U^\top\Big(\frac{x_i - x}{h}\Big) K\Big(\frac{x_i - x}{h}\Big).$$

If $B(x)$ is positive definite, then the solution $\hat{\theta}(x)$ of the quadratic program is given by

$$\hat{\theta}(x) = B(x)^{-1} a(x).$$

Consequently, we have

$$\hat{f}(x) = \sum_{i=1}^n Y_i W_i(x) \tag{4.10}$$

where

$$W_i(x) := \frac{1}{nh} U(0)^\top B(x)^{-1} U\Big(\frac{x_i - x}{h}\Big) K\Big(\frac{x_i - x}{h}\Big). \tag{4.11}$$

In particular, the LP(ℓ) estimator $\hat{f}(x)$ of $f(x)$ is a linear estimator (linear in the data Y_i).

4.4.2 Rates of estimation for local polynomial estimators

Before analyzing the rate of estimation achieved by the local polynomial estimator, we first show that the weights defined by (4.11) are able to "reproduce" any polynomial of degree $\le \ell$.

Proposition 4.12. *Suppose that $B(x)$ is positive definite. Let $Q(x)$ be a polynomial of degree $\le \ell$. Then we have*

$$\sum_{i=1}^n Q(x_i) W_i(x) = Q(x)$$

where the weights are defined in (4.11)*. In particular,*

$$\sum_{i=1}^n W_i(x) = 1, \qquad \sum_{i=1}^n (x_i - x)^k W_i(x) = 0 \text{ for all } k \in [\ell]. \tag{4.12}$$

Proof. Since Q is a polynomial of degree $\le \ell$, we have

$$Q(x_i) = Q(x) + Q'(x)(x_i - x) + \cdots + \frac{Q^{(\ell)}(x)}{\ell!}(x_i - x)^\ell = q(x)^\top U\Big(\frac{x_i - x}{h}\Big)$$

where $q(x) := \big(Q(x), Q'(x)h, \dots, Q^{(\ell)}(x)h^\ell\big)^\top \in \mathbb{R}^{\ell+1}$. Set $Y_i = Q(x_i)$. Then we have

$$\begin{aligned}\hat\theta(x) &= \operatorname*{argmin}_{\theta\in\mathbb{R}^{\ell+1}} \sum_{i=1}^n \Big[Q(x_i) - \theta^\top U\Big(\frac{x_i-x}{h}\Big)\Big]^2 K\Big(\frac{x_i-x}{h}\Big)\\ &= \operatorname*{argmin}_{\theta\in\mathbb{R}^{\ell+1}} \sum_{i=1}^n \Big[\big(q(x)-\theta\big)^\top U\Big(\frac{x_i-x}{h}\Big)\Big]^2 K\Big(\frac{x_i-x}{h}\Big)\\ &= \operatorname*{argmin}_{\theta\in\mathbb{R}^{\ell+1}} \big(q(x)-\theta\big)^\top B(x)\big(q(x)-\theta\big).\end{aligned}$$

Since $B(x)$ is positive, we have $\hat\theta(x) = q(x)$ and therefore $\hat f(x) = \hat\theta(x)_1 = Q(x)$. The reproducing property then follows from (4.10).

For (4.12), take respectively $Q(t) \equiv 1$ and $Q(t) = (t-x)^k$. □

In addition, we impose an assumption: The smallest eigenvalue $\lambda_{\min}(B(x))$ of $B(x)$ satisfies

$$\lambda_{\min}(B(x)) \ge \lambda_0 \tag{4.13}$$

for any $x \in [0,1]$ for a constant $\lambda_0 > 0$. In particular, $\|B(x)^{-1}v\|_2 \le \|v\|_2/\lambda_0$ for any $v \in \mathbb{R}^{\ell+1}$.

Lemma 4.13. *Under assumptions* (4.9) *and* (4.13)*, the weights defined in* (4.11) *satisfy*

- $W_i(x) = 0$ *if* $|x - x_i| > h$ *for any* $i \in [n]$*;*
- $|W_i(x)| \le \frac{2}{nh\lambda_0}$ *for any* $x \in [0,1]$ *and* $i \in [n]$*;*
- $\sum_{i=1}^n |W_i(x)| \le \frac{8}{\lambda_0}$ *for any* $x \in [0,1]$ *if* $h \ge 1/(2n)$*.*

Proof. The first statement is obvious since $\operatorname{supp}(K) \subset [-1,1]$.

Using $\|U(0)\|_2 = 1$ and $\|B(x)^{-1}v\|_2 \le \|v\|_2/\lambda_0$, we obtain

$$\begin{aligned}|W_i(x)| &\le \frac{1}{nh}\big\|U(0)\big\|_2\Big\|B(x)^{-1}U\Big(\frac{x_i-x}{h}\Big)K\Big(\frac{x_i-x}{h}\Big)\Big\|_2\\ &\le \frac{1}{nh\lambda_0}\Big\|U\Big(\frac{x_i-x}{h}\Big)K\Big(\frac{x_i-x}{h}\Big)\Big\|_2\\ &\le \frac{1}{nh\lambda_0}\big\|U(1)\big\|_2 \le \frac{1}{nh\lambda_0}\Big(1+1+\frac{1}{2^2}+\cdots+\frac{1}{(\ell!)^2}\Big)^{1/2} \le \frac{2}{nh\lambda_0}.\end{aligned}$$

In addition, we have

$$\sum_{i=1}^n |W_i(x)| \le \sum_{i=1}^n \frac{2}{nh\lambda_0}\mathbb{1}\{|x-x_i| \le h\} \le \frac{2}{nh\lambda_0}(2hn+1) \le \frac{4}{\lambda_0} + \frac{2}{nh\lambda_0},$$

finishing the proof. □

We now study the rate of estimation for the local polynomial estimator $\hat f(x)$ of $f(x)$ in terms of the mean squared risk $\mathbb{E}\big(\hat f(x) - f(x)\big)^2$. To this end, we consider the bias–variance decomposition:

$$\mathbb{E}\big(\hat f(x) - f(x)\big)^2 = \mathsf{Bias}(x)^2 + \mathsf{Var}(x),$$

where the bias and variance of $\hat f(x)$ are given, respectively, by

$$\mathsf{Bias}(x) := \mathbb{E}[\hat f(x)] - f(x), \quad \mathsf{Var}(x) := \mathbb{E}\big(\hat f(x) - \mathbb{E}[\hat f(x)]\big)^2.$$

Theorem 4.14. *Suppose that $f : [0,1] \to \mathbb{R}$ belongs to the Hölder class $\Sigma(\beta, L)$ for $\beta, L > 0$. Consider the model $Y_i = f(x_i) + \varepsilon_i$ where $i \in [n]$, $x_i = i/n$, and ε_i are independent with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] \le \sigma^2$. Let $\hat{f}$ be the LP(ℓ) estimator of f with $\ell = \lfloor \beta \rfloor$ and kernel K satisfying (4.9). Assume (4.13) and $h \ge 1/(2n)$.*

- *For any $x \in [0,1]$, we have the following upper bounds on the bias and the variance of $\hat{f}$:*

$$|\,\mathsf{Bias}(x)| \le \frac{8Lh^\beta}{\ell!\lambda_0}, \qquad \mathsf{Var}(x) \le \frac{16}{\lambda_0^2 nh}.$$

- *As a result, for $C = C(\beta, L, \lambda_0, \sigma) := \frac{32}{\lambda_0^2}\big(\frac{2L}{\ell!}\big)^{\frac{2}{2\beta+1}} \sigma^{\frac{4\beta}{2\beta+1}}$, we have*

$$\mathbb{E}\big(\hat{f}(x) - f(x)\big)^2 \le C n^{\frac{-2\beta}{2\beta+1}}.$$

- *For $g : [0,1] \to \mathbb{R}$, let $\|g\|_{L^2}^2 := \int_0^1 g(x)^2\,dx$. Then we have*

$$\mathbb{E}\,\|\hat{f} - f\|_{L^2}^2 \le C n^{\frac{-2\beta}{2\beta+1}}.$$

Proof. Applying (4.12) and the Taylor expansion, we obtain

$$\begin{aligned}
\mathsf{Bias}(x) &= \sum_{i=1}^n \mathbb{E}[Y_i] W_i(x) - f(x) = \sum_{i=1}^n \big(f(x_i) - f(x)\big) W_i(x) \\
&= \sum_{i=1}^n \bigg[\sum_{k=1}^{\ell-1} \frac{f^{(k)}(x)}{k!}(x_i - x)^k + \frac{f^{(\ell)}\big(x + \tau_i(x_i - x)\big)}{\ell!}(x_i - x)^\ell \bigg] W_i(x) \\
&= \sum_{i=1}^n \frac{f^{(\ell)}\big(x + \tau_i(x_i - x)\big) - f^{(\ell)}(x)}{\ell!}(x_i - x)^\ell W_i(x)
\end{aligned}$$

for some $\tau_i \in [0,1]$, where note that we could insert a term $-f^{(\ell)}(x)$ in the numerator since the sum vanishes. It follows from $f \in \Sigma(\beta, L)$ and Lemma 4.13 that

$$|\,\mathsf{Bias}(x)| \le \sum_{i=1}^n \frac{L|x_i - x|^\beta}{\ell!}|W_i(x)| \le \frac{Lh^\beta}{\ell!}\sum_{i=1}^n |W_i(x)| \le \frac{8Lh^\beta}{\ell!\lambda_0}.$$

The variance can be bounded, using Lemma 4.13 again, as

$$\mathsf{Var}(x) = \mathbb{E}\Big(\sum_{i=1}^n \varepsilon_i W_i(x)\Big)^2 = \sum_{i=1}^n W_i(x)^2\,\mathbb{E}[\varepsilon_i^2] \le \sigma^2\Big(\max_{i\in[n]}|W_i(x)|\Big)\sum_{i=1}^n |W_i(x)| \le \frac{16\sigma^2}{\lambda_0^2 nh}.$$

Therefore, choosing $h = (\frac{\ell!\sigma}{2L})^{\frac{2}{2\beta+1}} n^{\frac{-1}{2\beta+1}}$ yields

$$\mathbb{E}\big(\hat{f}(x) - f(x)\big)^2 \le \frac{64L^2}{(\ell!)^2\lambda_0^2}h^{2\beta} + \frac{16\sigma^2}{\lambda_0^2 n}\frac{1}{h} = \frac{32\sigma^2}{\lambda_0^2}\Big(\frac{2L}{\ell!\sigma}\Big)^{\frac{2}{2\beta+1}} n^{\frac{-2\beta}{2\beta+1}}.$$

Integrating over $x \in [0,1]$ gives the last statement. □

Some remarks:

- Note that we have the rate of estimation $n^{\frac{-2\beta}{2\beta+1}}$ for the pointwise risk $\mathbb{E}\left(\hat{f}(x) - f(x)\right)^2$ at each $x \in [0,1]$. This is stronger than bounding an average risk like $\mathbb{E}\,\|\hat{f} - f\|_2^2$.
- As β grows, the function becomes smoother, so the rate $n^{\frac{-2\beta}{2\beta+1}}$ improves as expected. In particular, as $\beta \to \infty$, the nonparametric rate $n^{\frac{-2\beta}{2\beta+1}}$ tends to the parametric rate $1/n$.
- Here we choose h depending on the smoothness parameters β and L. In practice, we may not know how smooth the function is a priori. To address this issue, we can in fact design adaptive estimators that do not depend these smoothness parameters.
- In dimension d, when we estimate a β-Hölder smooth function $f : [0,1]^d \to \mathbb{R}$ from n noisy observations, one can similarly establish the rate of estimation $n^{\frac{-2\beta}{2\beta+d}}$.

The name "nonparametric" simply refers to the setup where there is no obvious parameter (like β in linear regression). In fact, it is without loss of generality to focus on the framework of parametric estimation by viewing nonparametric estimation as a setup which has a "large" parameter space. For example, for nonparametric regression discussed in these two sections, we can view the Hölder class $\Sigma(\beta, L)$ as the parameter space, and view the function f as the parameter.

Chapter 5

Information-theoretic lower bounds

In the previous chapter, we studied several regression models and proved rates of estimations for specific estimators. That is, we established an upper bound on the minimax risk of the form

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} R\big(\theta, \hat{\theta}_n\big) \lesssim r_n,$$

for some rate r_n. Can we show that the minimax risk is also lower bounded by some rate s_n which is hopefully equal to r_n? If so, this suggests that the estimator that achieves the upper bound is the essentially the best we can hope for in the minimax sense.

5.1 Reduction to hypothesis testing

To establish such minimax lower bounds, we reduce the problem to hypothesis testing and use information-theoretic tools. Such lower bounds are referred to as statistical lower bounds or information-theoretic lower bounds.

Reduction to bounds in probability Suppose that the risk we would like to lower bound is of the form $R(\theta, \hat{\theta}) = \mathbb{E}_\theta[d(\theta, \hat{\theta})^2]$ for some pseudometric $d(\cdot, \cdot)$. If we can establish

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta\{d(\theta, \hat{\theta})^2 \geq s_n\} \geq c \tag{5.1}$$

for a universal constant $c > 0$, then by Markov's inequality,

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[d(\theta, \hat{\theta})^2] \geq c\, s_n.$$

Reduction to a finite number of parameters For any $\theta_1, \ldots, \theta_M \in \Theta$, if we can establish

$$\inf_{\hat{\theta}} \max_{i \in [M]} \mathbb{P}_{\theta_i}\{d(\theta_i, \hat{\theta})^2 \geq s_n\} \geq c, \tag{5.2}$$

then (5.1) obviously holds. The difficulty for proving lower bounds usually lies in how to appropriately choose $\theta_1, \ldots, \theta_M$, which we call hypotheses.

Reduction to hypothesis testing The crucial requirement is that $d(\theta_i, \theta_j)^2 \geq 4s_n$ for any distinct $i, j \in [M]$. Given any estimator $\hat{\theta}$, consider the minimum distance test

$$\psi(\hat{\theta}) := \operatorname*{argmin}_{i \in [M]} d(\theta_i, \hat{\theta}).$$

For any $i \in [M]$ such that $\psi(\hat{\theta}) \neq i$, we have

$$d(\theta_i, \hat{\theta}) \geq d(\theta_i, \theta_{\psi(\hat{\theta})}) - d(\theta_{\psi(\hat{\theta})}, \hat{\theta}) \geq d(\theta_i, \theta_{\psi(\hat{\theta})}) - d(\theta_i, \hat{\theta}),$$

so that

$$d(\theta_i, \hat{\theta}) \geq \frac{1}{2} d(\theta_i, \theta_{\psi(\hat{\theta})}) \geq \sqrt{s_n}.$$

Therefore, we obtain

$$\inf_{\hat{\theta}} \max_{i \in [M]} \mathbb{P}_{\theta_i}\{d(\theta_i, \hat{\theta})^2 \geq s_n\} \geq \inf_{\hat{\theta}} \max_{i \in [M]} \mathbb{P}_{\theta_i}\{\psi(\hat{\theta}) \neq i\} \geq \inf_{\psi} \max_{i \in [M]} \mathbb{P}_{\theta_i}\{\psi \neq i\},$$

where the infimum is taken over all tests ψ that are measurable with respect to the observations and take values in $[M]$. We have proved the following theorem.

Theorem 5.1. *Let $\theta_1, \ldots, \theta_M \in \Theta$ be such that $d(\theta_i, \theta_j)^2 \geq 4s_n$ for any distinct $i, j \in [M]$. Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[d(\theta, \hat{\theta})^2] \geq s_n \cdot \inf_{\psi} \max_{i \in [M]} \mathbb{P}_{\theta_i}\{\psi \neq i\},$$

where the infimum on the right-hand side is taken over all tests ψ that are measurable with respect to the observations and take values in $[M]$.

5.2 Le Cam's two-point method

5.2.1 General theory

To study lower bounds for hypothesis testing, let us start with the simplest case where we have two probability measures $\mathbb{P}_0$ and $\mathbb{P}_1$. Suppose that $\mathbb{P}_0$ and $\mathbb{P}_1$ have densities p_0 and p_1 respectively, with respect to a measure μ. We write $\int f = \int f(x)\, d\mu(x)$. Given an observation X from either $\mathbb{P}_0$ or $\mathbb{P}_1$, consider a test $\psi = \psi(X) \in \{0, 1\}$.

Lemma 5.2 (Neyman–Pearson)**.** *For any test ψ, the sum of the type I error and the type II error satisfies*

$$\mathbb{P}_0\{\psi = 1\} + \mathbb{P}_1\{\psi = 0\} \geq \int \min(p_0, p_1).$$

Moreover, the equality holds for the likelihood ratio test $\psi^* := \mathbb{1}\{p_1/p_0 \geq 1\}$.

This is the Neyman–Pearson lemma, although the name sometimes refers to a different formulation.

Proof. First note that

$$\begin{aligned}
\mathbb{P}_0\{\psi^* = 1\} + \mathbb{P}_1\{\psi^* = 0\} &= \int_{\{\psi^*=1\}} p_0 + \int_{\{\psi^*=0\}} p_1 \\
&= \int_{\{p_1 \geq p_0\}} p_0 + \int_{\{p_1 < p_0\}} p_1 \\
&= \int_{\{p_1 \geq p_0\}} \min(p_0, p_1) + \int_{\{p_1 < p_0\}} \min(p_0, p_1) \\
&= \int \min(p_0, p_1).
\end{aligned}$$

Next, for any test ψ, define $R := \{\psi = 1\}$. Also define $R^* := \{p_1 \geq p_0\}$. Then we have

$$\begin{aligned}
\mathbb{P}_0\{\psi = 1\} + \mathbb{P}_1\{\psi = 0\} &= \mathbb{P}_0\{R\} + 1 - \mathbb{P}_1\{R\} \\
&= 1 + \int_R (p_0 - p_1) \\
&= 1 + \int_{R \cap R^*} (p_0 - p_1) + \int_{R \cap (R^*)^c} (p_0 - p_1) \\
&= 1 - \int_{R \cap R^*} |p_0 - p_1| + \int_{R \cap (R^*)^c} |p_0 - p_1| \\
&= 1 - \int |p_0 - p_1| \big(\mathbb{1}\{R \cap R^*\} - \mathbb{1}\{R \cap (R^*)^c\} \big),
\end{aligned}$$

which is minimized at $R = R^*$. □

The total variation distance between $\mathbb{P}_0$ and $\mathbb{P}_1$ is defined as any of the following quantities:

$$\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) = \frac{1}{2} \int |p_0 - p_1| = 1 - \int \min(p_0, p_1) = 1 - \inf_{\psi} \big[\mathbb{P}_0\{\psi = 1\} + \mathbb{P}_1\{\psi = 0\} \big].$$

The equivalence of the first two definitions is proved as a homework problem. The above lemma gives the second equivalence.

Combining Theorem 5.1 and Lemma 5.2 with the definition of the total variation distance, we have established Le Cam's two-point bound.

Theorem 5.3 (Le Cam)**.** *For any $\theta_0, \theta_1 \in \Theta$, we have*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[d(\theta, \hat{\theta})^2] \geq \frac{1}{8} d(\theta_0, \theta_1)^2 \big[1 - \mathsf{TV}(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}) \big].$$

A couple of remarks: The constant 1/8 can be refined to 1/4 with an improved argument. Moreover, by the chain of inequalities between f-divergences proved in the homework, TV in the above theorem can be replaced by H, $\sqrt{\mathsf{KL}}$, or $\sqrt{\chi^2}$, which are typically easier to compute.

5.2.2 Lower bounds for nonparametric regression at a point

We establish a lemma which will be useful later.

Lemma 5.4. *We have*

$$\mathsf{KL}\big(\mathsf{N}(\mu_1,\sigma^2 I_d),\mathsf{N}(\mu_2,\sigma^2 I_d)\big)=\frac{\|\mu_1-\mu_2\|_2^2}{2\sigma^2}.$$

Proof. The one-dimensional case follows from direct computation

$$\mathsf{KL}\big(\mathsf{N}(\mu_1,\sigma^2),\mathsf{N}(\mu_2,\sigma^2)\big)=\int\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}\Big[-\frac{(x-\mu_1)^2}{2\sigma^2}+\frac{(x-\mu_2)^2}{2\sigma^2}\Big]dx=\frac{(\mu_1-\mu_2)^2}{2\sigma^2}.$$

The multivariate case follows from the tensorization property of KL established in the homework

$$\mathsf{KL}\big(\mathsf{N}(\mu_1,\sigma^2 I_d),\mathsf{N}(\mu_2,\sigma^2 I_d)\big)=\sum_{i=1}^d\mathsf{KL}\big(\mathsf{N}((\mu_1)_i,\sigma^2),\mathsf{N}((\mu_2)_i,\sigma^2)\big)=\frac{\|\mu_1-\mu_2\|_2^2}{2\sigma^2}.$$

□

Consider the nonparametric regression model (4.8), where we assume that $x_i=i/n$ and ε_i are i.i.d $\mathsf{N}(0,\sigma^2)$ noise for $i\in[n]$. We aim to establish a minimax lower bound over the Hölder class $\Sigma(\beta,L)$ where $\beta,L>0$, for the distance $d(f,g)=|f(x_0)-g(x_0)|$ at a fixed point $x_0\in[0,1]$.

Theorem 5.5. *For any $x_0\in[0,1]$, there exists a constant $c=c_2(\beta)L^{\frac{2}{2\beta+1}}\sigma^{\frac{4\beta}{2\beta+1}}>0$ such that*

$$\inf_{\hat f}\sup_{f\in\Sigma(\beta,L)}\mathbb{E}_f\big[\big(\hat f(x_0)-f(x_0)\big)^2\big]\ge c\,n^{\frac{-2\beta}{2\beta+1}}.$$

Proof. Let us consider the hypotheses $f_0\equiv 0$ and

$$f_1(x)=Lh^\beta K\Big(\frac{x-x_0}{h}\Big)$$

for all $x\in[0,1]$, where

$$h=c_0 n^{\frac{-1}{2\beta+1}},\quad c_0=c_0(\beta,L,\sigma)>0,$$

and the function K is defined as

$$K(u)=c_1\exp\Big(\frac{-1}{1-4u^2}\Big)\mathbb{1}\{|u|\le 1/2\},\quad c_1=c_1(\beta)>0. \tag{5.3}$$

Smoothness First, we check that $f_1\in\Sigma(\beta,L)$. For $\ell=\lfloor\beta\rfloor$, we have

$$f_1^{(\ell)}(x)=Lh^{\beta-\ell}K^{(\ell)}\Big(\frac{x-x_0}{h}\Big).$$

Moreover, we take $c_1>0$ to be sufficiently small depending on β, so that $K^{(\ell+1)}(u)\le 1$. Then for $-1/2\le u,u'\le 1/2$, it holds that

$$|K^{(\ell)}(u)-K^{(\ell)}(u')|\le|u-u'|\le|u-u'|^{\beta-\ell}.$$

Therefore, we obtain

$$|f_1^{(\ell)}(x)-f_1^{(\ell)}(x')|\le Lh^{\beta-\ell}\Big|\frac{x-x'}{h}\Big|^{\beta-\ell}=L|x-x'|^{\beta-\ell}.$$

Separation We have

$$d(f_0, f_1) = f_1(x_0) = Lh^\beta K(0) = Lc_0^\beta n^{\frac{-\beta}{2\beta+1}} c_1/e.$$

KL divergence The joint distribution of $(Y_1, \dots, Y_n)$ is $\mathsf{N}(0, \sigma^2 I_n)$ under f_0, and it is

$$\otimes_{i=1}^n \mathsf{N}(f_1(x_i), \sigma^2) = \mathsf{N}\Big(\big(f_1(x_i)\big)_{i=1}^n, \sigma^2 I_n\Big)$$

under f_1. By Lemma 5.4, we have

$$\begin{aligned}
&\mathsf{KL}(\mathbb{P}_{f_0}, \mathbb{P}_{f_1}) = \sum_{i=1}^n \frac{f_1(x_i)^2}{2\sigma^2} = \frac{L^2 h^{2\beta}}{2\sigma^2} \sum_{i=1}^n K\Big(\frac{x_i - x_0}{h}\Big)^2 \\
&\le \frac{L^2 h^{2\beta} c_1^2}{2\sigma^2 e^2} \sum_{i=1}^n \mathbb{1}\{|x_i - x_0| \le h/2\} \le \frac{L^2 h^{2\beta} c_1^2}{2\sigma^2 e^2}(nh+1) \le \frac{L^2 c_1^2}{\sigma^2 e^2} c_0^{2\beta+1} = \frac{1}{4},
\end{aligned}$$

if $c_0 = \big(\frac{\sigma^2 e^2}{4L^2 c_1^2}\big)^{\frac{1}{2\beta+1}}$. As a result, $\mathsf{TV}(\mathbb{P}_{f_0}, \mathbb{P}_{f_1}) \le \sqrt{\mathsf{KL}(\mathbb{P}_{f_0}, \mathbb{P}_{f_1})} \le 1/2$ by a homework problem.

The proof is complete thanks to Theorem 5.3. □

Note that this minimax lower bound matches the pointwise upper bound in Theorem 4.14 up to a constant factor. However, the two-point method is not sufficient for establishing a matching lower bound on the integrated error $\|\hat{f} - f\|_{L^2}^2$.

5.3 Assouad's lemma

5.3.1 General theory

To exhibit the difficulty of applying the two-point method in multivariate estimation, let us consider estimating $\mu \in \mathbb{R}^d$ given i.i.d. observations $X_1, \dots, X_n \sim \mathsf{N}(\mu, I_d)$. For the empirical mean $\bar{X}$, we achieve the minimax risk

$$\mathbb{E}\,\|\bar{X} - \mu\|_2^2 = \sum_{i=1}^d \mathbb{E}(\bar{X}_i - \mu_i)^2 = \frac{d}{n}.$$

To establish a lower bound, we apply the two-point method on the hypotheses $\mathbb{P}_0 = \otimes_{i=1}^n \mathsf{N}(v, I_d)$ and $\mathbb{P}_1 = \otimes_{i=1}^n \mathsf{N}(w, I_d)$. Then

$$\mathsf{KL}(\mathbb{P}_0, \mathbb{P}_1) = \sum_{i=1}^n \mathsf{KL}\big(\mathsf{N}(v, I_d), \mathsf{N}(w, I_d)\big) = \frac{n}{2}\|v - w\|_2^2.$$

Therefore, if we choose $v, w \in \mathbb{R}^d$ such that $\|v - w\|_2^2 = 1/n$, then

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathbb{R}^d} \mathbb{E}\,\|\hat{\mu} - \mu\|_2^2 \gtrsim \frac{1}{n}.$$

This is optimal in the sample size n but not in the dimension d unless it is a constant.

One powerful tool for proving high-dimensional lower bounds is the following theorem called Assouad's lemma.

Theorem 5.6 (Assouad)**.** *Let $\{\mathbb{P}_\omega : \omega \in \{0,1\}^d\}$ be a set of 2^d probability measures, and let $\mathbb{E}_\omega$ denote the corresponding expectations. Then*

$$\inf_{\hat{\omega}} \max_{\omega \in \{0,1\}^d} \mathbb{E}_\omega \, \rho(\hat{\omega}, \omega) \geq \frac{d}{2} \min_{\rho(\omega,\omega')=1} \big[1 - \mathsf{TV}(\mathbb{P}_\omega, \mathbb{P}_{\omega'})\big],$$

where the infimum is over all estimators $\hat{\omega}$ taking values in $\{0,1\}^d$, and $\rho(\hat{\omega}, \omega) = \sum_{i=1}^d \mathbb{1}\{\hat{\omega}_i \neq \omega_i\}$ denotes the Hamming distance between $\hat{\omega}$ and ω.

Proof. We have

$$\begin{aligned}
\max_{\omega \in \{0,1\}^d} \mathbb{E}_\omega \, \rho(\hat{\omega}, \omega) &\geq \frac{1}{2^d} \sum_{\omega \in \{0,1\}^d} \mathbb{E}_\omega \, \rho(\hat{\omega}, \omega) \\
&= \frac{1}{2^d} \sum_{\omega \in \{0,1\}^d} \mathbb{E}_\omega \sum_{i=1}^d \mathbb{1}\{\hat{\omega}_i \neq \omega_i\} = \frac{1}{2^d} \sum_{i=1}^d \sum_{\omega \in \{0,1\}^d} \mathbb{P}_\omega\{\hat{\omega}_i \neq \omega_i\}.
\end{aligned}$$

Let $\omega_{-i} \in \{0,1\}^{d-1}$ denote the subvector of ω with its ith entry removed. Let $(\omega_{-i}, 1)$ denote the vector ω whose ith entry is equal to 1. Then

$$\begin{aligned}
\max_{\omega \in \{0,1\}^d} \mathbb{E}_\omega \, \rho(\hat{\omega}, \omega) &\geq \frac{1}{2^d} \sum_{i=1}^d \sum_{\omega_{-i} \in \{0,1\}^{d-1}} \Big(\mathbb{P}_{(\omega_{-i},1)}\{\hat{\omega}_i = 0\} + \mathbb{P}_{(\omega_{-i},0)}\{\hat{\omega}_i = 1\} \Big) \\
&\geq \frac{1}{2^d} \sum_{i=1}^d \sum_{\omega_{-i} \in \{0,1\}^{d-1}} \Big(1 - \mathsf{TV}\left(\mathbb{P}_{(\omega_{-i},1)}, \mathbb{P}_{(\omega_{-i},0)} \right) \Big) \\
&\geq \frac{d}{2} \min_{\rho(\omega,\omega')=1} \Big(1 - \mathsf{TV}(\mathbb{P}_\omega, \mathbb{P}_{\omega'}) \Big).
\end{aligned}$$

□

Lemma 5.7. *In the problem of estimating $\theta \in \Theta$ where Θ is a closed set, let $\tilde{\theta}$ denote an estimator that takes values in Θ, and let $\hat{\theta}$ denote an arbitrary estimator. Then we have*

$$\frac{1}{4} \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} \mathbb{E}[d(\theta, \tilde{\theta})^2] \leq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}[d(\theta, \hat{\theta})^2] \leq \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} \mathbb{E}[d(\theta, \tilde{\theta})^2].$$

Proof. The second inequality is trivial. Let us focus on the first inequality. Consider an arbitrary estimator $\hat{\theta}$. Define

$$\tilde{\theta} := \operatorname*{argmin}_{\theta \in \Theta} d(\theta, \hat{\theta}),$$

which is an estimator that takes values in Θ. Then for any $\theta \in \Theta$,

$$d(\tilde{\theta}, \theta)^2 \leq \Big(d(\tilde{\theta}, \hat{\theta}) + d(\hat{\theta}, \theta) \Big)^2 \leq 2d(\tilde{\theta}, \hat{\theta})^2 + 2d(\hat{\theta}, \theta)^2 \leq 4d(\hat{\theta}, \theta)^2.$$

As a result,

$$\sup_{\theta \in \Theta} \mathbb{E}[d(\theta, \tilde{\theta})^2] \leq 4 \sup_{\theta \in \Theta} \mathbb{E}[d(\theta, \hat{\theta})^2].$$

Taking an infimum over $\hat{\theta}$ and then over $\tilde{\theta}$ completes the proof. □

Corollary 5.8. *Suppose that to each $\omega \in \{0,1\}^d$, we can associate a parameter $\theta_\omega \in \Theta$ such that*

$$d(\theta_\omega, \theta_{\omega'})^2 \geq c_n \rho(\omega, \omega')$$

for a constant $c_n > 0$ that may depend on the sample size n. Let $\mathbb{P}_\omega = \mathbb{P}_{\theta_\omega}$ denote the model at θ_ω for $\omega \in \{0,1\}^d$. Then we have

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\, d(\hat{\theta}, \theta)^2 \geq \frac{c_n d}{8} \min_{\rho(\omega,\omega')=1} \big[1 - \mathsf{TV}(\mathbb{P}_\omega, \mathbb{P}_{\omega'})\big],$$

Proof. By Assouad's lemma, we have

$$\inf_{\tilde{\theta}} \max_{\omega \in \{0,1\}^d} \mathbb{E}\, d(\tilde{\theta}, \theta_\omega)^2 \geq c_n \inf_{\hat{\omega}} \max_{\omega \in \{0,1\}^d} \mathbb{E}_\omega\, \rho(\hat{\omega}, \omega) \geq c_n \frac{d}{2} \min_{\rho(\omega,\omega')=1} \big[1 - \mathsf{TV}(\mathbb{P}_\omega, \mathbb{P}_{\omega'})\big],$$

where the first infimum is over all $\tilde{\theta}$ that takes values in $\big\{\theta_\omega : \omega \in \{0,1\}^d\big\}$. Lemma 5.7 implies

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\, d(\hat{\theta}, \theta)^2 \geq \frac{1}{4} \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\, d(\tilde{\theta}, \theta)^2 \geq \frac{1}{4} \inf_{\tilde{\theta}} \max_{\omega \in \{0,1\}^d} \mathbb{E}\, d(\tilde{\theta}, \theta_\omega)^2$$

where the first infimum is over arbitrary estimators. It suffices to combine the two bounds. □

5.3.2 Applications

Gaussian mean estimation The typically way of applying Assouad's lemma is to associate each ω with a parameter. For example, for the above Gaussian mean estimation problem, we define $\mu_\omega = \omega/\sqrt{n} \in \mathbb{R}^d$ for each $\omega \in \{0,1\}^d$. Then

$$\rho(\omega, \omega') = \|\omega - \omega'\|_2^2 = n\|\mu_\omega - \mu_{\omega'}\|_2^2,$$

and $\mathbb{P}_\omega = \otimes_{i=1}^n \mathsf{N}(\mu_\omega, I_d)$. If $1 = \rho(\omega, \omega') = n\|\mu_\omega - \mu_{\omega'}\|_2^2$, then by Lemma 5.4,

$$\mathsf{TV}(\mathbb{P}_\omega, \mathbb{P}_{\omega'}) \leq \sqrt{\mathsf{KL}(\mathbb{P}_\omega, \mathbb{P}_{\omega'})} = \sqrt{n \cdot (1/2) \cdot \|\mu_\omega - \mu_{\omega'}\|_2^2} = 1/\sqrt{2}.$$

Therefore, Corollary 5.8 implies that

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathbb{R}^d} \mathbb{E}\, \|\hat{\mu} - \mu\|_2^2 \gtrsim \frac{1}{n} \inf_{\hat{\omega}} \max_{\omega \in \{0,1\}^d} \mathbb{E}_\omega\, \rho(\hat{\omega}, \omega) \gtrsim \frac{d}{n}.$$

Linear regression Consider the linear regression model $Y = X\beta^* + \varepsilon$ where $\varepsilon \sim \mathsf{N}(0, \sigma^2 I_n)$. Let r be the rank of $X \in \mathbb{R}^{n \times d}$, and let $U \in \mathbb{R}^{n \times r}$ be a matrix whose columns form an orthonormal basis of the column space of X. We associate each $\omega \in \{0,1\}^r$ with a vector $\beta_\omega \in \mathbb{R}^d$ such that

$$U\omega = \frac{1}{\sigma^2} X\beta_\omega.$$

Then we have

$$\rho(\omega, \omega') = \|\omega - \omega'\|_2^2 = \|U\omega - U\omega'\|_2^2 = \frac{1}{\sigma^2}\|X\beta_\omega - X\beta_{\omega'}\|_2^2.$$

In addition, $\mathbb{P}_\omega = \mathsf{N}(X\beta_\omega, I_n)$. Hence, if $1 = \rho(\omega, \omega') = \frac{1}{\sigma^2}\|X\beta_\omega - X\beta_{\omega'}\|_2^2$, then by Lemma 5.4,

$$\mathsf{TV}(\mathbb{P}_\omega, \mathbb{P}_{\omega'}) \leq \sqrt{\mathsf{KL}(\mathbb{P}_\omega, \mathbb{P}_{\omega'})} = \sqrt{\frac{1}{2\sigma^2}\|X\beta_\omega - X\beta_{\omega'}\|_2^2} = 1/\sqrt{2}.$$

Applying Corollary 5.8, we obtain

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathbb{R}^d} \frac{1}{n} \mathbb{E}_\beta\, \|X\hat{\beta} - X\beta\|_2^2 \gtrsim \frac{\sigma^2}{n} \inf_{\hat{\omega}} \max_{\omega \in \{0,1\}^d} \mathbb{E}_\omega\, \rho(\hat{\omega}, \omega) \gtrsim \sigma^2 \frac{r}{n}.$$

5.4 Fano's inequality

5.4.1 General theory

We move on to study minimax lower bounds based on multiple hypothesis testing. Recall that to apply Theorem 5.1, we need to find separated parameters $\theta_1, \dots, \theta_M \in \Theta$ such that

$$\inf_{\psi} \max_{i \in [M]} \mathbb{P}_i\{\psi \neq i\} \geq c,$$

where we write $\mathbb{P}_i = \mathbb{P}_{\theta_i}$ and the infimum is over all tests ψ. To this end, we use a special case of Fano's inequality. Let us start with a lemma.

Lemma 5.9. *Define a function $h(p, q)$ for $p, q \in (0, 1)$ by*

$$h(p, q) := \mathsf{KL}(\mathsf{Ber}(p), \mathsf{Ber}(q)) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

Then h is convex.

Proof. We first show that the function $(p, q) \mapsto p \log \frac{p}{q}$ is convex for $p, q > 0$. The Hessian of the function is $H = \begin{pmatrix} 1/p & -1/q \\ -1/q & p/q^2 \end{pmatrix}$. We have $\det(H) = 0$ and $\mathsf{tr}(H) > 0$, so H is positive semidefinite.

Moreover, since the composition of a convex function with a linear function is convex, and a sum of two convex functions is convex, we see that h is convex. □

Theorem 5.10 (Data processing)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures that are absolutely continuous with respect to each other. For $X \sim \mathbb{P}$, $Y \sim \mathbb{Q}$, and a function g, we have*

$$\mathsf{KL}(g(X), g(Y)) \leq \mathsf{KL}(X, Y) = \mathsf{KL}(\mathbb{P}, \mathbb{Q}).$$

Proof. Let f_X denote the density of X. Then f_X can be identified with $f_{X,g(X)} = f_{g(X)} f_{X|g(X)}$. It follows that

$$\begin{aligned}
&\mathbb{E}_X \left[\log \frac{f_X}{f_Y} \right] = \mathbb{E}_{X,g(X)} \left[\log \frac{f_{X,g(X)}}{f_{Y,g(Y)}} \right] = \mathbb{E}_{X,g(X)} \left[\log \frac{f_{g(X)}}{f_{g(Y)}} + \log \frac{f_{X|g(X)}}{f_{Y|g(Y)}} \right] \\
&= \mathbb{E}_{g(X)} \left[\mathbb{E}_X \left[\log \frac{f_{g(X)}}{f_{g(Y)}} \,\middle|\, g(X) \right] \right] + \mathbb{E}_{g(X)} \left[\mathbb{E}_X \left[\log \frac{f_{X|g(X)}}{f_{Y|g(Y)}} \,\middle|\, g(X) \right] \right] \geq \mathbb{E}_{g(X)} \left[\log \frac{f_{g(X)}}{f_{g(Y)}} \right],
\end{aligned}$$

where we used that the conditional KL divergence $\mathbb{E}_X \left[\log \frac{f_{X|g(X)}}{f_{Y|g(Y)}} \,\middle|\, g(X) \right]$ is nonnegative. □

Theorem 5.11 (Fano's inequality)**.** *Let $\mathbb{P}_1, \dots, \mathbb{P}_M$ be probability measures that are absolutely continuous with respect to each other. Then we have*

$$\inf_{\psi} \max_{i \in [M]} \mathbb{P}_i\{\psi \neq i\} \geq 1 - \frac{\frac{1}{M^2} \sum_{i,j=1}^{M} \mathsf{KL}(\mathbb{P}_i, \mathbb{P}_j) + \log 2}{\log M},$$

where the infimum is over all tests ψ that take values in $[M]$.

Proof. Fix a test ψ. Let $p_i := \mathbb{P}_i\{\psi = i\}$ and $q_i := \frac{1}{M}\sum_{j=1}^M \mathbb{P}_j\{\psi = i\}$. Moreover, let

$$\bar{p} = \frac{1}{M}\sum_{i=1}^M p_i = \frac{1}{M}\sum_{i=1}^M \mathbb{P}_i\{\psi = i\}, \qquad \bar{q} = \frac{1}{M}\sum_{i=1}^M q_i = \frac{1}{M}.$$

We claim that for any test ψ,

$$1 - \max_{i\in[M]} \mathbb{P}_i\{\psi \neq i\} \le 1 - \frac{1}{M}\sum_{i=1}^M \mathbb{P}_i\{\psi \neq i\} = \bar{p} \le \frac{\frac{1}{M^2}\sum_{i,j=1}^M \mathsf{KL}(\mathbb{P}_i, \mathbb{P}_j) + \log 2}{\log M}.$$

It suffices to prove the last inequality.

Using the inequality

$$-\bar{p}\log\bar{p} - (1-\bar{p})\log(1-\bar{p}) \le \log 2$$

(which says that the entropy of $\mathsf{Ber}(p)$ is maximized at $p = 1/2$), we obtain

$$\begin{aligned} h(\bar{p},\bar{q}) + \log 2 &\ge \bar{p}\log\frac{\bar{p}}{\bar{q}} + (1-\bar{p})\log\frac{1-\bar{p}}{1-\bar{q}} - \bar{p}\log\bar{p} - (1-\bar{p})\log(1-\bar{p}) \\ &= \bar{p}\log\frac{1}{\bar{q}} + (1-\bar{p})\log\frac{1}{1-\bar{q}} \ge \bar{p}\log\frac{1}{\bar{q}} = \bar{p}\log M \end{aligned}$$

and thus

$$\bar{p} \le \frac{h(\bar{p},\bar{q}) + \log 2}{\log M}.$$

Moreover, the convexity of h yields

$$h(\bar{p},\bar{q}) \le \frac{1}{M}\sum_{i=1}^M h(p_i, q_i) \le \frac{1}{M^2}\sum_{i,j=1}^M h(\mathbb{P}_i\{\psi = i\}, \mathbb{P}_j\{\psi = i\}).$$

Hence it remains to show that

$$h(\mathbb{P}_i\{\psi = i\}, \mathbb{P}_j\{\psi = i\}) \le \mathsf{KL}(\mathbb{P}_i, \mathbb{P}_j).$$

Let X_i denote the observation under $\mathbb{P}_i$ for $i \in [M]$. The above inequality is equivalent to

$$\mathsf{KL}(\mathbb{1}\{\psi(X_i) = i\}, \mathbb{1}\{\psi(X_j) = i\}) \le \mathsf{KL}(X_i, X_j),$$

which holds thanks to the data processing inequality. □

Combining Theorem 5.1 with Fano's inequality, we obtain the following corollary.

Corollary 5.12. *Suppose that for $\theta_1, \dots, \theta_M \in \Theta$, we have*

$$d(\theta_i, \theta_j)^2 \ge 4s_n, \qquad \mathsf{KL}(\mathbb{P}_{\theta_i}, \mathbb{P}_{\theta_j}) \le \frac{1}{2}\log M - \log 2$$

for any distinct $i, j \in [M]$. Then it holds that

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta[d(\theta, \hat{\theta})^2] \ge s_n/2.$$

See Theorem 2.5 of [Tsy08] for a more precisely stated version.

5.4.2 Application to Gaussian mean estimation

Let us again consider estimating $\mu \in \mathbb{R}^d$ given i.i.d. $X_1, \dots, X_n \sim \mathsf{N}(\mu, I_d)$. Recall that

$$\mathsf{KL}(\mathbb{P}_\mu, \mathbb{P}_{\mu'}) = \frac{n}{2}\|\mu - \mu'\|_2^2.$$

Therefore, we would like to choose $\mu_1, \dots, \mu_M$ such that

$$4s_n \le \|\mu_i - \mu_j\|_2^2 \le \frac{1}{n}(\log M - 2\log 2).$$

On the one hand, we need many μ_i so that M is large. On the other hand, if there are too many μ_i packed together, the separation s_n becomes too small and so does the lower bound. We need to find a balance between these two tensions.

Let us introduce the notions of ε-packing and ε-net.

Definition 5.13. *A set $N \subset B \subset \mathbb{R}^d$ is called an ε-packing of B in the Euclidean distance if $\|\mu - \mu'\|_2 \ge \varepsilon$ for any distinct $\mu, \mu' \in N$.*

A set $N \subset B \subset \mathbb{R}^d$ is called an ε-net of B in the Euclidean distance if for every $\mu \in B$, there exists $\mu' \in N$ such that $\|\mu - \mu'\|_2 \le \varepsilon$.

We will not prove the following result, but the intuition is clear by considering the ratio of volumes.

Lemma 5.14. *Let $\mathcal{B}^d$ denote the unit ball in $\mathbb{R}^d$. There exists an ε-packing N of $\mathcal{B}^d$, which is also an ε-net of $\mathcal{B}^d$, such that*

$$(1/\varepsilon)^d \le |N| \le (3/\varepsilon)^d.$$

Note that in a homework problem, we assume that there is a $1/4$-net N of the unit sphere $\mathcal{S}^{d-1}$ in $\mathbb{R}^d$ such that $|N| \le 12^d$. This is simply replacing $\mathcal{B}^d$ with the subset $\mathcal{S}^{d-1}$ and setting $\varepsilon = 1/4$ in the above lemma.

With the lemma given, let us take a $1/4$-packing $N = \{\theta_1, \dots, \theta_M\} \subset \mathcal{B}^d$ where $M \ge 4^d$. Set $\mu_i = c\sqrt{\frac{d}{n}}\,\theta_i$ for each $i \in [M]$ and some constant $c > 0$ to be determined. Then by definition, we can set $s_n = \frac{c^2 d}{4^3 n}$ so that

$$\|\mu_i - \mu_j\|_2^2 = c^2\frac{d}{n}\|\theta_i - \theta_j\|_2^2 \ge \frac{c^2 d}{4^2 n} = 4s_n.$$

On the other hand,

$$\|\mu_i - \mu_j\|_2^2 = c^2\frac{d}{n}\|\theta_i - \theta_j\|_2^2 \le \frac{4c^2 d}{n} \le \frac{1}{n}(\log M - 2\log 2)$$

if we choose $c > 0$ to be a sufficiently small constant. We conclude that

$$\inf_{\hat\mu}\sup_{\mu\in\mathbb{R}^d}\mathbb{E}_\mu\left[\|\hat\mu - \mu\|_2^2\right] \ge \frac{s_n}{2} \gtrsim \frac{d}{n}.$$

5.4.3 Application to nonparametric regression

Lemma 5.15 (Hoeffding's lemma)**.** *Suppose that a random variable X has mean zero and satisfies $a \le X \le b$ for constants $a, b \in \mathbb{R}$. Then, for any $\lambda \in \mathbb{R}$, we have*

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

Proof. By convexity, it holds that

$$e^{\lambda X} \le \frac{b-X}{b-a} e^{\lambda a} + \frac{X-a}{b-a} e^{\lambda b},$$

so

$$\mathbb{E}[e^{\lambda X}] \le \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b} = e^{L(\lambda(b-a))}, \qquad L(t) := \frac{at}{b-a} + \log\Big(1 + \frac{a(1-e^t)}{b-a}\Big).$$

We have

$$L'(t) = \frac{ab(e^t-1)}{(a-b)(b-ae^t)}, \qquad L''(t) = \frac{-abe^t}{(b-ae^t)^2} \le \frac{1}{4}.$$

Using the second-order Taylor approximation, we obtain that $L(t) \le t^2/8$ for $t \in \mathbb{R}$. □

Lemma 5.16 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables taking values in $[0,1]$. Then for all $t > 0$,*

$$\mathbb{P}\Big\{ \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \le -t \Big\} \le \exp(-2t^2/n).$$

Proof. Use Hoeffding's lemma $\mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}] \le \exp(\frac{\lambda^2}{8})$ and Chernoff's bound. □

Lemma 5.17 (Varshamov–Gilbert bound)**.** *Let $d \ge 8$. There exists $\{\omega_1, \dots, \omega_M\} \subset \{0,1\}^d$ such that $\rho(\omega_i, \omega_j) \ge d/8$ for any distinct $i, j \in [M]$ and $M \ge e^{d/8}$, where $\rho(\cdot,\cdot)$ denotes the Hamming distance.*

Proof. Let $\omega_{i,k}$ be i.i.d. $\mathsf{Ber}(1/2)$ for $i \in [M]$ and $k \in [d]$. Consider the event

$$\mathcal{E} := \{\rho(\omega_i, \omega_j) \ge d/8 \text{ for any distinct } i, j \in [M]\}.$$

It suffices to show that $\mathbb{P}\{\mathcal{E}\} > 0$, i.e., $\mathbb{P}\{\mathcal{E}^c\} < 1$. (This is called the probabilistic method.)

For any distinct $i, j \in [M]$, $\rho(\omega_i, \omega_j) = \sum_{k=1}^d \mathbb{1}\{\omega_{i,k} \ne \omega_{j,k}\}$, so it is a sum of $\mathsf{Ber}(1/2)$ random variables. By a union bound and Hoeffding's inequality with $t = 3d/8$, we obtain

$$\mathbb{P}\{\mathcal{E}^c\} \le \sum_{i,j \in [M],\, i \ne j} \mathbb{P}\{\rho(\omega_i, \omega_j) < d/8\} \le M^2 \exp\Big(\frac{-9d}{32}\Big).$$

It is not hard to see that this is strictly less than 1, as taking the logarithm gives $\frac{d}{4} - \frac{9d}{32} < 0$. □

Theorem 5.18. *Consider the nonparametric regression model (4.8), where $f \in \Sigma(\beta, L)$ for $\beta, L > 0$, $x_i = i/n$, and ε_i are i.i.d $\mathsf{N}(0, \sigma^2)$ noise for $i \in [n]$. For a constant $c = c_3(\beta) L^{\frac{2}{2\beta+1}} \sigma^{\frac{4\beta}{2\beta+1}} > 0$, the following minimax lower bound in the squared L^2 distance holds over the Hölder class:*

$$\inf_{\hat f} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}\, \|\hat f - f\|_{L^2}^2 \ge c\, n^{\frac{-2\beta}{2\beta+1}}.$$

Proof. This time, for the squared L^2 distance $d(f,g)^2 = \|f-g\|_{L^2}^2 = \int_0^1 (f(x) - g(x))^2\, dx$, the proof is based on multiple hypothesis testing over $\{f_1, \dots, f_M\} \subset \Sigma(\beta, L)$.

Construction of hypotheses Fix a constant $C_0 = C_0(\beta, L, \sigma) > 0$ to be determined later. Let

$$d = \lceil C_0 n^{\frac{1}{2\beta+1}} \rceil, \quad h = \frac{1}{d}, \quad z_k = \frac{k-1/2}{d}, \quad \phi_k(x) = Lh^\beta K\Big(\frac{x - z_k}{h}\Big), \quad k \in [d],\, x \in [0,1],$$

where K is defined by (5.3). Recall that using the proof of Theorem 5.5, we can show that $\phi_k \in \Sigma(\beta, L/2)$ if the constant $c_1(\beta) > 0$ in (5.3) is taken to be sufficiently small. Moreover, ϕ_k is supported in $[z_k - \frac{h}{2}, z_k + \frac{h}{2}] = [\frac{k-1}{d}, \frac{k}{d}]$ for each $k \in [d]$.

Let $\omega_1, \dots, \omega_M$ be given by Lemma 5.17. For each $i \in [M]$, we define $f_i(x) := \sum_{k=1}^d \omega_{i,k}\phi_k(x)$. Since the supports of ϕ_k are disjoint (up to a set of measure zero), it is easily seen that $f_i \in \Sigma(\beta, L)$.

Separation For distinct $i, j \in [M]$, we have

$$\begin{aligned}\|f_i - f_j\|_{L^2}^2 &= \int_0^1 \big(f_i(x) - f_j(x)\big)^2\,dx = \int_0^1 \Big(\sum_{k=1}^d (\omega_{i,k} - \omega_{j,k})\phi_k(x)\Big)^2\,dx \\ &= \sum_{k=1}^d (\omega_{i,k} - \omega_{j,k})^2 \int_0^1 \phi_k(x)^2\,dx = L^2h^{2\beta+1}\|K\|_{L^2}^2 \rho(\omega_i, \omega_j).\end{aligned}$$

As a result of Lemma 5.17, for $c_2 = c_2(\beta) = \|K\|_{L^2}^2 > 0$,

$$\|f_i - f_j\|_{L^2}^2 \ge L^2h^{2\beta+1}c_2\, d/8 \gtrsim L^2c_2C_0^{-2\beta}n^{\frac{-2\beta}{2\beta+1}}.$$

KL divergence Finally, we check

$$\begin{aligned}\mathsf{KL}(\mathbb{P}_{f_i}, \mathbb{P}_{f_j}) &= \sum_{\ell=1}^n \frac{(f_i(x_\ell) - f_j(x_\ell))^2}{2\sigma^2} = \frac{1}{2\sigma^2}\sum_{\ell=1}^n\sum_{k=1}^d (\omega_{i,k} - \omega_{j,k})^2\phi_k(x_\ell)^2 \\ &\le \frac{L^2h^{2\beta}c_1^2}{2\sigma^2e^2}\sum_{\ell=1}^n\sum_{k=1}^d \mathbb{1}\{|x_\ell - z_k| \le h\} \lesssim \frac{L^2h^{2\beta}c_1^2}{\sigma^2}n \lesssim \frac{L^2c_1^2}{\sigma^2}C_0^{-2\beta}n^{\frac{1}{2\beta+1}}.\end{aligned}$$

To apply Corollary 5.12, we need this bound to be smaller than $\frac{1}{2}\log M - \log 2 \gtrsim d \ge C_0 n^{\frac{1}{2\beta+1}}$, i.e. $C_0 \gtrsim (\frac{L^2c_1^2}{\sigma^2})^{\frac{1}{2\beta+1}}$. Plugging this into the separation bound above finishes the proof. □

We have established matching upper and lower bounds for the minimax risk at a point or in the L^2 norm for nonparametric regression.

5.5 Generalization of the two-point method

One way to generalize the two-point method is through composite hypothesis testing. For a finite set Θ, consider the mixture

$$\bar{\mathbb{P}} := \frac{1}{|\Theta|}\sum_{\theta\in\Theta} \mathbb{P}_\theta, \tag{5.4}$$

where $\mathbb{P}_\theta$ is each a distribution. In other words, $Y \sim \bar{\mathbb{P}}$ can be generated as follows: First sample θ uniformly randomly from Θ and then, conditional on θ, sample $Y \sim \mathbb{P}_\theta$. We will study hypothesis

testing between a distribution $\mathbb{P}_0$ and the mixture $\bar{\mathbb{P}}$, where the latter is usually referred to as a composite hypothesis.

Let ψ denote a test, which equals 0 if it selects $\mathbb{P}_0$ and equals 1 if it selects $\bar{\mathbb{P}}$. By the Neyman–Pearson lemma and a homework problem, we have that for any test ψ,

$$\mathbb{P}_0\{\psi = 1\} + \bar{\mathbb{P}}\{\psi = 0\} \geq 1 - \mathsf{TV}(\mathbb{P}_0, \bar{\mathbb{P}}) \geq 1 - \sqrt{\chi^2(\bar{\mathbb{P}}, \mathbb{P}_0)}, \tag{5.5}$$

where $\mathsf{TV}(\cdot, \cdot)$ and $\chi^2(\cdot, \cdot)$ denote the total variation distance and the χ^2-divergence respectively.

To showcase how this inequality leads to a minimax lower bound for an estimation problem, we consider the following example. Suppose that we aim to estimate θ given the observation

$$Y = \theta + \varepsilon, \tag{5.6}$$

where θ is k-sparse and $\varepsilon \sim \mathsf{N}(0, \sigma^2 I_n)$. Recall that this is called the sparse sequence model in a homework problem, and it is a special case of sparse linear regression with $n = d$ and the design matrix X being orthogonal. Assume that $1 \leq k \leq \sqrt{n}$. We aim to prove a lower bound of order k/n up to a logarithmic factor, which then matches the upper bound.

Theorem 5.19. *Let $\mathbb{P}_\theta$ denote the probability associated with the model (5.6). For $\lambda \in (0, 1)$, set*

$$\mu := \frac{\sigma}{2}\sqrt{k \log\Big(1 + \frac{\lambda n}{k^2}\Big)}.$$

Define

$$\Theta := \Big\{\theta = \frac{\mu}{\sqrt{k}}\mathbf{1}_S : S \subset [n],\, |S| = k\Big\}.$$

In other words, each vector in Θ is k-sparse with support S, and its nonzero entries are all equal to $\mu/\sqrt{k}$. Let $\bar{\mathbb{P}}$ be defined as in (5.4). Then

$$\chi^2(\bar{\mathbb{P}}, \mathbb{P}_0) \leq 2\lambda.$$

Proof. Let p_0, p_θ, and $\bar{p}$ denote the densities of $\mathbb{P}_0$, $\mathbb{P}_\theta$, and $\bar{\mathbb{P}}$ respectively. Let θ and θ' be two independent uniform random variables over Θ. By the definition of the χ^2-divergence, we have

$$\chi^2(\bar{\mathbb{P}}, \mathbb{P}_0) = \int \frac{(\bar{p} - p_0)^2}{p_0} = \int \Big(\frac{\bar{p}}{p_0}\Big)^2 p_0 - 1 = \mathbb{E}_{\theta,\theta'} \int \frac{p_\theta}{p_0}\frac{p_{\theta'}}{p_0} p_0 - 1.$$

Let S and S' denote the supports of θ and θ' respectively; they are independent random subsets of $[n]$, each of size k. Since the noise is Gaussian, it holds that

$$\frac{p_\theta(x)}{p_0(x)} = \exp\bigg(-\frac{1}{2\sigma^2}\Big\|x - \frac{\mu}{\sqrt{k}}\mathbf{1}_S\Big\|_2^2 + \frac{1}{2\sigma^2}\|x\|_2^2\bigg) = \exp\bigg(\frac{1}{2\sigma^2}\Big(\frac{2\mu}{\sqrt{k}}x^\top \mathbf{1}_S - \mu^2\Big)\bigg).$$

For $Z \sim \mathsf{N}(0, I_n)$, define $Z_S := \sum_{i \in S} Z_i$. Let $\nu := \frac{\mu}{\sigma\sqrt{k}}$. Then we have

$$\begin{aligned}\int \frac{p_\theta}{p_0}\frac{p_{\theta'}}{p_0} p_0 &= \int \exp\bigg(\frac{1}{2\sigma^2}\Big(\frac{2\mu}{\sqrt{k}}(x^\top \mathbf{1}_S + x^\top \mathbf{1}_{S'}) - 2\mu^2\Big)\bigg) p_0(x)\, dx \\ &= \mathbb{E}\bigg[\exp\bigg(\frac{1}{2\sigma^2}\Big(\frac{2\mu\sigma}{\sqrt{k}}(Z_S + Z_{S'}) - 2\mu^2\Big)\bigg)\bigg] = \mathbb{E}\big[\exp\big(\nu(Z_S + Z_{S'}) - \nu^2 k\big)\big].\end{aligned}$$

Note that $Z_S + Z_{S'} = 2\sum_{i\in S\cap S'} Z_i + \sum_{i\in S\triangle S'} Z_i$ and $|S\triangle S'| \le 2k$. Therefore,

$$\begin{aligned}\int \frac{p_\theta}{p_0}\frac{p_{\theta'}}{p_0}p_0 &= \mathbb{E}\left[\exp\left(2\nu\sum_{i\in S\cap S'} Z_i\right)\right]\cdot\mathbb{E}\left[\exp\left(\nu\sum_{i\in S\triangle S'} Z_i\right)\right]\cdot\exp(-\nu^2 k)\\ &= \exp\left(2\nu^2|S\cap S'|\right)\cdot\exp\left(\nu^2|S\triangle S'|/2\right)\cdot\exp(-\nu^2 k) \le \exp(2\nu^2|S\cap S'|).\end{aligned}$$

It follows that

$$\begin{aligned}\chi^2(\bar{\mathbb{P}},\mathbb{P}_0) &\le \mathbb{E}_{S,S'}\left[\exp(2\nu^2|S\cap S'|)\right]-1\\ &\le \mathbb{E}_{S'}\left[\mathbb{E}_S\left[\exp(2\nu^2|S\cap S'|)\,\middle|\, S'\right]\right]-1 = \mathbb{E}_S\left[\exp(2\nu^2|S\cap[k]|)\right]-1.\end{aligned}$$

The random variable $|S\cap[k]|$ is a sampling-without-replacement version of $\mathsf{Bin}(k,k/n)$, and it can shown that the MGF of the former is dominated by the MGF of the latter. As a result,

$$\chi^2(\bar{\mathbb{P}},\mathbb{P}_0) \le \left(e^{2\nu^2}\frac{k}{n}+\left(1-\frac{k}{n}\right)\right)^k - 1 = \left(1+\frac{k}{n}(e^{2\nu^2}-1)\right)^k - 1.$$

Recall that

$$\nu = \frac{\mu}{\sigma\sqrt{k}} = \frac{1}{2}\sqrt{\log\left(1+\frac{\lambda n}{k^2}\right)}.$$

Hence, we conclude that

$$\chi^2(\bar{\mathbb{P}},\mathbb{P}_0) \le \left(1+\frac{k}{n}\frac{\lambda n}{k^2}\right)^k - 1 = \left(1+\frac{\lambda}{k}\right)^k - 1 \le 2\lambda$$

for any $\lambda\in(0,1)$ and $k\ge 1$. □

Corollary 5.20. *We have the following minimax lower bound for the model* (5.6)*:*

$$\inf_{\hat\theta}\sup_{\theta\in\mathbb{R}^n}\frac{1}{n}\mathbb{E}\,\|\hat\theta-\theta\|_2^2 \gtrsim \sigma^2\frac{k}{n}\log\left(1+\frac{n}{k^2}\right).$$

Proof. We continue to use the notation from above. Let $\lambda = 1/8$ and $\mu = \frac{\sigma}{2}\sqrt{k\log(1+\frac{n}{8k^2})}$. Let $\hat\theta$ be any estimator of θ. Define a test ψ by $\psi = 0$ if $\|\hat\theta\|_2 \le \mu/2$ and $\psi = 1$ if $\|\hat\theta\|_2 > \mu/2$. From (5.5) and the above theorem, we obtain

$$\max\Big\{\mathbb{P}_0\{\psi=1\}, \max_{\theta\in\Theta}\mathbb{P}_\theta\{\psi=0\}\Big\} \ge \frac{1}{2}\Big(\mathbb{P}_0\{\psi=1\}+\bar{\mathbb{P}}\{\psi=0\}\Big) \ge \frac{1-\sqrt{1/4}}{2} = \frac{1}{4}.$$

Suppose that $\frac{1}{n}\|\hat\theta-\theta\|_2^2 \le \frac{\mu^2}{9n}$ with probability at least 0.9. Then

$$\Big|\|\hat\theta\|_2 - \|\theta\|_2\Big| \le \|\hat\theta-\theta\|_2 \le \frac{\mu}{3}.$$

By the definition of ψ, if $\theta = 0$, then $\psi = 0$; if $\theta\in\Theta$, then $\|\theta\|_2 = \mu$ and thus $\psi = 1$. We reach a contradiction. Therefore, $\frac{1}{n}\|\hat\theta-\theta\|_2^2 > \frac{\mu^2}{9n}$ with probability at least 0.1, proving the conclusion. □

Bibliography

[Ber18] Quentin Berthet. *Principles of Statistics, Lecture Notes*. 2018.

[GL95] Richard D. Gill and Boris Y. Levit. Applications of the van trees inequality: a bayesian cramér-rao bound. *Bernoulli*, 1(1-2):59–79, 03 1995.

[Kee11] Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer, 2011.

[LC06] Erich L. Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[LR06] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[Mon15] Andrea Montanari. Computational implications of reducing data to sufficient statistics. *Electron. J. Statist.*, 9(2):2370–2390, 2015.

[RH19] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. 2019.

[Tsy08] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[vdV00] Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.