Introduction to Nonparametric Statistics

Bodhisattva Sen

March 24, 2020

Contents

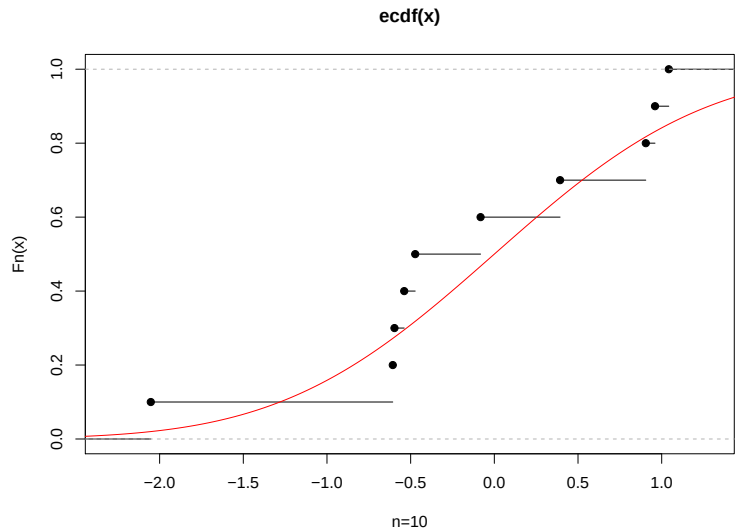
Abstract

This lecture note arose from a class I taught in Spring 2016 to our 2nd year PhD students (in Statistics) at Columbia University. The choice of topics is very eclectic and mostly reflect: (a) my background and research interests, and (b) some of the topics I wanted to learn more systematically in 2016. The first part of this lecture notes is on nonparametric function estimation — density and regression — and I borrow heavily from the book Tsybakov [14] and the course he taught at Yale in 2014. The second part of the course is a medley of different topics: (i) reproducing kernel Hilbert spaces (RKHSs; Section 5), (ii) bootstrap methods (Section 6), (iii) multiple hypothesis testing (Section 7), and (iv) an introduction to high dimensional linear regression (Section 8).

The content of Section 5 is greatly influenced by Arthur Gretton's lectures and slides on RKHSs and its applications in Machine Learning (see e.g., http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/rkhscourse.html for a more detailed course). I have borrowed the material in Section 7 from Emmanuel Candes's lectures on 'Theory of Statistics' (Stats 300C, Stanford), while the content of Section 8 is taken from Hastie et al. [5].

1 Kernel density estimation

Let $X_1, \ldots, X_n$ be i.i.d. random variables having a probability density p with respect to the Lebesgue measure on $\mathbb{R}$. The corresponding distribution function is $F(x) := \int_{-\infty}^{x} p(t)dt$.



A natural estimator of F is the *empirical distribution function*:

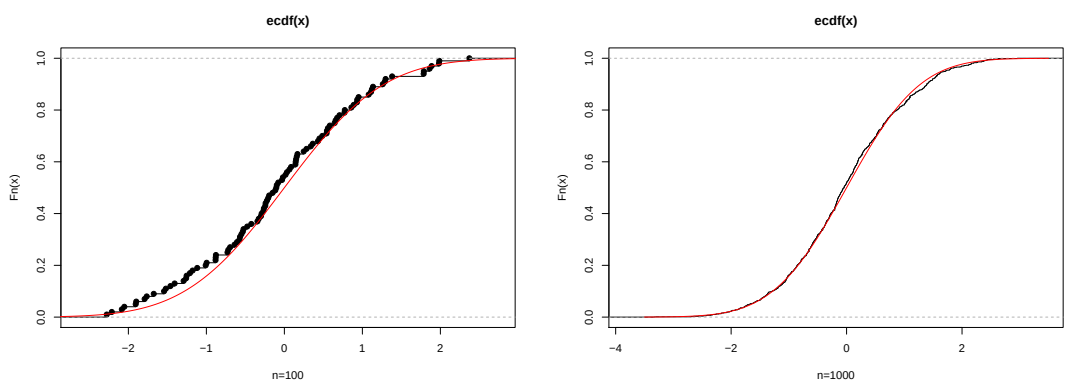
$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i), \tag{1}$$

where $I(\cdot)$ denotes the indicator function. The Glivenko-Cantelli theorem shows that

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \overset{a.s.}{\to} 0,$$

as $n \to \infty$ (Exercise (HW1)). Further we know that for every $x \in \mathbb{R}$,

$$\sqrt{n}(F_n(x) - F(x)) \overset{d}{\to} N(0, F(x)(1 - F(x))).$$



Exercise (HW1): Consider testing $F = F_0$ where F_0 is a known continuous strictly increasing distribution function (e.g., standard normal) when we observe i.i.d. data $X_1, \ldots, X_n$ from F. The Kolmogorov-Smirnov test statistic is to consider

$$D_n := \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|,$$

and reject H_0 when $D_n > c_\alpha$, for a suitable $c_\alpha > 0$ (where α is the level of the test). Show that, under H_0, D_n is distribution-free, i.e., the distribution of D_n does not depend on F_0 (as long as it is continuous and strictly increasing). How would you compute (approximate/simulate) the critical value c_α, for every n.

Let us come back to the estimation of p. As p is the derivative of F, for small $h > 0$, we can write the approximation

$$p(x) \approx \frac{F(x+h) - F(x-h)}{2h}.$$

As F_n is a natural estimator of F, it is intuitive to define the following (Rosenblatt) estimator of p:

$$\hat{p}_n^R(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}.$$

We can rewrite $\hat{p}_n^R$ as

$$\hat{p}_n^R(x) = \frac{1}{2nh} \sum_{i=1}^n I(x - h < X_i \leq x + h) = \frac{1}{nh} \sum_{i=1}^n K_0\left(\frac{X_i - x}{h}\right),$$

where $K_0(u) = \frac{1}{2} I_{(-1,1]}(u)$. A simple generalization of the Rosenblatt estimator is given by

$$\hat{p}_n(x) := \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right), \tag{2}$$

where $K : \mathbb{R} \to \mathbb{R}$ is an integrable function satisfying $\int K(u)du = 1$. Such a function K is called a *kernel* and the parameter h is called the *bandwidth* of the estimator (2). The function $\hat{p}_n$ is called the *kernel density estimator* (KDE) or the Parzen-Rosenblatt estimator. Some classical examples of kernels are the following:

$$\begin{aligned}
K(u) &= \tfrac{1}{2} I(|u| \leq 1) && \text{(the rectangular kernel)} \\
K(u) &= \frac{1}{\sqrt{2\pi}} \exp(-u^2/2) && \text{(the Gaussian kernel)} \\
K(u) &= \tfrac{3}{4}(1 - u^2) I(|u| \leq 1) && \text{(the Epanechnikov kernel).}
\end{aligned}$$

Note that if the kernel K takes only nonnegative values and if $X_1, \ldots, X_n$ are fixed, then $\hat{p}_n$ is a probability density.

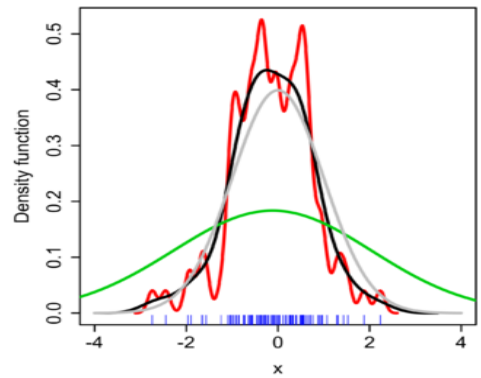


Figure 1: KDE with different bandwidths of a random sample of 100 points from a standard normal distribution. Grey: true density (standard normal). Red: KDE with h=0.05. Black: KDE with h=0.337. Green: KDE with h=2.

The Parzen-Rosenblatt estimator can be generalized to the multidimensional case easily. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. with (joint) density $p(\cdot, \cdot)$. A kernel estimator of p is then given by

$$\hat{p}_n(x, y) := \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right), \tag{3}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a kernel defined as above and $h > 0$ is the bandwidth.

1.1 The choice of the bandwidth and the kernel

It turns out that the choice of the bandwidth h is far more crucial for the quality of $\hat{p}_n$ as an estimator of p than the choice of the kernel K. We can view the KDE (for unimodal, nonnegative kernels) as the sum of n small "mountains" given by the functions

$$x \mapsto \frac{1}{nh} K\left(\frac{X_i - x}{h}\right).$$

Every small mountain is centered around an observation X_i and has area $1/n$ under it, for any bandwidth h. For a small bandwidth the mountain is very concentrated (peaked), while for a large bandwidth the mountain is low and fat. If the bandwidth is small, then the mountains remain separated and their sum is peaky. On the other hand, if the bandwidth is large, then the sum of the individual mountains is too flat. Intermediate values of the bandwidth should give the best results.

For a fixed h, the KDE $\hat{p}_n(x_0)$ is not consistent in estimating $p(x_0)$, where $x_0 \in \mathbb{R}$. However, if the bandwidth decreases with sample size at an appropriate rate, then it is, regardless of which kernel is used.

Exercise (HW1): Suppose that p is continuous at x_0, that $h_n \to 0$, and that $nh_n \to \infty$ as $n \to \infty$. Then, $\hat{p}_n(x_0) \xrightarrow{p} p(x_0)$ [Hint: Study the bias and variance of the estimator separately].

1.2 Mean squared error of kernel estimators

A basic measure of the accuracy of $\hat{p}_n$ is its *mean squared risk* (or *mean squared error*) at an arbitrary fixed point $x_0 \in \mathbb{R}$:

$$\mathrm{MSE} = \mathrm{MSE}(x_0) := \mathbb{E}_p\Big[(\hat{p}_n(x_0) - p(x_0))^2\Big].$$

Here $\mathbb{E}_p$ denotes the expectation with respect to the distribution of $(X_1, \ldots, X_n)$:

$$\mathrm{MSE}(x_0) := \int \cdots \int \Big(\hat{p}_n(x_0; z_1, \ldots, z_n) - p(x_0)\Big)^2 \Big[\prod_{i=1}^n p(z_i)\Big] dz_1 \ldots dz_n.$$

Of course,

$$\mathrm{MSE}(x_0) = b^2(x_0) + \sigma^2(x_0)$$

where

$$b(x_0) := \mathbb{E}_p[\hat{p}_n(x_0)] - p(x_0), \qquad \text{(bias)}$$

and

$$\sigma^2(x_0) := \mathbb{E}_p\Big[\Big(\hat{p}_n(x_0) - \mathbb{E}_p[\hat{p}_n(x_0)]\Big)^2\Big] \qquad \text{(variance)}.$$

To evaluate the mean squared risk of $\hat{p}_n$ we will analyze separately its variance and bias.

Proposition 1.1 (Variance of $\hat{p}_n$)**.** Suppose that the density p satisfies $p(x) \le p_{\max} < \infty$ for all $x \in \mathbb{R}$. Let $K : \mathbb{R} \to \mathbb{R}$ be the kernel function such that

$$\int K^2(u)du < \infty.$$

Then for any $x_0 \in \mathbb{R}$, $h > 0$, and $n \ge 1$ we have

$$\sigma^2(x_0) \le \frac{C_1}{nh},$$

where $C_1 = p_{\max} \int K^2(u)du$.

Proof. Observe that $\hat{p}_n(x_0)$ is an average of n i.i.d. random variables and so

$$\sigma^2(x_0) = \mathrm{Var}(\hat{p}_n(x_0)) = \frac{1}{n}\mathrm{Var}\left(\frac{1}{h}K\left(\frac{X_1 - x_0}{h}\right)\right) \le \frac{1}{nh^2}\mathbb{E}_p\left[K^2\left(\frac{X_1 - x_0}{h}\right)\right]$$

Now, observe that

$$\mathbb{E}_p\left[K^2\left(\frac{X_1 - x_0}{h}\right)\right] \;=\; \int K^2\left(\frac{z - x_0}{h}\right) p(z)dz \le p_{\max}\, h \int K^2(u)du.$$

Combining the above two displays we get the desired result.

□

Thus, we conclude that if the bandwidth $h \equiv h_n$ is such that $nh \to \infty$ as $n \to \infty$, then the variance of $\sigma^2(x_0)$ goes to 0 as $n \to \infty$.

To analyze the bias of the KDE (as a function of h) we need certain conditions on the density p and on the kernel K.

Definition 1.2. Let T be an interval in $\mathbb{R}$ and let β and L be two positive numbers. The *Hölder class* $\Sigma(\beta, L)$ on T is defined as the set of $\ell = \lfloor \beta \rfloor$ times differentiable functions $f : T \to \mathbb{R}$ whose derivative $f^{(\ell)}$ satisfies

$$|f^{(\ell)}(x) - f^{(\ell)}(x')| \le L|x - x'|^{\beta - \ell}, \qquad \text{for all } x, x' \in T.$$

Definition 1.3. Let $\ell \ge 1$ be an integer. We say that $K : \mathbb{R} \to \mathbb{R}$ is a *kernel of order* ℓ if the functions $u \mapsto u^j K(u),\ j = 0, 1, \ldots, \ell$, are integrable and satisfy

$$\int K(u)du = 1, \qquad \int u^j K(u)du = 0, \quad j = 1, \ldots, \ell.$$

Does bounded kernels of order ℓ exist? See Section 1.2.2 of [14] for constructing such kernels.

Observe that when $\ell \ge 2$ then the kernel has to take negative values which may lead to negative values of $\hat{p}_n$. This is sometimes mentioned as a drawback of using higher order kernels ($\ell \ge 2$). However, observe that we can always define the estimator

$$\hat{p}_n^+(x) = \max\{0, \hat{p}_n(x)\}$$

whose risk is smaller than or equal to the risk of $\hat{p}_n(x)$:

$$\mathbb{E}_p\Big[(\hat{p}_n^+(x_0) - p(x_0))^2\Big] \le \mathbb{E}_p\Big[(\hat{p}_n(x_0) - p(x_0))^2\Big], \qquad \forall x \in \mathbb{R}.$$

Suppose now that p belong to a class of densities $\mathcal{P} = \mathcal{P}(\beta, L)$ defined as follows:

$$\mathcal{P}(\beta, L) := \left\{p : p \ge 0, \int p(x)dx = 1, \text{and } p \in \Sigma(\beta, L) \text{ on } \mathbb{R}\right\}.$$

Proposition 1.4 (Bias of $\hat{p}_n$)**.** Assume that $p \in \mathcal{P}(\beta, L)$ and let K be a kernel of order $\ell = \lfloor \beta \rfloor$ satisfying

$$\int |u|^\beta |K(u)| du < \infty.$$

Then for any $x_0 \in \mathbb{R}$, $h > 0$, and $n \geq 1$ we have

$$|b(x_0)| \leq C_2 h^\beta, \tag{4}$$

where $C_2 = \frac{L}{\ell!} \int |u|^\beta |K(u)| du$.

Proof. We have

$$\begin{aligned} b(x_0) &= \frac{1}{h} \int K\left(\frac{z-x}{h}\right) p(z) dz - p(x_0) \\ &= \int K(u) \Big[p(x_0 + uh) - p(x_0) \Big] du. \end{aligned}$$

Next, using Taylor theorem[1] , we get

$$p(x_0 + uh) = p(x_0) + p'(x_0) uh + \ldots + \frac{(uh)^\ell}{\ell!} p^{(\ell)}(x_0 + \tau uh),$$

where $0 \leq \tau \leq 1$. Since K has order $\ell = \lfloor \beta \rfloor$, we obtain

$$\begin{aligned} b(x_0) &= \int K(u) \frac{(uh)^\ell}{\ell!} p^{(\ell)}(x_0 + \tau uh) du \\ &= \int K(u) \frac{(uh)^\ell}{\ell!} (p^{(\ell)}(x_0 + \tau uh) - p^{(\ell)}(x_0)) du, \end{aligned}$$

[1]**Taylor's theorem**: Let $k \geq 1$ be an integer and let the function $f : \mathbb{R} \to \mathbb{R}$ be k times differentiable at the point $a \in \mathbb{R}$. Then there exists a function $R_k : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + R_k(x),$$

where $R_k(x) = o(|x-a|^k)$ as $x \to a$.

Mean-value forms of the remainder: Let $f : \mathbb{R} \to \mathbb{R}$ be $k+1$ times differentiable on the open interval with $f^{(k)}$ continuous on the closed interval between a and x. Then

$$R_k(x) = \frac{f^{(k+1)}(\xi_L)}{(k+1)!}(x-a)^{k+1}$$

for some real number ξ_L between a and x. This is the *Lagrange form* of the remainder.

Integral form of the remainder: Let $f^{(k)}$ be absolutely continuous on the closed interval between a and x. Then

$$R_k(x) = \int_a^x \frac{f^{(k+1)}(t)}{k!}(x-t)^k \, dt. \tag{5}$$

Due to absolute continuity of $f^{(k)}$, on the closed interval between a and x, $f^{(k+1)}$ exists a.e.

and

$$
\begin{aligned}
|b(x_0)| &\leq \int |K(u)| \frac{|uh|^\ell}{\ell!} \Big| p^{(\ell)}(x_0 + \tau uh) - p^{(\ell)}(x_0) \Big| du \\
&\leq L \int |K(u)| \frac{|uh|^\ell}{\ell!} |\tau uh|^{\beta-\ell} du \leq C_2 h^\beta.
\end{aligned}
$$

□

From Propositions 1.1 and 1.4, we see that the upper bounds on the bias and variance behave in opposite ways as the bandwidth h varies. The variance decreases as h grows, whereas the bound on the bias increases. The choice of a small h corresponding to a large variance leads to *undersmoothing*. Alternatively, with a large h the bias cannot be reasonably controlled, which leads to *oversmoothing*. An optimal value of h that balances bias and variance is located between these two extremes. To get an insight into the optimal choice of h, we can minimize in h the upper bound on the MSE obtained from the above results.

If p and K satisfy the assumptions of Propositions 1.1 and 1.4, we obtain

$$
\text{MSE} \leq C_2^2 h^{2\beta} + \frac{C_1}{nh}. \tag{6}
$$

The minimum with respect to h of the right hand side of the above display is attained at

$$
h_n^* = \left(\frac{C_1}{2\beta C_2^2} \right)^{1/(2\beta+1)} n^{-1/(2\beta+1)}.
$$

Therefore, the choice $h = h_n^*$ gives

$$
\text{MSE}(x_0) = O\Big(n^{-\frac{2\beta}{2\beta+1}} \Big), \quad \text{as } n \to \infty,
$$

uniformly in x_0. Thus, we have the following result.

Theorem 1.5. Assume that the conditions of Proposition 1.4 hold and that $\int K^2(u)du < \infty$. Fix $\alpha > 0$ and take $h = \alpha n^{-1/(2\beta+1)}$. Then for $n \geq 1$, the KDE $\hat{p}_n$ satisfies

$$
\sup_{x_0 \in \mathbb{R}} \sup_{p \in \mathcal{P}(\beta, L)} \mathbb{E}_p \Big[(\hat{p}_n(x_0) - p(x_0))^2 \Big] \leq C n^{-\frac{2\beta}{2\beta+1}},
$$

where $C > 0$ is a constant depending only on β, L, α and on the kernel K.

Proof. We apply (14) to derive the result. To justify the application of Proposition 1.1, it remains to prove that there exists a constant $p_{\max} < \infty$ satisfying

$$
\sup_{x \in \mathbb{R}} \sup_{p \in \mathcal{P}(\beta, L)} p(x) \leq p_{\max}. \tag{7}
$$

To show that (7) holds, consider K^* which is a bounded kernel of order ℓ (not necessarily equal to K). Applying Proposition 1.4 with $h = 1$ we get that, for any $x \in \mathbb{R}$ and any $p \in \mathcal{P}(\beta, L)$,

$$\Big| \int K(z-x)\, p(z) dz - p(x) \Big| \le C_2^* := \frac{L}{\ell!} \int |u|^\beta |K^*(u)| du.$$

Therefore, for any $x \in \mathbb{R}$ and any $p \in \mathcal{P}(\beta, L)$,

$$p(x) \le C_2^* + \int |K^*(z-x)|\, p(z) dz \le C_2^* + K^*_{\max},$$

where $K^*_{\max} = \sup_{u \in \mathbb{R}} |K^*(u)|$. Thus, we get (7) with $p_{\max} = C_2^* + K^*_{\max}$. □

Under the assumptions of Theorem 1.5, the rate of convergence of the estimator $\hat{p}_n(x_0)$ is $\psi_n = n^{-\frac{\beta}{2\beta+1}}$, which means that for a finite constant C and for all $n \ge 1$ we have

$$\sup_{p \in \mathcal{P}(\beta, L)} \mathbb{E}_p \Big[(\hat{p}_n(x_0) - p(x_0))^2 \Big] \le C \psi_n^2.$$

Now the following two questions arise. Can we improve the rate ψ_n by using other density estimators? What is the best possible rate of convergence? To answer these questions it is useful to consider the minimax risk R_n^* associated to the class $\mathcal{P}(\beta, L)$:

$$R_n^*(\mathcal{P}(\beta, L)) = \inf_{T_n} \sup_{p \in \mathcal{P}(\beta, L)} \mathbb{E}_p \Big[(T_n(x_0) - p(x_0))^2 \Big],$$

where the infimum is over all estimators. One can prove a lower bound on the minimax risk of the form $R_n^*(\mathcal{P}(\beta, L)) \ge C' \psi_n^2 = C' n^{-\frac{2\beta}{2\beta+1}}$ with some constant $C' > 0$. This implies that under the assumptions of Theorem 1.5 the KDE attains the optimal rate of convergence $n^{-\frac{\beta}{2\beta+1}}$ associated with the class of densities $\mathcal{P}(\beta, L)$. Exact definitions and discussions of the notion of optimal rate of convergence will be given later.

Remark 1.1. Quite often in practice it is assumed that $\beta = 2$ and that p'' is continuous at x_0. Also, the kernel is taken to be of order one and symmetric around 0. Then it can be shown that (Exercise (HW1))

$$\mathrm{MSE}(x_0) = \frac{1}{nh} \int K^2(u) du p(x_0) + \frac{1}{4} h^4 \left(\int u^2 K(u) du \right)^2 p''(x_0)^2 + o((nh)^{-1} + h^4).$$

Remark 1.2. Since $2\beta/(2\beta+1)$ approaches 1 as k becomes large, Theorem 1.5 implies that, for sufficiently smooth densities, the convergence rate can be made arbitrarily close to the parametric n^{-1} convergence rate. The fact that higher-order kernels

can achieve improved rates of convergence means that they will eventually dominate first-order kernel estimators for large n. However, this does not mean that a higher-order kernel will necessarily improve the error for sample sizes usually encountered in practice, and in many cases, unless the sample size is very large there may actually be an increase in the error due to using a higher-order kernel.

1.3 Pointwise asymptotic distribution

Whereas the results from the previous sub-section have shown us that $\hat{p}_n(x_0)$ converges to $p(x_0)$ in probability under certain assumptions, we cannot straightforwardly use this for statistical inference. Ideally, if we want to estimate $p(x_0)$ at the point x_0, we would like to have exact confidence statements of the form

$$\mathbb{P}\left(p(x_0) \in [\hat{p}_n(x_0) - c(n, \alpha, x_0, K), \hat{p}_n(x_0) - c(n, \alpha, x_0, K)]\right) \geq 1 - \alpha,$$

where α is the significance level and $c(n, \alpha, x_0, K)$ sequence of constants that one would like to be as small as possible (given α).

Theorem 1.6. Assume that $p \in \mathcal{P}(\beta, L)$ and let K be a kernel of order $\ell = \lfloor \beta \rfloor$ satisfying

$$\int |u|^\beta |K(u)| du < \infty.$$

Suppose that p also satisfies $p(x) \leq p_{\max} < \infty$ for all $x \in \mathbb{R}$. Let K further satisfy (a) $\|K\|_2^2 := \int K^2(u)du < \infty$, (b) $\|K\|_\infty := \sup_{u \in \mathbb{R}} K(u) < \infty$. Suppose that the sequence of bandwidths $\{h_n\}_{n=1}^\infty$ satisfy $h_n \to 0$, $nh_n \to \infty$, and $n^{1/2} h_n^{\beta+1/2} \to 0$ as $n \to \infty$. Then, as $n \to \infty$,

$$\sqrt{nh}\left(\hat{p}_n(x_0) - p(x_0)\right) \xrightarrow{d} N\Big(0, p(x_0)\|K\|_2^2\Big).$$

Proof. We first find the limit for the 'variance term'. We use the Lindeberg-Feller central limit theorem for triangular arrays of independent random variables[2] with

$$Y_{ni} := \sqrt{nh}\frac{1}{nh}K\left(\frac{X_i - x_0}{h}\right) = \sqrt{\frac{1}{nh}}K\left(\frac{X_i - x_0}{h}\right), \qquad i = 1, \ldots, n,$$

[2] **Lindeberg-Feller CLT** (see e.g., [15, p.20]): For each n let $Y_{n1}, \ldots, Y_{nn}$ be independent random variables with finite variances. If, as $n \to \infty$, (i) $\sum_{i=1}^n \mathbb{E}[Y_{ni}^2 I(|Y_{ni}| > \epsilon)] \to 0$, for every $\epsilon > 0$, and (ii) $\sum_{i=1}^n \mathbb{E}[(Y_{ni} - \mathbb{E}(Y_{ni}))^2] \to \sigma^2$, then

$$\sum_{i=1}^n (Y_{ni} - \mathbb{E}(Y_{ni})) \xrightarrow{d} N(0, \sigma^2), \qquad \text{as } n \to \infty.$$

so that $Y_{n1}, \ldots, Y_{nn}$ are i.i.d. and we have

$$\sqrt{nh}\left(\hat{p}_n(x_0) - \mathbb{E}_p[\hat{p}_n(x_0)]\right) = \sum_{i=1}^{n}(Y_{ni} - \mathbb{E}(Y_{ni})).$$

Thus, we only need to show that the two conditions in the Lindeberg-Feller CLT hold. Clearly,

$$\begin{aligned} n\mathbb{E}(Y_{ni}^2) &= \frac{1}{h}\int K^2\left(\frac{z-x_0}{h}\right)p(z)dz \\ &= \int K^2(u)\,p(x_0+uh)du \to p(x_0)\int K^2(u)du, \qquad \text{as } n\to\infty, \end{aligned}$$

by the dominated convergence theorem (DCT), since $p(\cdot)$ is continuous at x_0 and bounded on $\mathbb{R}$. Now,

$$\begin{aligned} n\mathbb{E}(Y_{ni})^2 &= \frac{1}{h}\left(\int K\left(\frac{z-x_0}{h}\right)p(z)dz\right)^2 = h\left(\int K(u)p(x_0+uh)du\right)^2 \\ &\leq h\|K\|_2^2\, p_{\max} \to 0, \qquad \text{as } n\to\infty, \end{aligned}$$

which shows that $\sum_{i=1}^{n}\mathbb{E}[(Y_{ni}-\mathbb{E}(Y_{ni}))^2] \to p(x_0)\int K^2(u)du$. Furthermore,

$$|Y_{ni}| \leq \frac{1}{\sqrt{nh}}\|K\|_\infty \to 0, \qquad \text{as } n\to\infty,$$

by the assumption on the sequence of bandwidths. Thus, $I(|Y_{ni}|>\epsilon)\to 0$, for every $\epsilon>0$ and by the DCT

$$\sum_{i=}^{n}\mathbb{E}[Y_{ni}^2 I(|Y_{ni}|>\epsilon)] = \mathbb{E}[nY_{n1}^2 I(|Y_{n1}|>\epsilon)] \to 0.$$

By (4) we see that the bias term can be bounded above as

$$\sqrt{nh}|b(x_0)| \leq \sqrt{nh}h^\beta \to 0, \qquad \text{as } n\to\infty.$$

Therefore, we have the desired result. □

Exercise (HW1): Suppose that you are given an i.i.d. sample from a bounded density p with bounded derivatives at x_0. Suppose that $c(\alpha, x_0)$ is such that $\mathbb{P}(-c(\alpha,x_0) \leq Z \leq c(\alpha,x_0)) = 1-\alpha$ where $Z \sim N(0, p(x_0))$. Use a kernel density estimator (with a suitable kernel) to obtain a 95 percent confidence interval (CI) for $p(x_0)$ in such a way that the size of the interval shrinks at rate $1/\sqrt{nh_n}$ as $n\to\infty$, and that h_n can be chosen so that this rate is 'almost' (say, up to a $\log n$ term) of order $n^{-1/3}$.

Exercise (HW1): Under the setup of Remark 1.1 and the assumption that $h = \alpha n^{-1/5}$, where $\alpha > 0$, find the asymptotic distribution of $\sqrt{nh}(\hat{p}_n(x_0) - p(x_0))$. Can this be used to construct a CI for $p(x_0)$? What are the advantages/disadvantages of using this result versus the setup of Theorem 1.6 with $\beta = 2$ to construct a CI for $p(x_0)$?

1.4 Integrated squared risk of kernel estimators

In Section 1.2 we have studied the behavior of the KDE $\hat{p}_n$ at an arbitrary fixed point x_0. It is also interesting to analyze the *global risk* of $\hat{p}_n$. An important global criterion is the mean integrated squared error (MISE):

$$\text{MISE} := \mathbb{E}_p \int [(\hat{p}_n(x) - p(x))^2] dx.$$

By Fubini's theorem,

$$\text{MISE} = \int \text{MSE}(x) dx = \int b^2(x) dx + \int \sigma^2(x) dx. \tag{8}$$

Thus, the MISE is represented as a sum of the bias term $\int b^2(x)dx$ and the variance term $\int \sigma^2(x)dx$. To obtain bounds on these terms, we proceed in the same manner as for the analogous terms of the MSE. Let us study first the variance term.

Proposition 1.7 (Variance of $\hat{p}_n$)**.** Let $K : \mathbb{R} \to \mathbb{R}$ be the kernel function such that

$$\int K^2(u) du < \infty.$$

Then for any $h > 0$, and $n \geq 1$ and any probability density p we have

$$\int \sigma^2(x) dx \leq \frac{1}{nh} \int K^2(u) du.$$

Proof. As in the proof of Proposition 1.1,

$$\sigma^2(x) = \frac{1}{nh^2} \mathbb{E}_p[\eta_1^2(x)] \leq \frac{1}{nh^2} \mathbb{E}_p \left[K^2 \left(\frac{X_1 - x}{h} \right) \right]$$

for all $x \in \mathbb{R}$. Therefore,

$$\begin{aligned} \int \sigma^2(x) dx &\leq \frac{1}{nh^2} \int \left[\int K^2 \left(\frac{z - x}{h} \right) p(z) dz \right] dx \\ &= \frac{1}{nh^2} \int p(z) \left[\int K^2 \left(\frac{z - x}{h} \right) dx \right] dz \\ &= \frac{1}{nh} \int K^2(u) du. \end{aligned}$$

□

The upper bound for the variance term in Proposition 1.7 does not require any condition on p: The result holds for any density. For the bias term in (8) the situation is different: We can only control it on a restricted subset of densities. As above, we

specifically assume that p is smooth enough. Since the MISE is a risk corresponding to the $L_2(\mathbb{R})$-norm, it is natural to assume that p is smooth with respect to this norm. *Sobolev* classes provide a popular way to describe smoothness in $L_2(\mathbb{R})$.

Definition 1.8. Let $\beta \geq 1$ be an integer and $L > 0$. The Sobolev class $\mathcal{S}(\beta, L)$ is defined as the set of all $\beta - 1$ differentiable functions $f : \mathbb{R} \to \mathbb{R}$ having absolutely continuous derivative $f^{(\beta-1)}$ and satisfying

$$\int (f^{(\beta)}(x))^2 dx \leq L^2.$$

Theorem 1.9. Suppose that, for an integer $\beta \geq 1$:

(i) the function K is a kernel of order $\beta - 1$ satisfying the conditions

$$\int K^2(u)du < \infty, \qquad \int |u|^\beta |K(u)|du < \infty;$$

(ii) the density $p \in \mathcal{S}(\beta, L)$ for some $\beta \geq 1$ and $L > 0$.

Then for all $n \geq 1$ and all $h > 0$ the mean integrated squared error of the KDE $\hat{p}_n$ satisfies

$$\text{MISE} \leq \frac{1}{nh} \int K^2(u)du + \frac{L^2 h^{2\beta}}{(\ell!)^2} \left(\int |u|^\beta |K(u)|du \right)^2.$$

Proof. We bound the variance term as in Proposition 1.7. Let $\ell = \beta - 1$. For the bias term, first note that using the integral form of the remainder term in the Taylor's theorem (see (5) and make the transformation $t \mapsto \frac{t-x}{uh}$),

$$p(x + uh) = p(x) + p'(x)uh + \ldots + \frac{(uh)^\ell}{(\ell - 1)!} \int_0^1 (1 - \tau)^{\ell-1} p^{(\ell)}(x + \tau uh)d\tau.$$

Since the kernel K is of order $\beta - 1$, we obtain

$$\begin{aligned} b(x) &= \int K(u) \frac{(uh)^\ell}{(\ell - 1)!} \left[\int_0^1 (1 - \tau)^{\ell-1} p^{(\ell)}(x + \tau uh)d\tau \right] du \\ &= \int K(u) \frac{(uh)^\ell}{(\ell - 1)!} \left[\int_0^1 (1 - \tau)^{\ell-1} \Big(p^{(\ell)}(x + \tau uh) - p^{(\ell)}(x) \Big) d\tau \right] du \end{aligned}$$

Applying the generalized Minkowski inequality[3] twice and using the given assump-

[3] **Generalized Minkowski inequality**:

tions on p, we get the following upper bound for the bias term $\int b^2(x)dx$:

$$\begin{aligned}
&\int \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \int_0^1 (1-\tau)^{\ell-1} \left| p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x) \right| d\tau du \right)^2 dx \\
\leq \; & \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \left[\int \left(\int_0^1 (1-\tau)^{\ell-1} \left| p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x) \right| d\tau \right)^2 dx \right]^{1/2} du \right)^2 \\
\leq \; & \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \left[\int_0^1 (1-\tau)^{\ell-1} \left\{ \int \left(p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x) \right)^2 dx \right\}^{1/2} d\tau \right] du \right)^2 .
\end{aligned}$$

Now, for $t := \tau uh$,

$$\begin{aligned}
& \int \left(p^{(\ell)}(x+t) - p^{(\ell)}(x) \right)^2 dx \\
= \; & \int \left(t \int_0^1 p^{(\ell+1)}(x+\theta t) d\theta \right)^2 dx \\
\leq \; & t^2 \int \left(\int_0^1 \left[\int \left(p^{(\ell+1)}(x+\theta t) \right)^2 dx \right]^{1/2} d\theta \right)^2 = t^2 \int \left(p^{(\beta)}(x) \right)^2 dx
\end{aligned}$$

in view of the generalized Minskowski inequality. Therefore,

$$\begin{aligned}
\int b^2(x)dx \; &\leq \; \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \left[\int_0^1 (1-\tau)^{\ell-1} |\tau uh| L d\tau \right] du \right)^2 \\
&\leq \; \frac{L^2 h^{2(\ell+1)}}{[(\ell-1)!]^2} \left(\int |K(u)||u|^{\ell+1} \, du \right)^2 \left[\int_0^1 (1-\tau)^{\ell-1} d\tau \right]^2 \\
&\leq \; \frac{L^2 h^{2\beta}}{(\ell!)^2} \left(\int |u|^\beta |K(u)| \, du \right)^2
\end{aligned}$$

□

Exercise (HW1): Assume that:

(i) the function K is a kernel of order 1 satisfying the conditions

$$\int K^2(u)du < \infty, \qquad \int u^2 |K(u)| du < \infty, \qquad S_K := \int u^2 K(u) du \neq 0;$$

Lemma 1.10. For any Borel function g on $\mathbb{R} \times \mathbb{R}$, we have

$$\int \left(\int g(u,x) du \right)^2 dx \leq \left[\int \left(\int g^2(u,x) dx \right)^{1/2} du \right]^2 .$$

(ii) The density p is differentiable on $\mathbb{R}$, the first derivative p' is absolutely continuous on $\mathbb{R}$ and the second derivative satisfies $\int (p''(x))^2 dx < \infty$.

Then for all $n \geq 1$ the mean integrated squared error of the kernel estimator $\hat{p}_n$ satisfies

$$\text{MISE} = \left[\frac{1}{nh} \int K^2(u)du + \frac{h^4}{4} S_K^2 \int (p''(x))^2 dx \right] (1 + o(1)),$$

where the term $o(1)$ is independent of n (but depends on p) and tends to 0 as $h \to 0$.

1.5 Unbiased risk estimation: cross-validation

Let $\hat{p}_n$ be the KDE and let the kernel K be fixed. We already know that the bandwidth h is crucial to determine the behavior of the estimator. How to choose h in practice?

Consider the risk

$$\text{MISE}(h) := \mathbb{E}_p \int (\hat{p}_n^{(h)} - p)^2(x)dx.$$

The optimal value of h is the one that minimizes the MISE, i.e.,

$$h^* = \operatorname*{argmin}_{h>0} \text{MISE}(h).$$

This ideal bandwidth h depends on the true density p, so it is not available in practice. It is called the *oracle bandwidth*, and the estimator $\hat{p}_n$ with bandwidth $h = h^*$ is called the *oracle*. We would like to "mimic the oracle", i.e., to find a bandwidth $\hat{h}_n$ that only depends on the data $X_1, \ldots, X_n$, such that its risk is close to the risk of the oracle:

$$\mathbb{E}_p \int (\hat{p}_n^{(\hat{h}_n)} - p)^2(x)dx \;\approx\; \min_{h>0} \text{MISE}(h),$$

It turns out that this task can be achieved. The idea is to first estimate the MISE$(\cdot)$, and then to minimize in h the obtained estimator of MISE$(\cdot)$.

Note that the MISE can be written as

$$\text{MISE}(h) = \mathbb{E} \int (\hat{p}_n - p)^2 = \mathbb{E} \left[\int \hat{p}_n^2 - 2 \int \hat{p}_n p \right] + \int p^2.$$

Only the expression in the square brackets depends on h; the last term is constant in h. Let

$$J(h) := \mathbb{E}_p \left[\int \hat{p}_n^2 - 2 \int \hat{p}_n p \right].$$

Since we are minimizing over h, minimizing MISE(h) is equivalent to minimizing $J(h)$. Therefore, it is enough to look for an estimator of $J(h)$, denoted by $\hat{J}(h)$, because

MISE(h) and $J(h)$ have the same minimizers. A first idea is to take an unbiased estimator of $J(h)$.

[9] suggested the following estimator:

$$\hat{J}(h) \equiv \mathrm{CV}(h) = \int \hat{p}_n^2 - \frac{2}{n}\sum_{i=1}^n \hat{p}_{n,-i}(X_i),$$

where CV stands for cross-validation, and

$$\hat{p}_{n,-i}(x) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{X_j - x}{h}\right).$$

Now we prove that CV(h) is an unbiased estimator of $J(h)$, i.e., we show that

$$\mathbb{E}_p \int \hat{p}_n p = \mathbb{E}_p\left[\frac{1}{n}\sum_{i=1}^n \hat{p}_{n,-i}(X_i)\right]. \tag{9}$$

Since $X_1, \ldots, X_n$ are i.i.d., the right hand side of (9) is equal to

$$\begin{aligned}
\mathbb{E}_p[\hat{p}_{n,-1}(X_1)] &= \mathbb{E}_p\left[\frac{1}{(n-1)h}\sum_{j\neq 1}\int K\left(\frac{X_j - z}{h}\right)p(z)dz\right] \\
&= \frac{1}{h}\int p(x)\int K\left(\frac{x-z}{h}\right)p(z)dz\,dx
\end{aligned}$$

This integral is finite if K is bounded. The left hand side of (9) is equal to

$$\mathbb{E}_p\left[\frac{1}{nh}\sum_{i=1}^n \int K\left(\frac{X_i - x}{h}\right)p(x)dx\right] = \text{RHS of (9)}.$$

Define the cross-validated bandwidth and the cross-validated KDE:

$$\begin{aligned}
\hat{h}_{\mathrm{CV}} &= \operatorname*{argmin}_{h>0} \mathrm{CV}(h), \\
\tilde{p}_n^{\mathrm{CV}}(x) &= \frac{1}{n\hat{h}_{\mathrm{CV}}}\sum_{i=1}^n K\left(\frac{X_i - x}{\hat{h}_{\mathrm{CV}}}\right).
\end{aligned}$$

[12] was the first to investigate the issue of optimality in connection with cross-validation. He proved that the integrated squared error of the estimator $\tilde{p}_n^{\mathrm{CV}}$ is asymptotically equivalent to that of some oracle estimator:

$$\frac{\int(\tilde{p}_n^{\mathrm{CV}} - p)^2}{\min_{h>0}\int(\hat{p}_n^{(h)} - p)^2} \overset{a.s.}{\to} 1, \qquad n \to \infty,$$

under some assumptions (the density p is bounded, the kernel is compactly supported, essentially nonnegative, and satisfies the Hölder condition).

2 Nonparametric regression

Let (X, Y) be a pair of real-valued random variables such that $\mathbb{E}|Y| < \infty$. The regression function $f : \mathbb{R} \to \mathbb{R}$ of Y on X is defined as

$$f(x) = \mathbb{E}(Y|X = x).$$

Suppose that we have a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n i.i.d. pairs of random variables having the same distribution as (X, Y). We would like to estimate the regression function f from the data. The nonparametric approach only assumes that $f \in \mathcal{F}$, where $\mathcal{F}$ is a given nonparametric class of functions. The set of values $\{X_1, \ldots, X_n\}$ is called the *design*. Here the design is random.

The conditional residual $\xi := Y - \mathbb{E}(Y|X)$ has mean zero, $\mathbb{E}(\xi) = 0$, and we may write

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{10}$$

where ξ_i are i.i.d. random variables with the same distribution as ξ. In particular, $\mathbb{E}(\xi_i) = 0$ for all $i = 1, \ldots, n$. The variables ξ_i can therefore be interpreted as "errors".

The key idea we use in estimating f nonparametrically in this section is called *"local averaging"*. Given a kernel K and a bandwidth h, one can construct kernel estimators for nonparametric regression. There exist different types of kernel estimators of the regression function f. The most celebrated one is the Nadaraya-Watson estimator defined as follows:

$$f_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}, \qquad \text{if } \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0,$$

and $f_n^{NW}(x) = 0$ otherwise. This estimator was proposed separately in two papers by Nadaraya and Watson in the year 1964.

Example: If we choose $K(u) = \frac{1}{2} I(|u| \leq 1)$, then $f_n^{NW}(x)$ is the average of Y_i such that $X_i \in [x - h, x + h]$. Thus, for estimating $f(x)$ we define the "local" neighborhood as $[x - h, x + h]$ and consider the average of the observations in that neighborhood. For fixed n, the two extreme cases for the bandwidth are:

(i) $h \to \infty$. Then $f_n^{NW}(x)$ tends to $n^{-1} \sum_{i=1}^n Y_i$ which is a constant independent of x. The systematic error (bias) can be too large. This is a situation of *oversmoothing*.

(ii) $h \to 0$. Then $f_n^{NW}(X_i) = Y_i$ whenever $h < \min_{i,j} |X_i - X_j|$ and $\lim_{h \to 0} f_n^{NW}(x) = 0$, if $x \neq X_i$. The estimator f_n^{NW} is therefore too oscillating: it reproduces the

data Y_i at the points X_i and vanishes elsewhere. This makes the stochastic error (variance) too large. In other words, *undersmoothing* occurs.

Thus, the bandwidth h defines the "width" of the local neighborhood and the kernel K defines the "weights" used in averaging the response values in the local neighborhood. As we saw in density estimation, an appropriate choice of the bandwidth h is more important than the choice of the kernel K.

The Nadaraya-Watson estimator can be represented as a weighted sum of the Y_i:

$$f_n^{NW}(x) = \sum_{i=1}^{n} Y_i W_i^{NW}(x)$$

where the weights are

$$W_i^{NW}(x) := \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)} I\left(\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right) \neq 0\right).$$

Definition 2.1. An estimator $\hat{f}_n(x)$ of $f(x)$ is called a *linear nonparametric regression estimator* if it can be written in the form

$$\hat{f}_n(x) = \sum_{i=1}^{n} Y_i W_{ni}(x)$$

where the weights $W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n)$ depend only on n, i, x and the values $X_1, \ldots, X_n$.

Typically, the weights $W_{ni}(x)$ of linear regression estimators satisfy the equality $\sum_{i=1}^{n} W_{ni}(x) = 1$ for all x (or for almost all x with respect to the Lebesgue measure).

Another intuitive motivation of f_n^{NW} is given below. Suppose that the distribution of (X, Y) has density $p(x, y)$ with respect to the Lebesgue measure and $p^X(x) = \int p(x, y)dy > 0$. Then,

$$f(x) = \mathbb{E}(Y|X = x) = \frac{\int y p(x, y)dy}{p^X(x)}.$$

If we replace here $p(x, y)$ by the KDE $\hat{p}_n(x, y)$ of the density of (X, Y) defined by (3) and use the corresponding KDE $\hat{p}_n^X(x)$ to estimate $p^X(x)$, we obtain $\hat{f}_n^{NW}$ in view of the following result.

Exercise (HW1): Let $\hat{p}_n^X(x)$ and $\hat{p}_n(x, y)$ be the KDEs defined in (2) and (3) respectively, with a kernel K of order 1. Then

$$f_n^{NW}(x) = \frac{\int y \hat{p}_n(x, y)dy}{\hat{p}_n^X(x)}$$

if $\hat{p}_n^X(x) \neq 0$.

2.1 Local polynomial estimators

If the kernel K takes only nonnegative values, the Nadaraya-Watson estimator f_n^{NW} satisfies

$$f_n^{NW}(x) = \operatorname*{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \theta)^2 K\left(\frac{X_i - x}{h}\right). \tag{11}$$

Thus, f_n^{NW} is obtained by a local constant least squares approximation of the response values, i.e., Y_i's. The locality is determined by the bandwidth h and the kernel K which downweighs all the X_i that are not close to x whereas θ plays the role of a local constant to be fitted. More generally, we may define a local polynomial least squares approximation, replacing in (11) the constant θ by a polynomial of given degree ℓ. If $f \in \Sigma(\beta, L), \beta > 1, \ell = \beta$, then for z sufficiently close to x we may write

$$f(z) \approx f(x) + f'(x)(z - x) + \ldots + \frac{f^{(\ell)}(x)}{\ell!}(z - x)^\ell = \theta^\top(x) U\left(\frac{z - x}{h}\right),$$

where

$$\begin{aligned} U(u) &= (1, u, u^2/2!, \ldots, u^\ell/\ell!), \\ \theta(x) &= \left(f(x), f'(x)h, f''(x)h^2, \ldots, f^{(\ell)}(x)h^\ell\right)^\top. \end{aligned}$$

Definition 2.2. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel, $h > 0$ be a bandwidth, and $\ell \geq 0$ be an integer. A vector $\hat{\theta}_n(x) \in \mathbb{R}^{\ell+1}$ defined by

$$\hat{\theta}_n(x) = \operatorname*{argmin}_{\theta \in \mathbb{R}^{\ell+1}} \sum_{i=1}^{n} \left[Y_i - \theta^\top U\left(\frac{X_i - x}{h}\right)\right]^2 K\left(\frac{X_i - x}{h}\right) \tag{12}$$

is called a *local polynomial estimator of order ℓ of $\theta(x)$* or LP(ℓ) estimator of $\theta(x)$ for short. The statistic

$$\hat{f}_n(x) = U^\top(0)\hat{\theta}_n(x)$$

is called a *local polynomial estimator of order ℓ of f(x)* or LP(ℓ) estimator of $f(x)$ for short.

Note that $\hat{f}_n(x)$ is simply the first coordinate of the vector $\hat{\theta}_n(x)$. Comparing (11) and (12) we see that the Nadaraya-Watson estimator f_n^{NW} with kernel $K \geq 0$ is the $LP(0)$ estimator. Furthermore, properly normalized coordinates of $\hat{\theta}_n(x)$ provide estimators of the derivatives $f'(x), \ldots, f^{(\ell)}(x)$.

For a fixed x the estimator (12) is a weighted least squares estimator. Indeed, we can write $\hat{\theta}_n(x)$ as follows:

$$\hat{\theta}_n(x) = \underset{\theta \in \mathbb{R}^{\ell+1}}{\operatorname{argmin}}(-2\theta^\top a_{nx} + \theta^\top \mathcal{B}_{nx}\theta),$$

where the matrix $\mathcal{B}_{nx}$ and the vector a_{nx} are defined by the formulas:

$$\begin{aligned}
\mathcal{B}_{nx} &= \frac{1}{nh}\sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) U^\top\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right), \\
a_{nx} &= \frac{1}{nh}\sum_{i=1}^{n} Y_i U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right).
\end{aligned}$$

Exercise (HW1): If the matrix $\mathcal{B}_{nx}$ is positive definite, show that the local polynomial estimator $\hat{f}_n(x)$ of $f(x)$ is a linear estimator. Also, in this case, find an expression for $\hat{f}_n(x)$.

The local polynomial estimator of order ℓ has a remarkable property: It reproduces polynomials of degree $\leq \ell$. This is stated in the next proposition (Exercise (HW1)).

Proposition 2.3. Let $x \in \mathbb{R}$ be such that $\mathcal{B}_{nx} > 0$ (i.e., $\mathcal{B}_{nx}$ is positive definite) and let Q be a polynomial of degree $\leq \ell$. Then the LP(ℓ) weights W^*_{ni} are such that

$$\sum_{i=1}^{n} Q(X_i) W^*_{ni}(x) = Q(x),$$

for any sample $(X_1, \ldots, X_n)$. In particular,

$$\sum_{i=1}^{n} W^*_{ni}(x) = 1, \quad \text{and} \quad \sum_{i=1}^{n} (X_i - x)^k W^*_{ni}(x) = 0 \quad \text{for} \quad k = 1, \ldots, \ell.$$

2.2 Pointwise and integrated risk of local polynomial estimators

In this section we study statistical properties of the LP(ℓ) estimator constructed from observations $(X_i, Y_i), i = 1, \ldots, n$, such that

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{13}$$

where ξ_i are independent zero mean random variables ($\mathbb{E}(\xi_i) = 0$), the X_i are deterministic values belonging to $[0, 1]$, and f is a function from $[0, 1]$ to $\mathbb{R}$.

Let $\hat{f}_n(x_0)$ be an LP(ℓ) estimator of $f(x_0)$ at point $x_0 \in [0,1]$. The bias and the variance of $\hat{f}_n(x_0)$ are given by the formulas

$$b(x_0) = \mathbb{E}_f\Big[\hat{f}_n(x_0)\Big] - f(x_0), \qquad \sigma^2(x_0) = \mathbb{E}_f\Big[\hat{f}_n^2(x_0)\Big] - \Big(\mathbb{E}_f\Big[\hat{f}_n(x_0)\Big]\Big)^2,$$

respectively, where $\mathbb{E}_f$ denotes expectation with respect to the distribution of the random vector $(Y_1, \ldots, Y_n)$. We will sometimes write for brevity $\mathbb{E}$ instead of $\mathbb{E}_f$.

We will study separately the bias and the variance terms in this representation of the risk. First, we introduce the following assumptions.

Assumptions (LP)

(LP1) There exist a real number $\lambda_0 > 0$ and a positive integer n_0 such that the smallest eigenvalue $\lambda_{\min}(\mathcal{B}_{nx})$ of $\mathcal{B}_{nx}$ satisfies $\lambda_{\min}(\mathcal{B}_{nx}) \geq \lambda_0$ for all $n \geq n_0$ and any $x \in [0,1]$.

(LP2) There exists a real number $a_0 > 0$ such that for any interval $A \subset [0,1]$ and all $n \geq 1$,

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \leq a_0 \max(\mathrm{Leb}(A), 1/n)$$

where $\mathrm{Leb}(A)$ denotes the Lebesgue measure of A.

(LP3) The kernel K has compact support belonging to $[-1,1]$ and there exists a number $K_{\max} < \infty$ such that $|K(u)| \leq K_{\max}, \forall\, u \in \mathbb{R}$.

Assumption (LP1) is stronger than the condition $\mathcal{B}_{nx} > 0$ introduced before since it is uniform with respect to n and x. We will see that this assumption is natural in the case where the matrix $\mathcal{B}_{nx}$ converges to a limit as $n \to \infty$. Assumption (LP2) means that the points X_i are dense enough in the interval [0, 1]. It holds for a sufficiently wide range of designs. An important example is given by the regular design: $X_i = i/n$, for which (LP2) is satisfied with $a_0 = 2$. Finally, assumption (LP3) is not restrictive since the choice of K belongs to the statistician.

Exercise (HW1): Show that assumption (LP1) implies that, for all $n \geq n_0$, $x \in [0,1]$, and $v \in \mathbb{R}^{\ell+1}$,

$$\|\mathcal{B}_{nx}^{-1} v\| \leq \|v\|/\lambda_0,$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^{\ell+1}$. Hint: Use the fact that $\mathcal{B}_{nx}$ is symmetric and relate the eigenvalues of $\mathcal{B}_{nx}$ to that of $\mathcal{B}_{nx}^{-1}$ and $\mathcal{B}_{nx}^{-2}$ (note that for a square matrix $A \in \mathbb{R}^{r\times r}$, $\lambda_{max}(A) = \frac{v^\top A v}{\|v\|^2}$, where $v \neq 0 \in \mathbb{R}^r$).

We have the following result (Exercise (HW1)) which gives us some useful bounds on the weights $W^*_{ni}(x)$.

Lemma 2.4. Under assumptions (LP1)–(LP3), for all $n \geq n_0$, $h \geq 1/(2n)$, and $x \in [0,1]$, the weights $W^*_{ni}(x)$ of the LP(ℓ) estimator are such that:

(i) $\sup_{i,x} |W^*_{ni}(x)| \leq \frac{C_*}{nh}$;

(ii) $\sum_{i=1}^n |W^*_{ni}(x)| \leq C_*$;

(iii) $W^*_{ni}(x) = 0$ if $|X_i - x| > h$,

where the constant C_* depends only on λ_0, a_0, and $K_{\max}$.

We are now ready to find upper bounds on the MSE of the LP(ℓ) estimator.

Proposition 2.5. Suppose that $f \in \Sigma(\beta, L)$ on $[0,1]$, with $\beta > 0$ and $L > 0$. Let $\hat{f}_n$ be the LP(ℓ) estimator of f with $\ell = \lfloor \beta \rfloor$. Assume also that:

(i) the design points $X_1, \ldots, X_n$ are deterministic;

(ii) assumptions (LP1)–(LP3) hold;

(iii) the random variables ξ_i are independent and such that for all $i = 1, \ldots, n$,

$$\mathbb{E}(\xi_i) = 0, \qquad \mathbb{E}(\xi_i^2) \leq \sigma^2_{\max} < \infty.$$

Then for all $x_0 \in [0,1]$, $n \geq n_0$, and $h \geq 1/(2n)$ the following upper bounds hold:

$$|b(x_0) \leq q_1 h^\beta, \qquad \sigma^2(x_0) \leq \frac{q_2}{nh},$$

where $q_1 := C_* L/\ell!$ and $q_2 := \sigma^2_{\max} C_*^2$.

Thus, Proposition 2.5 implies that

$$\text{MSE} \leq q_1^2 h^{2\beta} + \frac{q_2}{nh}. \tag{14}$$

The minimum with respect to h of the right hand side of the above upper bound is attained at

$$h_n^* = \left(\frac{q_2}{2\beta q_2^2}\right)^{1/(2\beta+1)} n^{-1/(2\beta+1)}.$$

Therefore, the choice $h = h_n^*$ gives

$$\text{MSE}(x_0) = O\Big(n^{-\frac{2\beta}{2\beta+1}}\Big), \quad \text{as } n \to \infty,$$

uniformly in x_0. Thus, we have the following result.

Theorem 2.6. Assume that the assumptions of Proposition 2.5 hold. Suppose that for a fixed $\alpha > 0$ the bandwidth is chosen as $h = h_n = \alpha n^{-1/(2\beta+1)}$. Then the following holds:

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \sup_{x_0\in[0,1]} \mathbb{E}_f\left[\psi_n^{-2}(\hat{f}_n(x_0) - f(x_0))^2\right] \le C < \infty,$$

where $\psi_n := n^{-\frac{\beta}{2\beta+1}}$ is the rate of convergence and $C > 0$ is a constant depending only on $\beta, L, a_0, \sigma^2_{\max}, K_{\max}$ and α.

As the above upper bound holds for every $x_0 \in [0,1]$ we immediately get the following result on the integrated risk.

Corollary 2.7. Under the assumptions of Theorem 2.6 the following holds:

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f\left[\psi_n^{-2}\|\hat{f}_n(x_0) - f(x_0)\|_2^2\right] \le C < \infty,$$

where $\|f\|_2^2 = \int_0^1 f^2(x)dx$, $\psi_n := n^{-\frac{\beta}{2\beta+1}}$ is the rate of convergence and $C > 0$ is a constant depending only on $\beta, L, a_0, \sigma^2_{\max}, K_{\max}$ and α.

2.2.1 Assumption (LP1)

We now discuss assumption (LP1) in more detail. If the design is regular and n is large enough, $\mathcal{B}_{nx}$ is close to the matrix $\mathcal{B} := \int U(u)U^\top(u)K(u)du$, which is independent of n and x. Therefore, for Assumption (LP1) to hold we only need to assure that $\mathcal{B}$ is positive definite. This is indeed true, except for pathological cases, as the following lemma states.

Lemma 2.8. Let $K : \mathbb{R} \to [0,\infty)$ be a function such that the Lebesgue measure $\mathrm{Leb}(\{u : K(u) > 0\}) > 0$. Then the matrix

$$\mathcal{B} = \int U(u)U^\top(u)K(u)du$$

is positive definite.

Proof. It is sufficient to prove that for all $v \in \mathbb{R}^{\ell+1}$ satisfying $v \neq 0$, we have $v^\top\mathcal{B}v > 0$. Clearly,

$$v^\top\mathcal{B}v > 0 = \int (v^\top U(u))^2 K(u)du \ge 0.$$

If there exists $v \neq 0$ such that $\int (v^\top U(u))^2 K(u)du = 0$, then $v^\top U(u) = 0$ for almost all u on the set $\{u : K(u) > 0\}$, which has a positive Lebesgue measure by the assumption of the lemma. But the function $v \mapsto v^\top U(u)$ is a polynomial of degree

$\leq \ell$ which cannot be equal to zero except for a finite number of points. Thus, we come to a contradiction showing that $\int (v^\top U(u))^2 K(u)du = 0$ is impossible for $v \neq 0$. $\square$

Lemma 2.9. Suppose that there exist $K_{\min} > 0$ and $\Delta > 0$ such that

$$K(u) \geq K_{\min} I(|u| \leq \Delta), \quad \forall u \in \mathbb{R},$$

and that $X_i = i/n$ for $i = 1, \ldots, n$. Let $h = h_n$ be a sequence satisfying $h_n \to 0$ and $nh_n \to \infty$, as $n \to \infty$. Then assumption (LP1) holds.

3 Projection estimators

Consider data (X_i, Y_i), $i = 1, \ldots, n$, from a nonparametric regression model where

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{15}$$

with $X_i \in \mathfrak{X}$, a metric space, and $\mathbb{E}(\xi_i) = 0$. The goal is to estimate the function f based on the data. In what follows, we will also use the vector notation, writing the model as

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\xi},$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^\top, \mathbf{f} = (f(X_1), \ldots, f(X_n))^\top$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$.

The idea here is to approximate f by $f_{\boldsymbol{\theta}}$, a linear combination of N given functions $\varphi_1, \ldots, \varphi_N$ where $\varphi_j : \mathfrak{X} \to \mathbb{R}$, so that

$$f_{\boldsymbol{\theta}}(x) := \sum_{j=1}^{N} \theta_j \varphi_j(x).$$

Then we look for a suitable estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_N)$ of $\boldsymbol{\theta}$ based on the sample (X_i, Y_i), $i = 1, \ldots, n$, and construct an estimator of f having the form

$$\hat{f}(x) = f_{\hat{\boldsymbol{\theta}}}(x) = \sum_{j=1}^{N} \hat{\theta}_j \varphi_j(x). \tag{16}$$

Example 3.1. If $\mathfrak{X} = [0, 1]$ and $f \in L_2[0, 1]$, then a popular choice of $\{\varphi_j\}_{j=1}^N$ corresponds to the first N functions of an orthonormal basis in $L_2[0, 1]$. For example, $\{\varphi_j\}_{j=1}^\infty$ can be the trigonometric basis or the Legendre basis on $[0, 1]$. Let $\{\theta_j\}_{j=1}^\infty$ be the Fourier coefficients of f with respect to the orthonormal basis $\{\varphi_j\}_{j=1}^\infty$ of $L_2[0, 1]$, i.e.,

$$\theta_j = \int_0^1 f(x)\varphi(x)dx.$$

Assume that f can be represented as

$$f(x) = \sum_{j=1}^{\infty} \theta_j \varphi_j(x), \tag{17}$$

where the series converges for all $x \in [0,1]$. Observe that if X_i are scattered over $[0,1]$ in a sufficiently uniform way, which happens, e.g., in the case $X_i = i/n$, the coefficients θ_j are well approximated by the sums $n^{-1}\sum_{i=1}^{N} f(X_i)\varphi_j(X_i)$. Replacing in these sums the unknown quantities $f(X_i)$ by the observations Y_i we obtain the following estimator of θ_j:

$$\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{\infty} Y_i \varphi_j(X_i). \tag{18}$$

Remark 3.1. The parameter N (called the order of the estimator) plays the same role as the bandwidth h for kernel estimators: similar to h it is a smoothing parameter, i.e., a parameter whose choice is crucial for establishing the balance between bias and variance. The choice of very large N leads to undersmoothing, whereas for small values of N oversmoothing occurs.

An important class of estimators of the form (16) are projection estimators. Define the empirical norm $\|\cdot\|$ as:

$$\|\mathbf{f}\|^2 := \sum_{i=1}^{n} f^2(X_i), \qquad \|\mathbf{y}\|^2 := \sum_{i=1}^{n} Y_i^2.$$

The *projection estimator* is defined as follows:

$$\hat{f}^{LS}(x) = f_{\hat{\boldsymbol{\theta}}^{LS}}(x) = \sum_{j=1}^{N} \hat{\theta}_j^{LS} \varphi_j(x) \tag{19}$$

where $\hat{\boldsymbol{\theta}}^{LS}$ is the classical least squares estimator (LSE):

$$\hat{\boldsymbol{\theta}}^{LS} := \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^N} \|\mathbf{y} - \mathbf{f}_{\boldsymbol{\theta}}\|^2,$$

where $\mathbf{f}_{\boldsymbol{\theta}} = (f_{\boldsymbol{\theta}}(X_1), \ldots, f_{\boldsymbol{\theta}}(X_n))^\top$. Equivalently, we can write

$$\hat{\boldsymbol{\theta}}^{LS} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2,$$

where $\mathbf{X} := (\varphi_j(X_i))_{i,j}$ where $i = 1, \ldots, n$ and $j = 1, \ldots, N$. In other words, we construct a 'nonparametric' estimator based on a purely parametric idea. The question is whether such an estimator is good. We will see that this is indeed the case under

appropriate conditions on the functions $\{\varphi_j\}$, the function f, and N. Recall that, under the assumption that $\mathbf{X}^\top\mathbf{X} > 0$ (note that $\mathbf{X}^\top\mathbf{X}$ is an $N \times N$ matrix), we have

$$\hat{\boldsymbol{\theta}}^{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \quad \text{and} \quad \hat{\mathbf{f}}^{LS} = \mathbf{X}\hat{\boldsymbol{\theta}}^{LS} = A\mathbf{y}$$

where $A := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the so-called *hat matrix*. The hat matrix is the orthogonal projection matrix (in $\mathbb{R}^n$) onto the column-space of $\mathbf{X}$, i.e., the subspace of $\mathbb{R}^n$ spanned by the N columns of $\mathbf{X}$. Note that we can have $\mathbf{X}^\top\mathbf{X} > 0$ only if $N \leq n$. However, even if $\mathbf{X}^\top\mathbf{X}$ is not invertible $\hat{\mathbf{f}}^{LS}$ is uniquely defined by the Hilbert projection theorem[4] and can be expressed as $A\mathbf{y}$ where now $A = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^+\mathbf{X}^\top$; here A^+ stands for the Moore-Penrose pseudoinverse.

Indeed, $\text{rank}(\mathbf{X}^\top\mathbf{X}) = \text{rank}(\mathbf{X}) \leq \min(N, n)$. Under the assumption that $\mathbf{X}^\top\mathbf{X} > 0$, the projection estimator is unique and has the form

$$\hat{f}^{LS}(x) = \boldsymbol{\varphi}(x)^\top\hat{\boldsymbol{\theta}}^{LS} = \boldsymbol{\varphi}(x)^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \sum_{i=1}^{n} W_{ni}(x)Y_i,$$

where $\boldsymbol{\varphi}(x) = (\varphi_1(x), \ldots, \varphi_N(x))^\top$ and $W_{ni}(x)$ is the i-th component of the vector $\boldsymbol{\varphi}(x)^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

3.1 Risk bounds for projection estimators

Assume now that we have the regression model (48), where the points X_i are deterministic elements in the space $\mathfrak{X}$. Let us measure the accuracy of an estimator $\hat{f}$ of f by the following squared risk:

$$R(\mathbf{f}, \hat{\mathbf{f}}) := \mathbb{E}\|\mathbf{f} - \hat{\mathbf{f}}\|^2 = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - f(X_i))^2\right].$$

This choice of a loss function is quite natural and it measures the prediction accuracy of the estimator at the observed design points. Further, if the X_i are "equi-spaced" points then $R(\mathbf{f}, \hat{\mathbf{f}})$ is approximately equal to the MISE.

Let $\hat{f}(x)$ be a linear estimator, i.e., $\hat{f}(x) = \sum_{i=1}^{n} W_{ni}(x)Y_i$. Then we can write $\hat{\mathbf{f}} = S\mathbf{y}$ where $S := (W_{nj}(X_i))_{n\times n}$ is a deterministic matrix. Note that S does not depend on

[4]The Hilbert projection theorem is a famous result of convex analysis that says that for every point u in a Hilbert space H and every nonempty closed convex $C \subset H$, there exists a unique point $v \in C$ for which $\|x - y\|$ is minimized over C. This is, in particular, true for any closed subspace M of C. In that case, a necessary and sufficient condition for v is that the vector $u - v$ be orthogonal to M.

$\mathbf{y}$; it depends only on the X_i's. As particular cases, we can think of $\hat{f}$ as the LP(ℓ) estimator or the projection estimator in (16).

Proposition 3.2. Let ξ_i be random variables such that $\mathbb{E}(\xi_i) = 0$ and $\mathbb{E}(\xi_i\xi_j) = \sigma^2\delta_{ij}$ for $i, j = 1, \ldots, n$, where δ_{ij} is the Kronecker delta function. Let S be any $n \times n$ matrix. Then the risk of linear estimator $\mathbf{f} = S\mathbf{y}$ is given by

$$R(\mathbf{f}, \hat{\mathbf{f}}) = \|S\mathbf{f} - \mathbf{f}\|^2 + \frac{\sigma^2}{n}\mathrm{tr}(S^\top S).$$

Proof. By definition of the norm $\|\cdot\|$ and of the model,

$$\begin{aligned} \|\hat{\mathbf{f}} - \mathbf{f}\|^2 &= \|S\mathbf{f} + S\boldsymbol{\xi} - \mathbf{f}\|^2 \\ &= \|S\mathbf{f} - \mathbf{f}\|^2 + \frac{2}{n}(S\mathbf{f} - \mathbf{f})^\top S\boldsymbol{\xi} + \|S\boldsymbol{\xi}\|^2. \end{aligned}$$

Taking expectations and using that $E(\boldsymbol{\xi}) = 0$ we obtain

$$\mathbb{E}\|\hat{\mathbf{f}} - \mathbf{f}\|^2 = \|S\mathbf{f} - \mathbf{f}\|^2 + \frac{1}{n}\mathbb{E}(\boldsymbol{\xi}^\top S^\top S\boldsymbol{\xi}).$$

Set $V = S^\top S$ and denote the elements of this matrix by v_{ij}. We have

$$\mathbb{E}(\boldsymbol{\xi}^\top S^\top S\boldsymbol{\xi}) = \mathbb{E}\left(\sum_{i,j}^{n} \xi_i v_{ij} \xi_j\right) = \sum_{i=1}^{n} \sigma^2 v_{ii} = \sigma^2 \mathrm{tr}(V).$$

□

In particular, if $\hat{\mathbf{f}}$ is a projection estimator then S is an orthogonal projection matrix and $S^\top = S$ and thus, $V = S$ (as $S^2 = S$) which shows that

$$\mathrm{tr}(V) = \mathrm{tr}(S) = \mathrm{rank}(S) \leq \min(n, N).$$

Thus, we have

$$\begin{aligned} R(\mathbf{f}, \hat{\mathbf{f}}) &\leq \|S\mathbf{f} - \mathbf{f}\|^2 + \frac{\sigma^2}{n}\min(n, N) \\ &= \min_{\boldsymbol{\theta} \in \mathbb{R}^N} \|\mathbf{f}_{\boldsymbol{\theta}} - \mathbf{f}\|^2 + \frac{\sigma^2}{n}\min(n, N). \end{aligned} \tag{20}$$

In fact, a close inspection of the proof of Proposition 3.2 shows that for the above inequality to hold it is enough to assume that $\mathbb{E}(\xi_i^2) \leq \sigma^2$, and $\mathbb{E}(\xi_i\xi_j) = 0$ for $i \neq j$, where $i, j = 1, \ldots, n$.

In order to control this bias term and to analyze the rate of convergence of projection estimator, we need to impose some assumptions on the underlying function f and on the basis $\{\varphi_j\}_{j=1}^{\infty}$.

3.1.1 Projection estimator with trigonometric basis in $L_2[0,1]$

Here we continue to consider the nonparametric regression model (48) and we will assume that $\mathfrak{X} = [0,1]$. We will mainly focus on a particular case, $X_i = i/n$.

Definition 3.3. The *trigonometric basis* is the orthonormal basis of $L_2[0,1]$ defined by

$$\varphi_1(x) = 1, \quad \varphi_{2k}(x) = \sqrt{2}\cos(2\pi kx), \quad \varphi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx), \quad k = 1, 2 \ldots,$$

for $x \in [0,1]$.

We will assume that the regression function f is sufficiently smooth, or more specifically, that it belongs to a periodic Sobolev class of functions. First, we define the periodic Sobolev class for integer smoothness β.

Definition 3.4. Let $\beta \geq 1$ be an integer and let $L > 0$. The periodic Sobelev class $W(\beta, L)$ is defined as

$$\begin{aligned} W(\beta, L) \quad := \quad & \Big\{ f : [0,1] \to \mathbb{R} : f^{(\beta-1)} \text{ is absolutely continuous and} \\ & \int_0^1 (f^{(\beta)}(x))^2 dx \leq L^2, f^{(j)}(0) = f^{(j)}(1), j = 0, 1, \ldots, \beta - 1 \Big\} \end{aligned}$$

Any function f belonging to such a class is continuous and periodic ($f(0) = f(1)$) and thus admits the representation

$$f(x) = \theta_1 \varphi_1(x) + \sum_{k=1}^{\infty} (\theta_{2k}\varphi_{2k}(x) + \theta_{2k+1}\varphi_{2k+1}(x)) \tag{21}$$

where $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis given in Definition 3.3. The above infinite series converges pointwise, and the sequence $\theta = \{\theta_j\}_{j=1}^{\infty}$ of Fourier coefficients of f belongs to the space

$$\ell^2(\mathbb{N}) := \left\{ \theta : \sum_{j=1}^{\infty} \theta_j^2 < \infty \right\}.$$

We now state a necessary and sufficient condition on θ under which the function (21) belongs to the class $W(\beta, L)$. Define

$$a_j = \begin{cases} j^{\beta}, & \text{for even } j, \\ (j-1)^{\beta}, & \text{for odd } j. \end{cases} \tag{22}$$

Proposition 3.5. Let $\beta \in \{1, 2, \ldots\}, L > 0$, and let $\{\varphi_j\}_{j=1}^{\infty}$ be the trigonometric basis. A function $f \in L_2[0,1]$ belong to $W(\beta, L)$ if and only if the vector θ of the Fourier coefficients of f belongs to the following ellipsoid in $\ell^2(\mathbb{N})$:

$$\Theta(\beta, Q) := \left\{ \theta \in \ell^2(\mathbb{N}) : \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq Q \right\} \tag{23}$$

where $Q = L^2/\pi^{2\beta}$ and a_j is given by (22).

See [14, Lemma A.3] for a proof of the above result.

The set $\Theta(\beta, Q)$ defined by (23) with $\beta > 0$ (not necessarily an integer), $Q > 0$, and a_j satisfying (22) is called a *Sobolev ellipsoid*. We mention the following properties of these ellipsoids.

- The monotonicity with respect to inclusion:

$$0 < \beta' \leq \beta \text{ implies } \Theta(\beta, Q) \subset \Theta(\beta', Q).$$

- If $\beta > 1/2$, the function $f = \sum_{j=1}^{\infty} \theta_j \varphi_j$ with the trigonometric basis $\{\varphi_j\}_{j=1}^{\infty}$ and $\theta \in \Theta(\beta, Q)$ is continuous (check this as an exercise). In what follows, we will basically consider this case.

The ellipsoid $\Theta(\beta, Q)$ is well-defined for all $\beta > 0$. In this sense $\Theta(\beta, Q)$ is a more general object than the periodic Sobolev class $W(\beta, L)$, where β can only be an integer. Proposition 3.5 establishes an isomorphism between $\Theta(\beta, Q)$ and $W(\beta, L)$ for integer β. It can be extended to all $\beta > 0$ by generalizing the definition of $W(\beta, L)$ in the following way.

Definition 3.6. For any $\beta > 0$ and $L > 0$ the Sobolev class $W(\beta, L)$ is defined as:

$$W(\beta, L) = \left\{ f \in L_2[0,1] : \theta = \{\theta_j\}_{j=1}^{\infty} \in \Theta(\beta, Q) \right\}$$

where $\theta_j = \int_0^1 f \varphi_j$ and $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis. Here $\Theta(\beta, Q)$ is the Sobolev ellipsoid defined in (23), where $Q = L^2/\pi^{2\beta}$ and $\{a_j\}_{j=1}^{\infty}$ is given by (22).

For all $\beta > 1/2$, the functions belonging to $W(\beta, L)$ are continuous. On the contrary, they are not always continuous for $\beta < 1/2$; an example is given by the function $f(x) = \text{sign}(x - 1/2)$, whose Fourier coefficients θ_j are of order $1/j$.

Lemma 3.7. Let $\{\varphi_j\}_{j=1}^{\infty}$ be the trigonometric basis. Then,

$$\frac{1}{n} \sum_{s=1}^{n} \varphi_j(s/n) \varphi_k(s/n) = \delta_{jk}, \qquad 1 \leq j, k \leq n-1, \tag{24}$$

where δ_{jk} is the Kronecker delta.

See [14, Lemma 1.7] for a proof of the above result.

We are now ready to establish an upper bound on the bias of the projection estimator.

Proposition 3.8. Let $f \in W(\beta, L), \beta \geq 1, L > 0$. Assume that $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis and $X_i = i/n$, $i = 1, \ldots, n$. Then, for all $n \geq 1, N \geq 1$,

$$\inf_{\theta \in \mathbb{R}^N} \mathbb{E}\|\mathbf{f}_{\boldsymbol{\theta}} - \mathbf{f}\|^2 \leq C(\beta, L) \left(\frac{1}{N^{2\beta}} + \frac{1}{n} \right),$$

where $C(\beta, L)$ is a constant that depends only on β and L.

The proof of the above result was given in class.

Theorem 3.9. Let $f \in W(\beta, L), \beta \geq 1, L > 0$ and $N = \lceil \alpha n^{1/(2\beta+1)} \rceil$ for $\alpha > 0$. Assume that $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis and $X_i = i/n$, $i = 1, \ldots, n$. Let ξ_i be random variables such that $\mathbb{E}(\xi_i) = 0$, $\mathbb{E}(\xi_i^2) \leq \sigma^2$ and $\mathbb{E}(\xi_i \xi_j) = 0$ for $i \neq j \in \{1, \ldots, n\}$. Then, for all $n \geq 1$,

$$\sup_{f \in W(\beta, L)} \mathbb{E}\|\hat{\mathbf{f}}^{LS} - \mathbf{f}\|^2 \leq C n^{-\frac{2\beta}{2\beta+1}},$$

where C is a constant that depends only on σ^2, β, L and α.

Proof. In view of (20) and Proposition 3.8,

$$\mathbb{E}\|\hat{\mathbf{f}}^{LS} - \mathbf{f}\|^2 \leq C(\beta, L) \left(\frac{1}{N^{2\beta}} + \frac{1}{n} \right) + \frac{\sigma^2 N}{n} = O(n^{-\frac{2\beta}{2\beta+1}}).$$

□

4 Minimax lower bounds

We have a family $\{P_\theta : \theta \in \Theta\}$ of probability measures, indexed by Θ, on a measurable space $(\mathcal{X}, \mathcal{A})$ associated with the data. Usually, in nonparametric statistics, Θ is a nonparametric class of functions (e.g., $\Theta = \Sigma(\beta, L)$ or $\Theta = W(\beta, L)$). Thus, in the density estimation model, P_θ is the probability measure associated with a sample $\mathbf{X} = (X_1, \ldots, X_n)$ of size n when the density of X_i is $p(\cdot) \equiv \theta$.

Given a semi-distance[5] the performance of an estimator $\hat{\theta}_n$ of θ is measured by the *maximum risk* of this estimator on Θ:

$$r(\hat{\theta}_n) := \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[d^2(\hat{\theta}_n, \theta) \right].$$

The aim of this section is to complement the upper bound results (i.e., $\sup_{\theta \in \Theta} \mathbb{E}_\theta \left[d^2(\hat{\theta}_n, \theta) \right] \leq C\psi_n^2$ for certain estimator $\hat{\theta}_n$) by the corresponding lower bounds:

$$\forall\ \hat{\theta}_n : \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[d^2(\hat{\theta}_n, \theta) \right] \geq c\psi_n^2$$

for sufficiently large n, where c is a positive constant.

The *minimax risk* associated with a statistical model $\{P_\theta : \theta \in \Theta\}$ and with a semi-distance d is defined as

$$\mathcal{R}_n^* := \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[d^2(\hat{\theta}_n, \theta) \right],$$

where the infimum is over all estimators. The upper bounds established previously imply that there exists a constant $C < \infty$ such that

$$\limsup_{n \to \infty} \psi_n^{-2} \mathcal{R}_n^* \leq C$$

for a sequence $\{\psi_n\}_{n \geq 1}$ converging to zero. The corresponding lower bounds claim that there exists a constant $c > 0$ such that, for the same sequence $\{\psi_n\}_{n \geq 1}$,

$$\liminf_{n \to \infty} \psi_n^{-2} \mathcal{R}_n^* \geq c. \tag{25}$$

4.1 Distances between probability measures

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let P and Q be two probability measures on $(\mathcal{X}, \mathcal{A})$. Suppose that ν is a σ-finite measure on $(\mathcal{X}, \mathcal{A})$ satisfying $P \ll \nu$ and $Q \ll \nu$.

[5] We will call the semi-distance $d : \Theta \times \Theta \to [0, +\infty)$ on Θ as a function that satisfies $d(\theta, \theta') = d(\theta', \theta)$, $d(\theta, \theta'') \leq d(\theta, \theta') + d(\theta', \theta'')$, and $d(\theta, \theta) = 0$, where $\theta, \theta', \theta'' \in \Theta$. The following are a few common examples of d: $d(f, g) = |f(x_0) - g(x_0)|$ (for some fixed x_0), $d(f, g) = \|f - g\|_2$, etc.

Define $p = dP/d\nu, q = dQ/d\nu$. Observe that such a measure ν always exists since we can take, for example, $\nu = P + Q$.

Definition 4.1. The *Hellinger distance* between P and Q is defined as follows:

$$H^2(P,Q) := \int (\sqrt{p} - \sqrt{q})^2 d\nu = 2\left(1 - \int \sqrt{pq}\, d\nu\right).$$

Exercise (HW2): The following are some properties of the Hellinger distance:

1. $H(P,Q)$ does not depend on the choice of the dominating measure ν.
2. $H(P,Q)$ satisfies the axioms of distance.
3. $0 \le H^2(P,Q) \le 2$.
4. If P and Q are product measures, $P = \otimes_{i=1}^n P_i$, $Q = \otimes_{i=1}^n Q_i$, then

$$H^2(P,Q) = 2\left[1 - \prod_{i=1}^n \left(1 - \frac{H^2(P_i,Q_i)}{2}\right)\right].$$

Definition 4.2. The *total variation distance* between P and Q is defined as follows:

$$V(P,Q) := \sup_{A\in\mathcal{A}} |P(A) - Q(A)| = \sup_{A\in\mathcal{A}} \left|\int_A (p-q) d\nu\right|.$$

Note that $0 \le V(P,Q) \le 1$ and $V(P,Q)$ satisfies the axioms of distance.

Lemma 4.3 (Scheffé's theorem)**.**

$$V(P,Q) = \frac{1}{2}\int |p-q| d\nu = 1 - \int \min(p,q) d\nu.$$

Lemma 4.4 (Le Cam's inequalities)**.**

$$\begin{aligned} \int \min(p,q) d\nu &\ge \frac{1}{2}\left(\int \sqrt{pq}\, d\nu\right)^2 = \frac{1}{2}\left(1 - \frac{H^2(P,Q)}{2}\right)^2, \\ \frac{1}{2}H^2(P,Q) &\le V(P,Q) \le H(P,Q)\sqrt{1 - \frac{H^2(P,Q)}{2}}. \end{aligned}$$

Exercise (HW2): Prove the above two lemmas.

Definition 4.5. The *Kullback divergence* between P and Q is defined by:

$$K(P,Q) := \begin{cases} \int \log\left(\frac{p}{q}\right) p\, d\nu, & \text{if } P \ll Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

It can be shown that the above definition always makes sense if $P \ll Q$. Here are some properties of the Kullback divergence:

1. $K(P,Q) \geq 0$. Indeed, by Jensen's inequality,
$$\int_{pq>0} \log\left(\frac{p}{q}\right) p\, d\nu = -\int_{pq>0} p \log\left(\frac{q}{p}\right) d\nu \geq -\log\left(\int q d\nu\right) \geq 0.$$

2. $K(P,Q)$ is not a distance (for example, it is not symmetric).

3. [Show this (Exercise (HW2))] If P and Q are product measures, $P = \otimes_{i=1}^n P_i$, $Q = \otimes_{i=1}^n Q_i$, then
$$K(P,Q) = \sum_{i=1}^n K(P_i, Q_i).$$

The next lemma links the Hellinger distance with the Kullback divergence.

Lemma 4.6.
$$H^2(P,Q) \leq K(P,Q).$$

The following lemma links the total variation distance with the Kullback divergence.

Lemma 4.7 (Pinsker's inequality)**.**
$$V(P,Q) \leq \sqrt{K(P,Q)/2}.$$

Exercise (HW2): Prove the above two lemmas.

Definition 4.8. The χ^2 *divergence* between P and Q is defined by:
$$\chi^2(P,Q) := \begin{cases} \int \frac{(p-q)^2}{p} d\nu, & \text{if } P \ll Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

Lemma 4.9.
$$\int \min(p,q) d\nu = 1 - \frac{1}{2}\int |p-q| d\nu \geq 1 - \frac{1}{2}\sqrt{\chi^2(P,Q)}.$$

Proof. Since p and q are probability densities,
$$2 = \int p\, d\nu + \int q\, d\nu = 2\int \min(p,q) d\nu + \int |p-q| d\nu$$
which shows the first equality. To show the inequality, we use Cauchy-Schwarz inequality to obtain
$$\int |p-q| d\nu = \int \frac{1}{\sqrt{p}} |p-q| \sqrt{p}\, d\nu \leq \chi^2(P,Q).$$

□

Lemma 4.10. If P and Q are product measures, $P = \otimes_{i=1}^n P_i$ and $Q = \otimes_{i=1}^n Q_i$, then

$$\chi^2(P, Q) = \prod_{i=1}^n (\chi^2(P_i, Q_i) + 1) - 1.$$

The proof is left as an exercise (HW2).

4.2 Lower Bounds on the risk of density estimators at a point

Our aim is to obtain a lower bound for the minimax risk on (Θ, d) where Θ is a Sobolev density:

$$\Theta = \mathcal{P}(\beta, L), \quad \beta > 0, L > 0,$$

and where d is a distance at a fixed point $x_0 \in \mathbb{R}$:

$$d(f, g) = |f(x_0) - g(x_0)|.$$

The rate that we would like to obtain is $\psi_n = n^{-\frac{\beta}{2\beta+1}}$. Indeed, this is the same rate as in the upper bounds which will enable us to conclude that ψ_n is optimal on (Θ, d).

Thus, we want to show that

$$\inf_{T_n} \sup_{p \in \mathcal{P}(\beta, L)} \mathbb{E}_p \left[(T_n(x_0) - p(x_0))^2 \right] \geq c n^{-\frac{2\beta}{2\beta+1}}, \tag{26}$$

for all n sufficiently large, where T_n ranges over all density estimators and $c > 0$ is a constant. For brevity we write $T_n = T_n(x_0)$. For any $p_0, p_1 \in \mathcal{P}(\beta, L)$, we may write

$$\begin{aligned} \sup_{p \in \mathcal{P}(\beta, L)} \mathbb{E}_p[(T_n - p(x_0))^2] &\geq \max \left\{ \mathbb{E}_{p_0}[(T_n - p_0(x_0))^2], \mathbb{E}_{p_1}[(T_n - p_1(x_0))^2] \right\} \\ &\geq \frac{1}{2} \left\{ \mathbb{E}_{p_0}[(T_n - p_0(x_0))^2] + \mathbb{E}_{p_1}[(T_n - p_1(x_0))^2] \right\}. \end{aligned} \tag{27}$$

Note that

$$\mathbb{E}_p[(T_n - p(x_0))^2] = \int \dots \int [T_n(x_1, \dots, x_n) - p(x_0)]^2 \left(\prod_{i=1}^n p(x_i) dx_i \right).$$

Let $\mathbf{x} := (x_1, \dots, x_n)$ and $\pi_n(\mathbf{x}) = \prod_{i=1}^n p(x_i)$. Also, let $\pi_{0,n}$, $\pi_{1,n}$ be the joint densities corresponding to the chosen densities p_0 and p_1. The expression in (27) is then equal to

$$\begin{aligned} & \frac{1}{2} \left[\int (T_n(\mathbf{x}) - p_0(x_0))^2 \pi_{0,n}(\mathbf{x}) d\mathbf{x} + \int (T_n(\mathbf{x}) - p_1(x_0))^2 \pi_{1,n}(\mathbf{x}) d\mathbf{x} \right] \\ \geq & \frac{1}{2} \left[\int (T_n(\mathbf{x}) - p_0(x_0))^2 + (T_n(\mathbf{x}) - p_1(x_0))^2 \right] \min\{\pi_{0,n}(\mathbf{x}), \pi_{1,n}(\mathbf{x})\} \, d\mathbf{x} \\ \geq & \frac{1}{4} (p_0(x_0) - p_1(x_0))^2 \int \min\{\pi_{0,n}(\mathbf{x}), \pi_{1,n}(\mathbf{x})\} \, d\mathbf{x}, \end{aligned}$$

where we have used the fact that $u^2 + v^2 \geq (u-v)^2/2$, for $u, v \in \mathbb{R}$.

In view of the above, to prove (26) it suffices to find densities p_0 and p_1 such that

(i) $p_0, p_1 \in \mathcal{P}(\beta, L)$,

(ii) $|p_0(x_0) - p_1(x_0)| \geq c_1 n^{-\frac{\beta}{2\beta+1}}$,

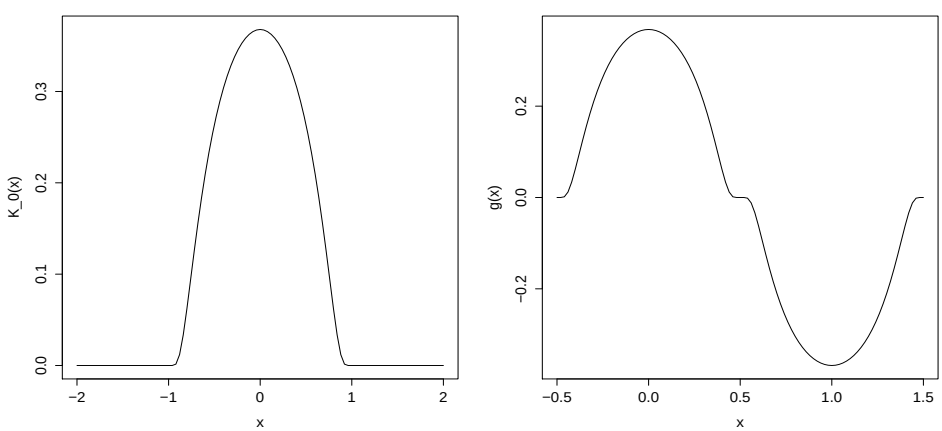
(iii) $\int \min\{\pi_{0,n}(\mathbf{x}), \pi_{1,n}(\mathbf{x})\}\, d\mathbf{x} \geq c_2$, where the constant c_2 does not depend on n.

We take p_0 to be a density on $\mathbb{R}$ such that $p_0 \in \Sigma(\beta, L/2)$ and $p_0(x_0) > 0$; e.g., p_0 can be the $N(0, \sigma^2)$ density with σ^2 chosen is such a suitable way. Obviously $p_0 \in \Sigma(\beta, L)$. Construct p_1 by adding a small perturbation to p_0:

$$p_1(x) := p_0(x) + h^\beta K\left(\frac{x - x_0}{h}\right),$$

where $h = \alpha n^{-1/(2\beta+1)}$ (for $\alpha > 0$), the support of K is $[-\frac{1}{2}, \frac{3}{2}]$, K is infinitely differentiable (i.e., $K \in C^\infty(\mathbb{R})$), $K(0) > 0$ and $\int K(u)du = 0$. Thus, p_1 is a density for h small enough.

Figure 2: Graphs of K_0 and g.



Lemma 4.11. $p_1 \in \Sigma(\beta, L)$ is a density for $h > 0$ small enough.

Proof. Let

$$K_0(u) := e^{-\frac{1}{1-u^2}} I_{[-1,1]}(u).$$

Then, $K_0 \in C^\infty(\mathbb{R})$ and the support of K_0 is $[-1, 1]$. Let $g : [-\frac{1}{2}, \frac{3}{2}] \to \mathbb{R}$ be defined as

$$g(u) := K_0(2u) - K_0(2(u-1)).$$

Observe that

1. $g(0) \neq 0$,
2. $\int g(u)du = 0$,
3. $g \in C^\infty(\mathbb{R})$, which implies that $g \in \Sigma(\beta, L')$ for a certain $L' > 0$.

Define $K : [-\frac{1}{2}, \frac{3}{2}] \to \mathbb{R}$ such that $K(u) := ag(u)$ for $a > 0$ small enough so that $K \in \Sigma(\beta, L/2)$.

Using the fact that $\int g(u)du = 0$ it is easy to see that $\int p_1(x)dx = 1$. Next we show that $p_1 \geq 0$ for $h > 0$ small enough. For $x \in [x_0 - \frac{h}{2}, x_0 + \frac{3h}{2}]$,

$$\begin{aligned} p_1(x) &\geq \min_{t \in [x_0 - \frac{h}{2}, x_0 + \frac{3h}{2}]} p_0(t) - \sup_{t \in [x_0 - \frac{h}{2}, x_0 + \frac{3h}{2}]} h^\beta \left| K\left(\frac{x - x_0}{h}\right)\right| \\ &\geq \min_{t \in [x_0 - \frac{h}{2}, x_0 + \frac{3h}{2}]} p_0(t) - h^\beta \sup_{t \in \mathbb{R}} |K(t)| . \end{aligned}$$

Since p_0 is continuous, $p_0(x_0) > 0$, we obtain that $p_1(x) > 0$ for all $x \in [x_0 - \frac{h}{2}, x_0 + \frac{3h}{2}]$, if h is smaller than some constant $h_0 > 0$. Note that for $x \notin [x_0 - \frac{h}{2}, x_0 + \frac{3h}{2}]$, $p_1(x) = p_0(x) \geq 0$. Thus, p_1 is a density.

We now have to show that $p_1 \in \Sigma(\beta, L)$. Set $\ell := \lfloor \beta \rfloor$. Clearly, p_1 is ℓ times differentiable. Further,

$$p_1^{(\ell)}(x) = p_0^{(\ell)}(x) + h^{\beta - \ell} K^{(\ell)}\left(\frac{x - x_0}{h}\right).$$

Hence,

$$\begin{aligned} |p_1^{(\ell)}(x) - p_1^{(\ell)}(x')| &\leq |p_0^{(\ell)}(x) - p_0^{(\ell)}(x)| + h^{\beta - \ell}\left| K^{(\ell)}\left(\frac{x - x_0}{h}\right) - K^{(\ell)}\left(\frac{x' - x_0}{h}\right)\right| \\ &\leq \frac{L}{2}|x - x'|^{\beta - \ell} + \frac{L}{2} h^{\beta - \ell}\left|\frac{x - x'}{h}\right|^{\beta - \ell} \leq L|x - x'|^{\beta - \ell}, \end{aligned}$$

where we have used the fact that both $p_0, K \in \Sigma(\beta, L/2)$. □

Thus,

$$|p_0(x_0) - p_1(x_0)| = h^\beta K(0) = K(0) n^{-\frac{\beta}{2\beta + 1}}.$$

Next we will try to show that (iii) holds. In view of Lemma 4.9, it suffices to bound

$\chi^2(\pi_{0,n}, \pi_{1,n})$ from above by a constant strictly less than 1. First write $\chi^2(p_0, p_1)$ as

$$\begin{aligned}
\int \frac{(p_0 - p_1)^2}{p_0} &= \int_{x_0-h/2}^{x_0+3h/2} \left[\frac{h^{2\beta} K^2((x-x_0)/h)}{p_0(x)}\right] dx \\
&\leq \frac{h^{2\beta+1}}{\min_{x\in[x_0-h/2, x_0+3h/2]} p_0(x)} \int K^2(u)du \\
&\leq \frac{h^{2\beta+1}}{\min_{x\in[x_0-1/2, x_0+3/2]} p_0(x)} \int K^2(u)du
\end{aligned}$$

where we have assumed that $h \leq \alpha$ and $\alpha \leq 1$. Plugging the choice of h we obtain

$$\chi^2(p_0, p_1) \leq c_* \alpha^{2\beta+1} n^{-1}$$

where the constant c_* depends only on p_0 and K. Therefore, applying Lemma 4.10 we find

$$\chi^2(\pi_{0,n}, \pi_{1,n}) \leq (1 + c_* \alpha^{2\beta+1} n^{-1})^n - 1 \leq \exp(c_* \alpha^{2\beta+1}) - 1,$$

where we have used the fact that $1 + v < e^v$, for $v \in \mathbb{R}$. Now, we choose α small enough so that $\exp(c_* \alpha^{2\beta+1}) - 1 < 1$. Then,

$$\int \min(\pi_0, \pi_1) \geq 1 - \frac{1}{2} = \frac{1}{2},$$

and thus, condition (iii) is satisfied.

Theorem 4.12. Let $\beta > 0, L > 0$. There exists a constant $c > 0$ that only depends on β and L such that, for all $x_0 \in \mathbb{R}$, $n \geq 1$,

$$\inf_{T_n} \sup_{p\in\mathcal{P}(\beta,L)} \mathbb{E}_p \left[(T_n(x_0) - p(x_0))^2\right] \geq c n^{-\frac{2\beta}{2\beta+1}},$$

where T_n ranges over all density estimators.

Since the choice of x_0 is arbitrary, we can equivalently put $\inf_{x_0\in\mathbb{R}}$before the minimax risk.

Definition 4.13. Let x_0 be fixed, and let $\mathcal{P}$ be a class of densities on $\mathbb{R}$. A sequence $\{\psi_n\}_{n\geq 1}, \psi_n > 0$, is called an *optimal rate of convergence* of mean squared error (risk) on the class $\mathcal{P}$ if the following two conditions are satisfied:

(i) $\inf_{T_n} \sup_{p\in\mathcal{P}} \mathbb{E}_p[(T_n(x_0) - p(x_0))^2] \geq c\psi_n^2$, where $c > 0$ is a constant independent of n.

(ii) There exist an estimator $p_n(\cdot)$, and a constant $C > 0$ independent of n such that

$$\sup_{p\in\mathcal{P}} \mathbb{E}_p[(p_n(x_0) - p(x_0))^2] \leq C\psi_n^2.$$

If (i) and (ii) hold, then p_n is called a *rate optimal estimator* for the risk on the class $\mathcal{P}$.

Corollary 4.14. Let $\beta > 0, L > 0$. The KDE with bandwidth $h = \alpha n^{-1/(2\beta+1)}$, $\alpha > 0$, and kernel of order $\ell = \lfloor \beta \rfloor$ is rate optimal for the mean squared error on the Hölder class $\mathcal{P}(\beta, L)$, and $\psi_n = n^{-\beta/(2\beta+1)}$ is the corresponding optimal rate.

Summary: We have seen that the following issues play the key role in nonparametric estimation.

- *Bias-variance trade-off*: For nonparametric estimation, the bias is not negligible, which brings in the problem of optimal choice of the smoothing parameter. For the KDE, the smoothing parameter is the bandwidth.
- *Optimality in a minimax sense*: Is the upper bound obtained from bias-variance trade-off indeed optimal? We need minimax lower bounds to answer this question.
- *Adaptation*: What is the optimal data-driven choice of the smoothing parameter? An adaptive estimator is an estimator which is rate optimal on a large scale of classes without any knowledge about the parameters of the classes. Cross-validation is an example of a successful adaptation procedure.

4.3 Lower bounds on many hypotheses

The lower bounds based on two hypotheses turn out to be inconvenient when we deal with estimation in L_p distances; see e.g., the start of Section 2.6 of [14].

Let us consider the nonparametric density estimation problem under the L_2 risk. Then,

$$d(f, g) = \|f - g\|_2 = \left(\int (f(x) - g(x))^2 \, dx \right)^{1/2}.$$

Our aim is to prove an optimal lower bound on the minimax risk for the Sobolev class of densities $\Theta = \mathcal{S}(\beta, L)$ (where $\beta \geq 1$ is an integer and $L > 0$) and the above L_2 distance with the rate $\psi_n = n^{-\beta/(2\beta+1)}$.

The proof is based on a construction of subsets $\mathcal{F}_n \subset \mathcal{S}(\beta, L)$, consisting of 2^{r_n} functions, where $r_n = \lfloor n^{1/(2\beta+1)} \rfloor$, and on bounding the supremum over $\mathcal{S}(\beta, L)$ by the average over $\mathcal{F}_n$.

The subset $\mathcal{F}_n$ is indexed by the set of all vectors $\theta \in \{0, 1\}^{r_n}$ consisting of sequences

of r_n zeros and ones. For $h = n^{-1/(2\beta+1)}$, let $x_{n,1} < x_{n,2} < \ldots < x_{n,n}$ be a regular grid of mesh width $2h$ (i.e., $x_{n,i} - x_{n,i-1} > 2h$, for $i = 2, \ldots, n$).

For a fixed probability density $p \in \mathcal{S}(\beta, L/2)$ (e.g., let p be the density of $N(0, \sigma^2)$ where σ^2 is such that $p \in \mathcal{S}(\beta, L/2)$). Consider a fixed function $K \in \mathcal{S}(\beta, L')$ with support $(-1, 1)$, and define, for every $\theta \in \{0,1\}^{r_n}$,

$$p_{n,\theta}(x) := p(x) + h^\beta \sum_{j=1}^{r_n} \theta_j K\left(\frac{x - x_{n,j}}{h}\right). \tag{28}$$

If p is bounded away from zero on a interval containing the grid, $|K|$ is bounded, and $\int K(x)dx = 0$, then $p_{n,\theta}$ is a p.d.f, at least for large n. Furthermore,

$$\int \left|p_{n,\theta}^{(\beta)}(x)\right|^2 dx \le 2\int \left|p^{(\beta)}(x)\right|^2 dx + 2hr_n \int \left|K^{(\beta)}(x)\right|^2 dx \le 2\frac{L^2}{4} + 2\frac{L^2}{4} \le L^2.$$

Observe that in the above we have used the fact that the mesh width is more than $2h$ so that for $j \neq k$,

$$\int K\left(\frac{x - x_{n,j}}{h}\right) K\left(\frac{x - x_{n,k}}{h}\right) dx = 0.$$

Thus, $p_{n,\theta} \in \mathcal{S}(\beta, L)$ for every θ.

Of course, there exists many choices of p and K such that $p_{n,\theta} \in \mathcal{S}(\beta, L)$ for every θ.

Theorem 4.15. There exists a constant $c_{\beta,L}$ such that for any density estimator $\hat{p}_n$,

$$\sup_{p \in \mathcal{S}(\beta,L)} \mathbb{E}_p\left[\int (\hat{p}_n - p)^2\right] \ge c_{\beta,L} n^{-2\beta/(2\beta+1)}.$$

We will use the following result crucially to prove the above theorem.

4.3.1 Assouad's lemma

The following lemma gives a lower bound for the maximum risk over the parameter set $\{0,1\}^r$, in an abstract form, applicable to the problem of estimating an arbitrary quantity $\psi(\theta)$ belonging to a semi-metric space (with semi-distance d). Let

$$H(\theta, \theta') := \sum_{i=1}^{r} |\theta_i - \theta_i'|$$

be the *Hamming distance* on $\{0,1\}^r$, which counts the number of positions at which θ and θ' differ.

For two probability measures P and Q with densities p and q let $\|P \wedge Q\| := \int p \wedge q \, d\nu$. Before we state and prove Assouad's lemma we give a simple result which will be useful later.

Lemma 4.16 (Lemma from hypothesis testing)**.** Suppose that we are given two models P_{θ_0} and P_{θ_1} on a measurable space $(\mathcal{X}, \mathcal{A})$ with densities p_0 and p_1 with respect to a σ-finite measure ν. Consider testing the hypothesis

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_0 : \theta = \theta_1.$$

The power function π_ϕ of any test ϕ satisfies

$$\pi_\phi(\theta_1) - \pi_\phi(\theta_0) \leq \frac{1}{2} \|P_{\theta_1} - P_{\theta_0}\|.$$

Proof. The difference on the left hand side can be written as $\int \phi(p_1 - p_0) d\nu$. The expression is maximized for the test function $I\{p_1 > p_0\}$ (Exercise (HW2): Show this). Thus,

$$\int \phi \, (p_1 - p_0) \, d\nu \leq \int_{p_1 > p_0} (p_1 - p_0) \, d\nu = \frac{1}{2} \int |p_1 - p_0| d\nu,$$

as

$$\begin{aligned} \int |p_1 - p_0| \, d\nu &= \int_{p_1 > p_0} (p_1 - p_0) \, d\nu + \int_{p_0 > p_1} (p_0 - p_1) \, d\nu \\ &= \int_{p_1 > p_0} (p_1 - p_0) \, d\nu + \left(\int (p_0 - p_1) \, d\nu - \int_{p_1 > p_0} (p_0 - p_1) \, d\nu \right) \\ &= 2 \int_{p_1 > p_0} (p_1 - p_0) \, d\nu. \end{aligned}$$

□

Lemma 4.17 (Assouad's lemma)**.** For any estimator T of $\psi(\theta)$ based on an observation in the experiment $\{P_\theta : \theta \in \{0,1\}^r\}$, and any $p > 0$,

$$\max_\theta 2^p \, \mathbb{E}_\theta[d^p(T, \psi(\theta))] \geq \min_{H(\theta,\theta') \geq 1} \frac{d^p(\psi(\theta), \psi(\theta'))}{H(\theta, \theta')} \, \frac{r}{2} \min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\|. \tag{29}$$

Proof. Define an estimator S, taking values in $\Theta = \{0,1\}^r$, by letting $S = \theta$ is $\theta' \mapsto d(T, \psi(\theta'))$ is minimal over Θ at $\theta' = \theta$ (if the minimum is not unique, choose a point of minimum in any consistent way), i.e.,

$$S = \operatorname*{argmin}_{\theta \in \Theta} d(T, \psi(\theta)).$$

By the triangle inequality, for any θ,

$$d(\psi(S), \psi(\theta)) \leq d(\psi(S), T) + d(\psi(\theta), T),$$

which is (upper) bounded by $2d(\psi(\theta), T)$, by the definition of S (as $d(\psi(S), T) \leq d(\psi(\theta), T)$). If

$$d^p(\psi(\theta), \psi(\theta')) \geq \gamma H(\theta, \theta') \tag{30}$$

for all pairs $\theta, \theta' \in \Theta$ (for some γ to be defined later), then

$$2^p \, \mathbb{E}_\theta[d^p(T, \psi(\theta))] \geq \mathbb{E}_\theta[d^p(\psi(S), \psi(\theta))] \geq \gamma \mathbb{E}_\theta[H(S, \theta)].$$

The maximum of this expression over Θ is bounded below by the average, which, apart from the factor γ, can be written as

$$\begin{aligned} \frac{1}{2^r} \sum_\theta \sum_{j=1}^r \mathbb{E}_\theta |S_j - \theta_j| &= \frac{1}{2} \sum_{j=1}^r \left(\frac{1}{2^{r-1}} \sum_{\theta : \theta_j = 0} \int S_j dP_\theta + \frac{1}{2^{r-1}} \sum_{\theta : \theta_j = 1} \int (1 - S_j) dP_\theta \right) \\ &= \frac{1}{2} \sum_{j=1}^r \left(\int S_j d\bar{P}_{0,j} + \int (1 - S_j) d\bar{P}_{1,j} \right), \end{aligned}$$

where

$$\bar{P}_{0,j} = \frac{1}{2^{r-1}} \sum_{\theta : \theta_j = 0} P_\theta \qquad \text{and} \qquad \bar{P}_{1,j} = \frac{1}{2^{r-1}} \sum_{\theta : \theta_j = 1} P_\theta.$$

This is minimized over S by choosing S_j for each j separately to minimize the j-th term in the sum. The expression within brackets is the sum of the error probabilities of a test of

$$H_0 : P = \bar{P}_{0,j} \qquad \text{versus} \qquad H_1 : P = \bar{P}_{1,j}.$$

Equivalently, it is equal to 1 minus the difference of power and level. By Lemma 4.16 it can be shown that this is at least $1 - \frac{1}{2} \|\bar{P}_{0,j} - \bar{P}_{1,j}\| = \|\bar{P}_{0,j} \wedge \bar{P}_{1,j}\|$ (by Lemma 4.9). Hence, the preceding display is bounded below by

$$\frac{1}{2} \sum_{j=1}^r \|\bar{P}_{0,j} \wedge \bar{P}_{1,j}\|.$$

Note that for two sequences $\{a_i\}_{i=1}^m$ and $\{b_i\}_{i=1}^m$,

$$\min \left(\frac{1}{m} \sum_{i=1}^m a_i, \frac{1}{m} \sum_{i=1}^m b_i \right) \geq \frac{1}{m} \sum_{i=1}^m \min(a_i, b_i) \geq \min_{i=1,\ldots,m} \min(a_i, b_i).$$

The 2^{r-1} terms P_θ and $P_{\theta'}$ in the averages $\bar{P}_{0,j}$ and $\bar{P}_{1,j}$ can be ordered and matched such that each pair θ and θ' differ only in their j-th coordinate. Conclude that

$$\frac{1}{2} \sum_{j=1}^r \|\bar{P}_{0,j} \wedge \bar{P}_{1,j}\| \geq \frac{1}{2} \sum_{j=1}^r \min_{H(\theta,\theta')=1, \theta_j \neq \theta'_j} \|P_\theta \wedge P_{\theta'}\| \geq \frac{r}{2} \min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\|.$$

Observing that the γ in (30) can always be taken as $\min_{H(\theta,\theta')\geq 1} \frac{d^p(\psi(\theta),\psi(\theta'))}{H(\theta,\theta')}$, we obtain the desired result. $\square$

Exercise (HW2): Complete the proof of Theorem 4.15.

4.3.2 Estimation of a monotone function

Consider data (X_i, Y_i), $i = 1, \ldots, n$, from a nonparametric regression model where

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{31}$$

with $0 \leq X_1 < X_2 < \ldots < X_n \leq 1$ being deterministic design points, $f : [0,1] \to \mathbb{R}$ is *nondecreasing* and the (unobserved) errors $\xi_1, \ldots, \xi_n$ are i.i.d. $N(0, \sigma^2)$. In what follows, we will also use the vector notation, writing the model as

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\xi},$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^\top, \mathbf{f} = (f(X_1), \ldots, f(X_n))^\top$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$. The goal of this section is to find the (optimal) lower bound on the rate of convergence for any estimator of f based on the loss

$$d^2(\mathbf{f}, \mathbf{g}) := \frac{1}{n}\|\mathbf{f} - \mathbf{g}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2 \tag{32}$$

where f, g are real valued functions defined on $[0,1]$.

For every $V > 0$ we define by $\mathcal{M}_V$ the class of nondecreasing functions $f : [0,1] \to R$ such that $f(1) - f(0) \leq V$.

Theorem 4.18. For any $V > 0$, there exists a constant $c_V > 0$, only depending on σ^2 and V, such that for any estimator $\hat{f}_n$, and for all $n \geq n_0 \in \mathbb{N}$,

$$\sup_{f \in \mathcal{M}_V} \mathbb{E}_f[d^2(\hat{\mathbf{f}}_n, \mathbf{f})] \geq c_V n^{-2/3},$$

where $\hat{\mathbf{f}}_n := (\hat{f}_n(X_1), \ldots, \hat{f}_n(X_n))^\top$.

Proof. We will use Assouad's lemma to prove the desired result. Fix an integer $1 \leq k \leq n$ (be chosen later) and let $r_n := \lfloor n/k \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x. Let us define $\mathbf{f} \in \mathbb{R}^n$ as

$$\mathbf{f}_i = \begin{cases} V\frac{(j-1)}{r_n}, & \text{if } (j-1)k < i \leq jk; \\ V\frac{(r_n-1)}{r_n}, & \text{if } r_n k < i \leq n. \end{cases}$$

Take f to be any nondecreasing function on $[0,1]$ such that $f(X_i) = \mathbf{f}_i$, for $i = 1, \ldots, n$. Also, it can be assumed that $f \in \mathcal{M}_V$. Let $\Theta = \{0,1\}^{r_n}$ and let $\psi(\theta) \in \mathbb{R}^n$, for $\theta \in \Theta$, be defined as:

$$\psi(\theta)_i = \mathbf{f}_i + \frac{V}{2r_n} \sum_{j=1}^{r_n} (2\theta_j - 1) I\{(j-1)k < i \leq jk\}. \tag{33}$$

Note that $\psi(\theta)$ induces a nondecreasing function that belongs to $\mathcal{M}_V$.

For $\theta, \theta' \in \Theta$, we have

$$\begin{aligned} d^2(\psi(\theta), \psi(\theta')) &= \frac{1}{n} \sum_{j=1}^{r_n} \sum_{(j-1)k < i \leq jk} [\psi(\theta)_i - \psi(\theta')_i]^2 \\ &= \frac{1}{n} \sum_{j=1}^{r_n} k |\theta_j - \theta'_j|^2 \frac{V^2}{r_n^2} = \frac{V^2 k}{r_n^2 n} H(\theta, \theta'). \end{aligned}$$

Therefore, this implies that for $\theta, \theta' \in \Theta$,

$$\min_{H(\theta,\theta') \geq 1} \frac{d^2(\psi(\theta), \psi(\theta'))}{H(\theta, \theta')} = \frac{V^2 k}{r_n^2 n}.$$

Further, by Pinsker's inequality (see Lemma 4.7), and using the fact that the Kullback-Leibler divergence $K(P_\theta, P_{\theta'})$ has a simple expression in terms of $d^2(\psi(\theta), \psi(\theta'))$ [Show this (Exercise (HW2))]:

$$V^2(P_\theta, P_{\theta'}) \leq \frac{1}{2} K(P_\theta, P_{\theta'}) = \frac{n}{4\sigma^2} d^2(\psi(\theta), \psi(\theta')) = \frac{V^2 k}{4\sigma^2 r_n^2} H(\theta, \theta').$$

Let $k := \lfloor n^{2/3} \left(\frac{\sigma}{V}\right)^{2/3} \rfloor$. As, $\int \min(p_\theta, p_{\theta'}) d\nu = 1 - V(P_\theta, P_{\theta'})$,

$$\min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\| \geq 1 - \frac{V\sqrt{k}}{2\sigma r_n} \geq c > 0,$$

for $c > 0$ and n sufficiently large (in fact c can be taken to be close to $1/2$). Therefore, using Assouad's lemma, we get the following lower bound:

$$\inf_{\hat{f}_n} \sup_{\theta \in \mathcal{M}_V} \mathbb{E}_f[d^2(\hat{\mathbf{f}}_n, \mathbf{f})] \geq \frac{V^2 k}{r_n^2 n} \frac{r_n}{8} c \geq c_V n^{-2/3},$$

where c_V is a constant that depends only on σ and V. □

4.4 A general reduction scheme

We can consider a more general framework where the goal is to find lower bounds of the following form:

$$\liminf_{n \to \infty} \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[w(\psi^{-1} d(\hat{\theta}_n, \theta)) \right] \geq c > 0,$$

where $w : [0,\infty) \to [0,\infty)$ is nondecreasing, $w(0) = 0$ and $w \neq 0$ (e.g., $w(u) = u^p, p > 0$). A general scheme for obtaining lower bounds is based on the following three remarks:

(a) *Reduction to bounds in probability.* For any $A > 0$ satisfying $w(A) > 0$ we have

$$\mathbb{E}_\theta \left[w(\psi_n^{-1} d(\hat{\theta}_n, \theta)) \right] \geq w(A)\, P_\theta \left[\psi_n^{-1} d(\hat{\theta}_n, \theta) \geq A \right]. \tag{34}$$

We will usually take $s \equiv s_n = A\psi_n$. Therefore, instead of searching for a lower bound on the minimax risk $\mathbb{R}_n^*$, it is sufficient to find a lower bound on the *minimax probabilities* of the form

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P_\theta \left(d(\hat{\theta}_n, \theta) \geq s \right)$$

where $s \equiv s_n = A\psi_n$.

(b) *Reduction to a finite number of hypotheses.* It is clear that

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P_\theta \left(d(\hat{\theta}_n, \theta) \geq s \right) \geq \inf_{\hat{\theta}_n} \max_{\theta \in \{\theta_0, \ldots, \theta_M\}} P_\theta \left(d(\hat{\theta}_n, \theta) \geq s \right) \tag{35}$$

for any finite set $\{\theta_0, \ldots, \theta_M\}$ contained in Θ. In the examples we have already seen that the finite set $\{\theta_0, \ldots, \theta_M\}$ has to be chosen appropriately. We call the $M+1$ elements $\theta_0, \ldots, \theta_M$ as *hypotheses*. We will call a *test* any $\mathcal{A}$-measurable function $\Psi : \mathcal{X} \to \{0, 1, \ldots, M\}$.

(c) *Choice of* $2s$*-separated hypotheses.* If

$$d(\theta_j, \theta_k) \geq 2s, \qquad k \neq j, \tag{36}$$

then for any estimator $\hat{\theta}_n$,

$$P_\theta \left(d(\hat{\theta}_n, \theta) \geq s \right) \geq P_\theta \left(\Psi^* \neq j \right), \qquad j = 0, 1, \ldots, M,$$

where $\Psi^* : \mathcal{X} \to \{0, 1, \ldots, M\}$ is the *minimum distance test* defined by

$$\Psi^* = \operatorname*{argmin}_{0 \leq k \leq M} d(\hat{\theta}_n, \theta_k).$$

Therefore,

$$\inf_{\hat{\theta}_n} \max_{\theta \in \{\theta_0, \ldots, \theta_M\}} P_\theta \left(d(\hat{\theta}_n, \theta) \geq s \right) \geq \inf_{\Psi} \max_{0 \leq j \leq M} P_j(\Psi \neq j) =: p_{e,M}, \tag{37}$$

where $P_j \equiv P_{\theta_j}$ and $\inf_\Psi$ denotes the infimum over all tests.

Thus, in order to obtain lower bounds it is sufficient to check that

$$p_{e,M} \geq c',$$

where the hypotheses θ_j satisfy (36) with $s = A\psi_n$ and where the constant $c' > 0$ is independent of n. The quantity $p_{e,M}$ is called the *minimum probability of error* for the problem of testing $M+1$ hypotheses $\theta_0, \theta_1, \ldots, \theta_M$.

Remark 4.1. Let $P_0, P_1, \ldots, P_M$ be probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. For a test $\Psi : \mathcal{X} \to \{0, 1, \ldots, M\}$, define the *average probability of error* and the minimum average probability of error by

$$\bar{p}_{e,M}(\Psi) := \frac{1}{M+1} \sum_{j=0}^{M} P_j(\Psi \neq j), \qquad \text{and} \qquad \bar{p}_{e,M} := \inf_{\Psi} \bar{p}_{e,M}(\Psi).$$

Note that as

$$p_{e,M} \geq \bar{p}_{e,M},$$

we can then use tools (from multiple hypotheses testing) to lower bound $\bar{p}_{e,M}$.

Example 4.19. Let $\Theta = [0, 1]$. Consider data $X_1, \ldots, X_n$ i.i.d. Bernoulli(θ), where $\theta \in \Theta$. Thus, here P_θ is the joint distribution of $\mathbf{X} = (X_1, \ldots, X_n)$. The goal is to find the minimax lower bound for the estimation of θ under the loss $d(\hat{\theta}_n, \theta) := |\hat{\theta}_n - \theta|$. We want to show that there exists $c > 0$ such that

$$\liminf_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta \left[n d^2(\hat{\theta}_n, \theta) \right] \geq c > 0.$$

Consider $M = 1$ and let $\theta_0 = \frac{1}{2} - s$ and $\theta_1 = \frac{1}{2} + s$, where $s \in [0, 1/4]$. Using Lemma 4.16 we can show that

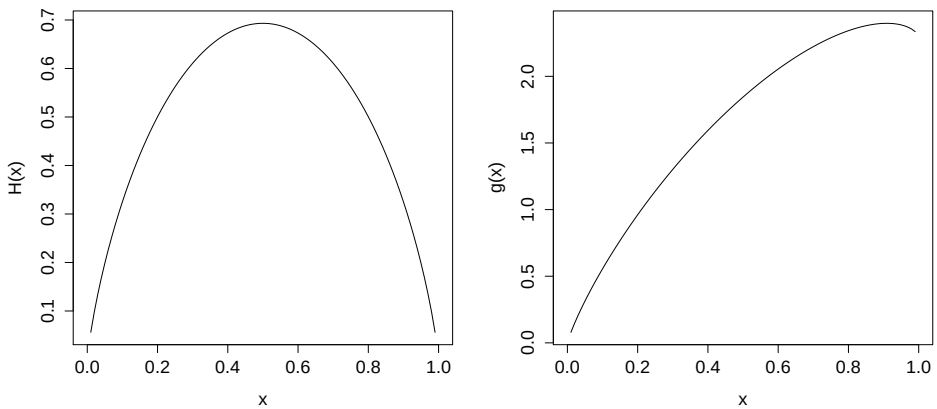
$$\inf_{\hat{\theta}_n} \max_{\theta\in\{\theta_0,\theta_1\}} P_\theta \left(d^2(\hat{\theta}_n, \theta) \geq s \right) \geq p_{e,M} \geq \bar{p}_{e,M} \geq 1 - V(P_0, P_1).$$

We can bound $V(P_0, P_1)$ using Pinsker's inequality (see Lemma 4.7) and then use Property (3) of the Kullback divergence to show that

$$V^2(P_0, P_1) \leq \frac{1}{2} K(P_0, P_1) \leq n K(\mathrm{Ber}(\theta_0), \mathrm{Ber}(\theta_1)) = 2s \log \left(\frac{1+2s}{1-2s} \right).$$

Using the fact that $x \log\left(\frac{1+x}{1-x}\right) \leq 3x^2$ for $x \in [0, \frac{1}{2}]$, we can now show the desired result for $c = \frac{1}{48}$.

Figure 3: Graphs of H and g with $M = 10$.



4.5 Fano's lemma

Lemma 4.20 (Fano's lemma)**.** Let $P_0, P_1, \ldots, P_M$ be probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$, $M \geq 1$. Then, $\bar{p}_{e,M} \leq M/(M+1)$, and

$$g(\bar{p}_{e,M}) \geq \log(1+M) - \frac{1}{M+1} \sum_{j=0}^{M} K(P_j, \overline{P}), \tag{38}$$

where

$$\overline{P} = \frac{1}{M+1} \sum_{j=0}^{M} P_j,$$

and, for $x \in [0,1]$,

$$g(x) = x \log M + H(x), \qquad H(x) = -x \log x - (1-x) \log(1-x).$$

Proof. We have

$$\bar{p}_{e,M}(\Psi) = \frac{1}{M+1} \mathbb{E}_{\overline{P}} \left[\sum_{j=0}^{M} I(A_j) \frac{dP_j}{d\overline{P}} \right] = \mathbb{E}_{\overline{P}} \left[\sum_{j=0}^{M} b_j p_j \right] \tag{39}$$

where

$$p_j := (M+1)^{-1} \frac{dP_j}{d\overline{P}}, \qquad A_j := \{\Psi \neq j\}, \qquad b_j = I(A_j)$$

and $\mathbb{E}_{\overline{P}}$ denotes the expectation with respect to $\overline{P}$. The random variables b_j and p_j satisfy $\overline{P}$-a.s. the following conditions:

$$\sum_{j=0}^{M} b_j = M, \qquad b_j \in \{0,1\}, \qquad \text{and} \qquad \sum_{j=0}^{M} p_j = 1, \qquad p_j \geq 0.$$

Then we have that, $\overline{P}$-a.s.,

$$\sum_{j=0}^{M} b_j p_j = \sum_{j \neq j_0} p_j, \tag{40}$$

where j_0 is a random number, $0 \leq j_0 \leq M$. We now apply the following lemma (see [14, Lemma 2.11] for a proof).

Lemma 4.21. For all $j_0 \in \{0, 1, \ldots, M\}$ and all real numbers $p_0, p_1, \ldots, p_M$, such that $\sum_{j=0}^{M} p_j = 1$, $p_j \geq 0$, we have

$$g\left(\sum_{j \neq j_0} p_j\right) \geq -\sum_{j=0}^{M} p_j \log p_j, \tag{41}$$

where $0 \log 0 := 0$.

Note that the function g is concave on $0 \leq x \leq 1$. Using (39), Jensen's inequality, and (40) and (41), we obtain that, for any test Ψ,

$$\begin{aligned} g(\bar{p}_{e,M}(\Psi)) &= g\left(\mathbb{E}_{\overline{P}}\left[\sum_{j=0}^{M} b_j p_j\right]\right) \geq \mathbb{E}_{\overline{P}}\left[g\left(\sum_{j=0}^{M} b_j p_j\right)\right] \\ &\geq \mathbb{E}_{\overline{P}}\left[-\sum_{j=0}^{M} p_j \log p_j\right] \\ &= \log(1+M) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \overline{P}). \end{aligned}$$

Since there exists a sequence of tests $\{\Psi_k\}_{k \geq 1}$ such that $\bar{p}_{e,M}(\Psi_k) \to \bar{p}_{e,M}$ as $k \to \infty$, we obtain, by the continuity of g,

$$g(\bar{p}_{e,M}) = \lim_{k \to \infty} g(\bar{p}_{e,M}(\Psi_k)) \geq \log(1+M) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \overline{P}).$$

It remains to show that $\bar{p}_{e,M} \leq M/(M+1)$. For this purpose, we define a degenerate test $\Psi^* \equiv 1$, and observe that

$$\inf_{\Psi} \bar{p}_{e,M}(\Psi) \leq \bar{p}_{e,M}(\Psi^*) = \frac{1}{M+1}\sum_{j=0}^{M} P_j(j \neq 1) = \frac{M}{M+1}.$$

□

Using Fano's lemma we can bound from below the minimax probability of error $p_{e,M}$ in the following way:

$$\begin{aligned} p_{e,M} &= \inf_{\Psi} \max_{0\le j\le M} P_j(\Psi \neq j) \ge \inf_{\Psi} \bar{p}_{e,M} \\ &\ge g^{-1}\left(\log(M+1) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \overline{P})\right), \end{aligned} \tag{42}$$

where $g^{-1}(t) := 0$ for $t < 0$, and for $0 < t < \log(M+1)$, $g^{-1}(t)$ is a solution of the equation $g(x) = t$ with respect to $x \in [0, M/(M+1)]$ — this solution exists as g is continuous and strictly increasing on $[0, M/(M+1)]$ and $g(0) = 0$, $g(M/(M+1)) = \log(M+1)$.

The following corollary gives a more workable lower bound on $p_{e,M}$.

Corollary 4.22. Let $P_0, P_1, \ldots, P_M$ be probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$, $M \ge 2$. Let

$$I(M) := \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \overline{P}). \tag{43}$$

Then,

$$p_{e,M} \ge \bar{p}_{e,M} \ge 1 - \frac{I(M) + \log 2}{\log(M+1)}. \tag{44}$$

Proof. As $H(x) \le \log 2$ for all $x \in [0,1]$, and $g(x) = x \log M + H(x)$, we have, from (38),

$$\bar{p}_{e,M} \log(M+1) \ge \bar{p}_{e,M} \log M \ge \log(M+1) - I(M) - \log 2$$

which yields the desired result. □

Determining $I(M)$ exactly is usually intractable however and one typically works with appropriate bounds on $I(M)$. In fact, (42) is going to be useful if we can show that $\log(M+1) - I(M) > 0$. The following corollary gives a sufficient condition for this and gives a non-trivial lower bound on $p_{e,M}$.

Corollary 4.23. Let $P_0, P_1, \ldots, P_M$ be probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$, $M \ge 2$. If

$$\frac{1}{M+1}\sum_{j=0}^{M} K(P_j, P_0) \le \alpha \log(M+1) \tag{45}$$

with $0 < \alpha < 1$, then

$$p_{e,M} \ge \bar{p}_{e,M} \ge 1 - \frac{\log 2}{\log(M+1)} - \alpha. \tag{46}$$

Proof. We will use the elementary fact (show this; Exercise (HW2)):

$$\frac{1}{M+1}\sum_{j=0}^{M} K(P_j, P_0) = \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \overline{P}) + K(\overline{P}, P_0). \tag{47}$$

Thus, using the above display, (38) and the fact that $K(\overline{P}, P_0) \geq 0$, we get

$$\begin{aligned} g(\bar{p}_{e,M}) &\geq \log(M+1) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \overline{P}), \\ &\geq \log(M+1) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, P_0) \\ &\geq \log(M+1) - \alpha \log(M+1). \end{aligned}$$

A similar calculation as in the proof of Corollary 4.22 now yields the desired result. □

4.5.1 Estimation of a regression function under the supremum loss

Consider data (X_i, Y_i), $i = 1, \ldots, n$, from a nonparametric regression model where

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{48}$$

with $f : [0,1] \to \mathbb{R}$, the ξ_i's being i.i.d. $N(0, \sigma^2)$, and the X_i's are arbitrary random variables taking values in $[0,1]$ such that $(X_1, \ldots, X_n)$ is independent of $(\xi_1, \ldots, \xi_n)$.

Theorem 4.24. Let $\beta > 0$ and $L > 0$. Consider data from the above model where $f \in \Sigma(\beta, L)$. Let

$$\psi_n = \left(\frac{\log n}{n}\right)^{\beta/(2\beta+1)}.$$

Then,

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f \in \Sigma(\beta,L)} \mathbb{E}_f[\psi_n^{-2} \|T_n - f\|_\infty^2] \geq c$$

where $\inf_{T_n}$ denotes the infimum over all estimators and where the constant $c > 0$ depends only on β, L and σ^2.

Proof. The proof was mostly done in class; also see [14, Theorem 2.11]. □

4.6 Covering and packing numbers and metric entropy

In Section 4.4 we described a general scheme for proving lower bounds. In step (c) of the scheme it is important to choose the hypotheses θ_j's in Θ such that they are

$2s$-separated. Further, the choice of the number of such points M depends on how large the space Θ is. In this section we define a concept that has been successfully employed in many fields of mathematics to capture the size of the an underlying set (with a semi-metric). We also give a few examples from parametric models to show how this concept can be used in conjunction with Fano's lemma (as discussed in the last section) to yield useful lower bounds that do not need specification of the exact θ_j's (the perturbation functions).

Let (Θ, d) be an arbitrary semi-metric space.

Definition 4.25 (Covering number)**.** A *δ-cover* of the set Θ with respect to the semi-metric d is a set $\{\theta_1, \ldots, \theta_N\} \subset \Theta$ such that for any point $\theta \in \Theta$, there exists some $v \in \{1, \ldots, N\}$ such that $d(\theta, \theta_v) < \delta$.

The δ-covering number of Θ is

$$N(\delta, \Theta, d) := \inf\{N \in \mathbb{N} : \ \exists \text{ a } \delta\text{-cover } \theta_1, \ldots, \theta_N \text{ of } \Theta\}.$$

Equivalently, the δ-covering number $N(\delta, \Theta, d)$ is the minimal number of balls $B(x; \delta) := \{y \in \Theta : d(x, y) < \delta\}$ of radius δ needed to cover the set Θ.

A semi-metric space (Θ, d) is said to be *totally bounded* if the δ-covering number is finite for every $\delta > 0$.

The *metric entropy* of the set Θ is the logarithm of its covering number: $\log N(\delta, \Theta, d)$.

We can define a related measure — more useful for constructing our lower bounds — of size that relates to the number of disjoint balls of radius $\delta > 0$ that can be placed into the set Θ.

Definition 4.26 (Packing number)**.** A *δ-packing* of the set Θ with respect to the semi-metric d is a set $\{\theta_1, \ldots, \theta_D\}$ such that for all distinct $v, v' \in \{1, \ldots, D\}$, we have $d(\theta_v, \theta_{v'}) \geq \delta$.

The δ-packing number of Θ is

$$D(\delta, \Theta, d) := \inf\{D \in \mathbb{N} : \ \exists \text{ a } \delta\text{-packing } \theta_1, \ldots, \theta_D \text{ of } \Theta\}.$$

Equivalently, call a collection of points *δ-separated* if the distance between each pair of points is larger than δ. Thus, the *packing number* $D(\delta, \Theta, d)$ is the maximum number of δ-separated points in Θ.

Exercise (HW2): Show that

$$D(2\delta, \Theta, d) \leq N(\delta, \Theta, d) \leq D(\delta, \Theta, d), \qquad \text{for every } \delta > 0.$$

Thus, packing and covering numbers have the same scaling in the radius δ.

Remark 4.2. As shown in the preceding exercise, covering and packing numbers are closely related, and we can use both in the following. Clearly, they become bigger as $\delta \to 0$.

We can now provide a few more complex examples of packing and covering numbers, presenting two standard results that will be useful for constructing the packing sets used in our lower bounds to come.

Our first bound shows that there are (exponentially) large packings of the d-dimensional hypercube of points that are $O(d)$-separated in the Hamming metric.

Lemma 4.27 (Varshamov-Gilbert Lemma)**.** Fix $k \geq 1$. There exists a subset $\mathcal{V}$ of $\{0,1\}^k$ with $|\mathcal{V}| \geq \exp(k/8)$ such that the Hamming distance, $H(\tau, \tau') := \sum_{i=1}^k I\{\tau_i \neq \tau_i'\} > k/4$ for all $\tau, \tau' \in \mathcal{V}$ with $\tau \neq \tau'$.

Proof. Consider a maximal subset $\mathcal{V}$ of $\{0,1\}^k$ that satisfies:

$$H(\tau, \tau') \geq k/4 \qquad \text{for all } \tau, \tau' \in \mathcal{V} \text{ with } \tau \neq \tau'. \tag{49}$$

The meaning of maximal here is that if one tries to expand $\mathcal{V}$ by adding one more element, then the constraint (49) will be violated. In other words, if we define the closed ball, $B(\tau, k/4) := \{\theta \in \{0,1\}^k : H(\theta, \tau) \leq k/4\}$ for $\tau \in \{0,1\}^k$, then we must have

$$\bigcup_{\tau \in \mathcal{V}} B(\tau, k/4) = \{0,1\}^k.$$

This implies that

$$\sum_{\tau \in \mathcal{V}} |B(\tau, k/4)| \geq 2^k. \tag{50}$$

Let $T_1, \ldots, T_k$ denote i.i.d. Bernoulli random variables with probability of success $1/2$. For every $A \subseteq \{0,1\}^k$, we have $\mathbb{P}\left((T_1, \ldots, T_k) \in A\right) = |A| 2^{-k}$. Therefore, for each $\tau \in \mathcal{V}$, we can write

$$2^{-k}|B(\tau, k/4)| = \mathbb{P}\Big((T_1, \ldots, T_k) \in B(\tau, k/4)\Big) = \mathbb{P}\left(\sum_{i=1}^k \{T_i \neq \tau_i\} \leq k/4\right).$$

If $S_i := \{T_i \neq \tau_i\}$, then it is easy to see that $S_1, \ldots, S_k$ are also i.i.d. Bernoulli random

variables with probability of success 1/2. Thus,

$$\begin{aligned}
2^{-k}|B(\tau, k/4)| &= \mathbb{P}\left(S_1 + \cdots + S_k \leq k/4\right) \\
&= \mathbb{P}\left(S_1 + \cdots + S_k \geq 3k/4\right) \\
&\leq \inf_{\lambda > 0} \exp(-3\lambda k/4) \left(\mathbb{E}\exp(\lambda S_1)\right)^k \\
&= \inf_{\lambda > 0} \exp(-3\lambda k/4) 2^{-k}(1 + e^{\lambda})^k.
\end{aligned}$$

Taking $\lambda = \log 3$, we get

$$|B(\tau, k/4)| \leq 3^{-3k/4} 4^k \qquad \text{for every } \tau \in \mathcal{V}.$$

Finally, from (50), we obtain

$$|\mathcal{V}| \geq \frac{3^{3k/4}}{2^k} = \exp\left(k \log(3^{3/4}/2)\right) \geq \exp\left(k/8\right).$$

□

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

Lemma 4.28. Let $\mathcal{B} := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$ denote the unit Euclidean ball in $\mathbb{R}^d$. Then

$$\left(\frac{1}{\delta}\right)^d \leq N(\delta, \mathcal{B}, \|\cdot\|_2) \leq \left(1 + \frac{2}{\delta}\right)^d. \tag{51}$$

As a consequence of Lemma 4.28, we see that for any $\delta < 1$, there is a packing $\mathcal{V}$ of $\mathcal{B}$ such that $\|\theta - \theta'\|_2 \geq \delta$ for all distinct $\theta, \theta' \in \mathcal{V}$ and $|\mathcal{V}| \geq (1/\delta)^d$, because we know $D(\delta, \mathcal{B}, \|\cdot\|_2) \geq N(\delta, \mathcal{B}, \|\cdot\|_2)$. In particular, the lemma shows that any norm ball has a 1/2-packing in its own norm with cardinality at least 2^d. We can also construct exponentially large packings of arbitrary norm-balls (in finite dimensions) where points are of constant distance apart.

Smoothly parameterized functions: Let $\mathcal{F}$ be a parameterized class of functions, i.e.,

$$\mathcal{F} := \{f_\theta : \theta \in \Theta\}.$$

Let $\|\cdot\|_\Theta$ be a norm on Θ, and let $\|\cdot\|_\mathcal{F}$ be a norm on $\mathcal{F}$. Suppose that the mapping $\theta \mapsto f_\theta$ is L-Lipschitz, i.e.,

$$\|f_\theta - f_{\theta'}\|_\mathcal{F} \leq L\|\theta - \theta'\|_\Theta.$$

Lemma 4.29 (Exercise (HW2))**.** $N(\delta, \mathcal{F}, \|\cdot\|_{\mathcal{F}}) \le N(\delta/L, \Theta, \|\cdot\|_{\Theta})$ for all $\delta > 0$.

A Lipschitz parameterization allows us to translates a cover of the parameter space Θ into a cover of the function space $\mathcal{F}$. For example, if $\mathcal{F}$ is smoothly parameterized by (compact set of) d parameters, then $N(\delta, \mathcal{F}, \|\cdot\|_{\mathcal{F}}) = O(\delta^{-d})$.

Exercise (HW2): Let $\mathcal{F}$ be the set of L-Lipschitz functions mapping from $[0,1]$ to $[0,1]$. Then in the supremum norm $\|f\|_\infty := \sup_{x\in[0,1]} |f(x)|$,

$$\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp L/\delta.$$

Hint: (Proof idea) Form an δ grid of the y-axis, and an δ/L grid of the x-axis, and consider all functions that are piecewise linear on this grid, where all pieces have slopes $+L$ or $-L$. There are $1/\delta$ starting points, and for each starting point there are $2^{L/\delta}$ slope choices. Show that this set is an $O(\delta)$ packing and an $O(\delta)$ cover.

4.6.1 Two examples

Example 4.30 (Normal mean estimation)**.** Consider the d-dimensional normal location family $\mathcal{N}_d := \{N(\theta, \sigma^2 I_d) : \theta \in \mathbb{R}^d\}$, where $\sigma^2 > 0$ and $d \ge 2$. We wish to estimate the mean θ in the squared error loss, i.e., $d^2(\hat{\theta}_n, \theta) = \|\hat{\theta}_n - \theta\|_2^2$, given n i.i.d. observations $X_1, \ldots, X_n$ from a member in $\mathcal{N}_d$ with mean θ. Let P_θ denote the joint distribution of the data.

Let $\mathcal{V}$ be a 1/2-packing of the unit $\|\cdot\|_2$-ball with cardinality at least 2^d, as guaranteed by Lemma 4.28. Now we construct our local packing. Fix $\delta > 0$, and for each $v \in \mathcal{V}$, set $\theta_v = \delta v \in \mathbb{R}^d$. Then we have

$$\|\theta_v - \theta_{v'}\|_2 = \delta\|v - v'\|_2 \ge \frac{\delta}{2} =: 2s$$

for each distinct pair $v, v' \in \mathcal{V}$, and moreover, we note that $\|\theta_v - \theta_{v'}\|_2 \le 2\delta$ for such pairs as well. Thus, $\{\theta_v\}_{v\in\mathcal{V}}$ is a $2s$-separated set with cardinality at least 2^d. Let $\theta_{v_0}, \theta_{v_1}, \ldots, \theta_{v_M}$ be an enumeration of the $2s$-separated points, and we take $P_j \equiv P_{\theta_{v_j}}$, for $j = 0, 1, \ldots, M$. Note that for $j \in \{0, \ldots, M\}$ such that $P_j \equiv P_v$, for some $v \in \mathcal{V}$,

$$K(P_j, P_0) = \frac{n}{2\sigma^2}\|\theta_v - \theta_{v_0}\|_2^2 \le \frac{2n\delta^2}{\sigma^2}.$$

Therefore, taking $\delta^2 := d\sigma^2 \log 2/(8n)$,

$$\frac{1}{M+1}\sum_{j=0}^{M} K(P_j, P_0) \le \frac{2n\delta^2}{\sigma^2 d \log 2} \cdot d \log 2 \le \alpha \log(M+1)$$

where $\alpha := 1/4$. This shows that (45) holds. Hence, by (34), (35), (37) and Corollary 4.23 we have

$$\begin{aligned}
\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta[d^2(\hat{\theta}_n, \theta)] &\geq \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} s^2 P_\theta[d(\hat{\theta}_n, \theta) \geq s] \\
&\geq s^2 \left(1 - \frac{\log 2}{\log(M+1)} - \alpha\right) \\
&\geq \frac{1}{4^2} \cdot \frac{d\sigma^2 \log 2}{8n} \left(1 - \frac{1}{d} - \frac{1}{4}\right).
\end{aligned}$$

As $d \geq 2$, the above inequality implies the minimax lower bound

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta[d^2(\hat{\theta}_n, \theta)] \geq \frac{1}{64} \cdot \frac{d\sigma^2 \log 2}{8n} = c\frac{d\sigma^2}{n},$$

where $c > 0$. While the constant c is not sharp, we do obtain the right scaling in d, n and the variance σ^2. The sample mean attains the same risk.

Example 4.31 (Linear regression)**.** In this example, we show how local packings can give (up to some constant factors) sharp minimax rates for standard linear regression problems. In particular, for fixed matrix $X \in \mathbb{R}^{n \times d}$, we observe

$$Y = X\theta + \epsilon,$$

where $\epsilon \in \mathbb{R}^n$ consists of independent random variables ϵ_i with variance bounded by $\mathrm{Var}(\epsilon_i) \leq \sigma^2$, and $\theta \in \mathbb{R}^d$ is allowed to vary over $\mathbb{R}^d$. For the purposes of our lower bound, we may assume that $\epsilon \sim N(0, \sigma^2 I_n)$. Let $\mathcal{P} := \{N(X\theta, \sigma^2 I_n) : \theta \in \mathbb{R}^d\}$ denote the family of such normally distributed linear regression problems, and assume for simplicity that $d \geq 32$.

In this case, we use the Varshamov-Gilbert bound (Lemma 4.27) to construct a local packing and attain minimax rates. Indeed, let $\mathcal{V}$ be a packing of $\{0,1\}^d$ such that $\|v - v'\|_1 \geq d/4$ for distinct elements of $\mathcal{V}$, and let $|\mathcal{V}| \geq \exp(d/8)$ as guaranteed by the Varshamov-Gilbert bound. For fixed $\delta > 0$, if we set $\theta_v = \delta v$, then we have the packing guarantee for distinct elements v, v' that

$$\|\theta_v - \theta_{v'}\|_2^2 = \delta^2 \|v - v'\|_2^2 = \delta^2 \|v - v'\|_1 \geq d\delta^2/4.$$

Moreover, we have the upper bound

$$K(P_{\theta_v}, P_{\theta_{v'}}) = \frac{1}{2\sigma^2} \|X(\theta_v - \theta_{v'})\|_2^2 \leq \frac{\delta^2}{2\sigma^2} \Lambda_{max}(X^\top X) \|\theta_v - \theta_{v'}\|_2^2 \leq \frac{d\delta^2}{2\sigma^2} \Lambda_{max}(X^\top X),$$

where $\Lambda_{max}(X^\top X)$ denotes the maximum singular value of $X^\top X$.

Consequently, taking $\delta^2 := \frac{\sigma^2}{16\Lambda_{max}(X^\top X)}$, we obtain that

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta[d^2(\hat{\theta}_n, \theta)] \geq \frac{d\sigma^2}{(16)^2 n \Lambda_{max}(X^\top X/n)} \left(1 - \frac{8}{d}\log 2 - \frac{1}{4}\right) \geq c\frac{d\sigma^2}{n\Lambda_{max}(X^\top X/n)},$$

for some $c > 0$ if $d > (32/3)\log 2$. Thus, the convergence rate is (roughly) $\sigma^2 d/n$ after rescaling the singular values of $X^\top X$ by $n^{-1/2}$. This bound is sharp in terms of the dimension d, dependence on n, and the variance σ^2, but it does not fully capture the dependence on $X^\top X$, as it depends only on the maximum singular value. An exact calculation can show that the minimax value of the problem is exactly $\sigma^2 \mathrm{tr}((X^\top X)^{-1})$.

4.7 Global Fano method: Bounding $I(M)$ based on metric entropy

Observe that, from (47), it follows that

$$I(M) \leq \frac{1}{M+1} \inf_Q \sum_{j=0}^{M} K(P_j, Q). \tag{52}$$

Different choices of Q in (52) yield different upper bounds on $I(M)$. One gets, for example,

$$I(M) \leq \min_{k=0,1\ldots,M} \frac{\sum_{j=0}^{M} K(P_j, P_k)}{M+1} \leq \frac{\sum_{j,k=0}^{M} K(P_j, P_k)}{(M+1)^2} \leq \max_{j,k\in\{0,1\ldots,M\}} K(P_j, P_k). \tag{53}$$

These bounds are very frequently used in conjunction with Fano's inequality; see e.g., the two examples in Section 4.6.1. The last bound $\max_{j,k\in\{0,1\ldots,M\}} K(P_j, P_k)$ is called the Kullback-Leibler diameter of $\{P_j\}_{j=0}^M$.

We will see that quite often (in nonparametric problems) the bounds in (53) are, in general, quite inaccurate and describe an improved bounds due to [17].

Let $\mathcal{P}$ be a collection of distributions. In analogy with Definition 4.25, we say that the collection of distributions $\{Q_i\}_{i=1}^N$ form an *ϵ-cover of $\mathcal{P}$* in KL-divergence if for all $P \in \mathcal{P}$, there exists some i such that $K(P, Q_i) \leq \epsilon^2$. With this, we may define the KL-covering number of the set $\mathcal{P}$ as

$$N(\epsilon, \mathcal{P}, K) := \inf\left\{N \in \mathbb{N} : \exists\, Q_i, i = 1, \ldots, N, \text{ such that } \sup_{P\in\mathcal{P}} \min_i K(P, Q_i) \leq \epsilon^2\right\},$$

where $N(\epsilon, \mathcal{P}, K) = +\infty$ if no such cover exists.

Let $P_0, P_1, \ldots, P_M$ be probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. Recall that

$$I(M) := \frac{1}{M+1} \sum_{j=0}^{M} K(P_j, \overline{P}).$$

Let $\mathcal{P}$ be a collection of distributions such that $P_j \in \mathcal{P}$, for all $j = 0, 1, \ldots, M$.

Proposition 4.32. $I(M) \leq \inf_{\epsilon > 0}\{\epsilon^2 + \log N(\epsilon, \mathcal{P}, K)\}$.

Proof. By carefully choosing the distribution Q in the upper bound in (52) above, we will obtain the desired. Now, assume that the distributions $\{Q_i\}_{i=1}^N$, form an ϵ-cover of the family $\mathcal{P}$, meaning that $\min_i K(P, Q_i) \leq \epsilon^2$, for all $P \in \mathcal{P}$. Let p_j and q_i denote the densities of P_j and Q_i with respect to some fixed base measure ν on $\mathcal{X}$ (the choice of based measure does not matter). Then defining the distribution $Q := (1/N) \sum_{i=1}^N Q_i$ (with density q with respect to ν), we obtain for any j,

$$\begin{aligned} K(P_j, Q) &= \int \log\left(\frac{p_j}{q}\right) p_j \, d\nu = \int \log\left(\frac{p_j}{N^{-1}\sum_{i=1}^N q_i}\right) p_j \, d\nu \\ &= \log N + \int \log\left(\frac{p_j}{\sum_{i=1}^N q_i}\right) p_j \, d\nu \leq \log N + \int \log\left(\frac{p_j}{\max_i q_i}\right) p_j \, d\nu \\ &\leq \log N + \min_i \int \log\left(\frac{p_j}{q_i}\right) p_j \, d\nu = \log N + \min_i K(P_j, Q_i). \end{aligned}$$

By our assumption that the Q_i's form a cover which gives the desired result, as $\epsilon > 0$ was arbitrary (as was our choice of the cover). □

4.7.1 A general scheme for proving minimax bounds using global packings

There is now a four step process to proving minimax lower bounds using the global Fano method. Our starting point is to recall the Fano minimax lower bound in (46) of Corollary 4.22 and (37), which begins with the construction of a set of points $\{\theta(P_j)\}_{j=0}^M$ that form a $2s$-packing of a set Θ in the semi-metric d. With this in mind, we perform the following four steps:

(i) *Bound the packing entropy.* Give a lower bound on the packing number of the set Θ with $2s$-separation (call this lower bound $D(s) \equiv M + 1$).

(ii) *Bound the metric entropy.* Give an upper bound on the KL-metric entropy of the class $\mathcal{P}$ of distributions containing all the distributions $\{P_j\}_{j=0}^M$, i.e., an upper bound on $\log N(\epsilon, \mathcal{P}, K)$.

(iii) *Find the critical radius.* Using Proposition 4.32 we can now balance $I(M)$ and the packing entropy $\log D(s)$. To that end, we choose ϵ_n and $s_n > 0$ at the *critical radius*, defined as follows: choose any ϵ_n such that

$$\epsilon_n^2 \geq \log N(\epsilon_n, \mathcal{P}, K), \tag{54}$$

and choose the largest $s_n > 0$ such that

$$\log D(s_n) \geq 4\epsilon_n^2 + 2\log 2. \tag{55}$$

Then,

$$\log D(s_n) \geq 2\log N(\epsilon_n, \mathcal{P}, K) + 2\epsilon_n^2 + 2\log 2 \geq 2(I(M) + \log 2).$$

(iv) *Apply the Fano minimax bound* (46). Having chosen s_n and ϵ_n as above, we immediately obtain that

$$p_{e,M} \geq 1 - \frac{I(M) + \log 2}{\log D(s_n)} \geq 1 - \frac{1}{2} = \frac{1}{2},$$

and thus, we obtain

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[w(s_n^{-1} d(\hat{\theta}_n, \theta)) \right] \geq \frac{1}{2} w(s_n).$$

4.7.2 An example

Example 4.33 (Lipschitz regression)**.** Consider data (X_i, Y_i), $i = 1, \ldots, n$, from a nonparametric regression model 31 with $X_i = i/n$, $f : [0,1] \to [0,1]$ is *L-Lipschitz* and the (unobserved) errors $\xi_1, \ldots, \xi_n$ are i.i.d. $N(0, \sigma^2)$. The goal of this section is to find the (optimal) lower bound on the rate of convergence for any estimator of f based on the discrete L_2-loss $d(\cdot, \cdot)$ defined in (32). Let

$$\mathcal{F} := \{f : [0,1] \to [0,1] | \, f \text{ is } L\text{-Lipschitz}\}.$$

Result: Note that for $\delta > 0$,

$$c_1 \frac{L}{\delta} \leq \log D(\delta, \mathcal{F}, \|\cdot\|_\infty) \leq c_2 \frac{L}{\delta},$$

where $c_2 \geq c_1 > 0$.

Exercise (HW2): Show that $\log(\epsilon, \mathcal{P}, K) \leq c_2 \sqrt{\frac{n}{2\sigma^2}} L \epsilon^{-1}$. This completes step (ii). Further show that (54) holds for $\epsilon_n \geq \left(\frac{c_2 L \sqrt{n}}{\sqrt{2\sigma^2}} \right)^{1/3}$ and (55) holds for $s_n = c(\sigma^2 L/n)^{1/3}$, for some $c > 0$. Hence, show that the lower bound on the minimax rate is $(\sigma^2 L/n)^{1/3}$ which involves the right scaling in n, L and the variance σ^2.

5 Reproducing kernel Hilbert spaces

5.1 Hilbert spaces

A vector space in $\mathbb{R}^n$ can be spanned by a finite set of vectors. Classes of functions may also form vector spaces over $\mathbb{R}$, but these spaces are rarely spanned by a finite set of functions. In this chapter we study a special class of functions that form a Hilbert space (a generalization of the notion of Euclidean space) and admit expansions like that as in a finite dimensional vector space.

Definition 5.1 (Hilbert space)**.** Let $\mathcal{H}$ be a (real) vector space together with a function $\langle\cdot,\cdot\rangle : \mathcal{H}\times\mathcal{H}\to\mathbb{R}$ (the inner product) for which

$$\begin{aligned}
\langle x, y\rangle &= \langle y, x\rangle, \qquad \forall\, x, y \in \mathcal{H} \text{ (symmetric)},\\
\langle x, ay + bz\rangle &= a\langle x, y\rangle + b\langle x, z\rangle, \qquad \forall\, x, y, z \in \mathcal{H},\ \alpha, \beta \in \mathbb{R} \text{ (bilinear)},\\
\langle x, x\rangle &\geq 0, \qquad x \in \mathcal{H}, \ \text{ with equality if and only if } x = 0.
\end{aligned}$$

Suppose that the norm in $\mathcal{H}$ is defined by

$$\|x\| := \sqrt{\langle x, x\rangle}$$

and $\mathcal{H}$ is complete[6] in the metric $d(x, y) := \|x - y\|$. Then $\mathcal{H}$ forms a Hilbert space equipped with the inner product $\langle\cdot,\cdot\rangle$.

Example 5.2 (Euclidean space)**.** Let $\mathcal{H} = \mathbb{R}^m$ and $\langle x, y\rangle := \sum_{i=1}^m x_i y_i$ (where $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$); or more generally $\langle x, y\rangle = x^\top A y$ where A is a symmetric positive definite matrix.

Example 5.3 (Euclidean matrices)**.** Let $\mathcal{H} = \mathbb{R}^{m\times m}$ be the set of all $m\times m$ matrices. Define $\langle x, y\rangle := \mathrm{tr}(xy^\top)$. Then $\langle\cdot,\cdot\rangle$ defines a Hilbert space over $m \times m$ matrices.

Example 5.4 (L_2 space)**.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $L_2(\Omega, \mathcal{A}, \mu)$ be the set (of equivalence classes) of all square integrable functions with

$$\langle f, g\rangle := \int f g \, d\mu.$$

Example 5.5 (Sobolev space)**.** The Sobolev space $W_m[0, 1]$ is the collection of all functions $f : [0, 1] \to \mathbb{R}$ with $m - 1$ continuous derivatives, $f^{(m-1)}$ absolutely continuous, and $\|f^{(m)}\| < \infty$. With an inner product $\langle\cdot,\cdot\rangle$ defined by

$$\langle f, g\rangle := \sum_{k=0}^{m-1} f^{(k)}(0) g^{(k)}(0) + \int_0^1 f^{(m)}(x) g^{(m)}(x) dx, \qquad f, g \in W_m[0, 1], \tag{56}$$

[6] A metric space $\mathcal{H}$ is said to be *complete* if every Cauchy sequence in $\mathcal{H}$ has a limit in $\mathcal{H}$.

$W_m[0,1]$ is a Hilbert space.

Here are some properties of any Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$:

- The *Cauchy-Schwarz* inequality holds:

$$|\langle x, y\rangle| \leq \|x\|\|y\|, \qquad \forall\, x, y \in \mathcal{H}.$$

- The *Parallelogram laws* assert that

$$\|x+y\|^2+\|x-y\|^2 = 2(\|x\|^2+\|y\|^2) \;\text{ and }\; \|x+y\|^2-\|x-y\|^2 = 4\langle x, y\rangle \quad \forall\, x, y \in \mathcal{H}.$$

- (Linear functional) A function $\varphi : \mathcal{H} \to \mathbb{R}$ is said to be a *linear functional* if $\varphi(\alpha x + \beta y) = \alpha\varphi(x) + \beta\varphi(y)$ whenever $x, y \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$. For example, for a fixed $y \in \mathcal{H}$,

$$\varphi_y(x) := \langle x, y\rangle, \qquad \forall\, x \in \mathcal{H}, \tag{57}$$

  defines a continuous linear functional, a linear functional that is continuous with respect to the metric induced by the inner product.

- (Dual space) The *dual space* $\mathcal{H}^*$ (of $\mathcal{H}$) is the space of all continuous linear functions from $\mathcal{H}$ into $\mathbb{R}$. It carries a natural norm[7], defined by

$$\|\varphi\|_{\mathcal{H}^*} = \sup_{\|x\|=1, x\in\mathcal{H}} |\varphi(x)|, \qquad \varphi \in \mathcal{H}^*.$$

  This norm satisfies the parallelogram laws.

  Result: The *Riesz representation theorem* gives a convenient description of the dual. It states that *any continuous linear functional can be represented in the form* (57) *for some* $y \in \mathcal{H}$ *depending on the linear functional.*

[7] Exercise (HW3): Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces over $\mathbb{R}$. A function $T : \mathcal{X} \to \mathcal{Y}$ is called a linear operator if

$$T(cx_1 + x_2) = cT(x_1) + T(x_2), \qquad \forall\, x_1, x_2 \in \mathcal{X}, c \in \mathbb{R}.$$

The *operator norm* (or spectral norm) of T is defined as

$$\|T\| := \sup\{\|T(x)\| : \|x\| \leq 1\},$$

and T is called *bounded* if $\|T\| < \infty$.

(a) Show that a bounded operator T is continuous: If $\|x_n - x\| \to 0$, then $\|T(x_n) - T(x)\| \to 0$.

(b) Show that a continuous linear operator T is bounded.

(c) Let $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = \mathbb{R}^n$, with the usual Euclidean norms. Let A be an $n \times m$ matrix, and define a linear operator T by $T(x) = Ax$. Relate the operator norm $\|T\|$ to the eigenvalues of $A^\top A$.

Thus to every element φ of the dual $\mathcal{H}^*$ there exists one and only one $u_\varphi \in \mathcal{H}$ such that $\langle x, u_\varphi \rangle = \varphi(x)$, for all $x \in \mathcal{H}$. The inner product on the dual space $\mathcal{H}^*$ satisfies

$$\langle \varphi, \psi \rangle_{\mathcal{H}^*} := \langle u_\psi, u_\varphi \rangle_{\mathcal{H}}.$$

So the dual space is also an inner product space. The dual space is also complete, and so it is a Hilbert space in its own right.

- (Convex sets) Recall that a subset $\mathcal{H}_0 \subset \mathcal{H}$ is called a *linear subspace* if it is closed under addition and scalar multiplication; i.e., $\alpha x + \beta y \in \mathcal{H}_0$ whenever $x, y \in \mathcal{H}_0$ and $\alpha, \beta \in \mathbb{R}$.

  A subset $C \subset \mathcal{H}$ is said to be *convex* if it contains the line joining any two of its elements, i.e., $\alpha x + (1-\alpha) y \in C$ whenever $x, y \in C$ and $0 \le \alpha \le 1$.

  A set $C \subset \mathcal{H}$ is said to be a *cone* if $\alpha x \in C$ whenever $x \in C$ and $\alpha \ge 0$. Thus, C is a convex cone if $\alpha x + \beta y \in C$ whenever $x, y \in C$ and $0 \le \alpha, \beta < \infty$. Any linear subspace is, by definition, also a convex cone. Any ball, $B = \{x \in \mathcal{H} : \|x\| \le c\}$, $c > 0$, is a convex set, but not a convex cone.

- (Projection theorem) If $C \subset \mathcal{H}$ is a closed convex set and $z \in \mathcal{H}$, then there is a unique $x \in C$ for which

  $$\|x - z\| = \inf_{z \in C} \|y - z\|.$$

  In fact, $x \in C$ satisfies the condition

  $$\langle z - x, y - x \rangle \le 0, \qquad \forall\, y \in C. \tag{58}$$

  The element $x \in C$ is called the *projection* of z onto C and denoted by $\Pi_C(z)$. Prove the projection theorem. (Exercise (HW3))

  In particular, if C is a convex cone, setting $y = x/2$ and $y = 2x$ in (58) shows that $\langle z - x, x \rangle = 0$. Thus, x is the unique element of C for which

  $$\langle z - x, x \rangle = 0 \qquad \text{and} \qquad \langle z - x, y \rangle \le 0 \quad \forall\, y \in C.$$

  If C is a linear subspace, then $z - x$ is *orthogonal* to C, i.e.,

  $$\langle z - x, y \rangle = 0 \quad \forall\, y \in C.$$

- (Orthogonal complement) Suppose that $\mathcal{H}_0 \subset \mathcal{H}$. The *orthogonal complement* of $\mathcal{H}_0$ is

  $$\mathcal{H}_0^\perp := \{x \in \mathcal{H} : \langle x, y \rangle = 0,\ \forall\, y \in \mathcal{H}_0\}.$$

Result: The orthogonal complement of a subset of a Hilbert space is a closed linear subspace.

The projection theorem states that if $C \subset \mathcal{H}$ is a closed subspace, then any $z \in C$ may be uniquely represented as $z = x + y$, where $x \in C$ is the best approximation to z, and $y \in C^{\perp}$.

Result: If $C \subset \mathcal{H}$ is a closed subspace, then $\mathcal{H} = C \oplus C^{\perp}$, where

$$A \oplus B := \{x + y : x \in A, y \in B\}.$$

Thus, every closed subspace C of $\mathcal{H}$ has a closed complementary subspace $C^{\perp}$.

- (Orthonormal basis) A collection $\{e_t : t \in T\} \subset \mathcal{H}$ (where T is any index set) is said to be *orthonormal* if $e_s \perp e_t$ (i.e., $\langle e_s, e_t \rangle = 0$) for all $s \neq t$ and $\|e_t\| = 1$, for all $t \in T$.

  As in the finite-dimensional case, we would like to represent elements in our Hilbert space as linear combinations of elements in an orthonormal collection, but extra care is necessary because some infinite linear combinations may not make sense.

  The *linear span* of $S \subset \mathcal{H}$, denoted $\text{span}(S)$, is the collection of all finite linear combinations $\alpha_1 x_1 + \cdots + \alpha_n x_n$ with $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in S$. The closure of this set is denoted by $\overline{\text{span}}(S)$.

  An orthonormal collection $\{e_t, t \in T\}$, is called an *orthonormal basis* for the Hilbert space $\mathcal{H}$ if $\langle e_t, x \rangle \neq 0$ for some $t \in T$, for every nonzero $x \in \mathcal{H}$.

  Result: *Every Hilbert space has an orthonormal basis.*

  When $\mathcal{H}$ is *separable*[8], a basis can be found by applying the Gram-Schmidt algorithm to a countable dense set, and in this case the basis will be countable.

  Result: If $\{e_n\}_{n \geq 1}$, is an orthonormal basis of $\mathcal{H}$, then each $x \in \mathcal{H}$ may be written as $x = \sum_{k=1}^{\infty} \langle x, e_k \rangle e_k$. Show this. (Exercise (HW3))

[8] A topological space is called *separable* if it contains a countable, dense subset; i.e., there exists a sequence $\{x_n\}_{n=1}^{\infty}$ of elements of the space such that every nonempty open subset of the space contains at least one element of the sequence.

5.2 Reproducing Kernel Hilbert Spaces

Definition 5.6 (Reproducing kernel Hilbert space)**.** Let $\mathcal{X}$ be an arbitrary set and $\mathcal{H}$ a Hilbert space of real-valued functions on $\mathcal{X}$. The *evaluation functional* over the Hilbert space of functions $\mathcal{H}$ is a linear functional that evaluates each function at a point $x \in \mathcal{X}$,

$$L_x : f \mapsto f(x) \quad \forall f \in \mathcal{H}.$$

We say that $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS) if L_x is continuous at any f in $\mathcal{H}$, for all $x \in \mathcal{X}$ (equivalently, if for all $x \in \mathcal{X}$, L_x is a bounded[9] operator on $\mathcal{H}$).

Thus, a RKHS is a Hilbert space of functions in which point evaluation is a continuous linear functional. Roughly speaking, this means that if two functions f and g in the RKHS are close in norm, i.e., $\|f - g\|$ is small, then f and g are also pointwise close, i.e., $|f(x) - g(x)|$ is small for all $x \in \mathcal{X}$.

The Riesz representation theorem implies that for all $x \in \mathcal{X}$ there exists a unique element K_x of $\mathcal{H}$ with the *reproducing property*:

$$f(x) = L_x(f) = \langle f, K_x \rangle \quad \forall\, f \in \mathcal{H}. \tag{59}$$

Since K_y is itself a function in $\mathcal{H}$ we have that for each $y \in \mathcal{X}$,

$$K_y(x) = \langle K_y, K_x \rangle.$$

This allows us to define the *reproducing kernel* of $\mathcal{H}$ as a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$K(x, y) = \langle K_x, K_y \rangle.$$

From this definition it is easy to see (Exercise (HW3)) that $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is both symmetric and *positive definite*, i.e.,

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \geq 0, \tag{60}$$

for any $n \in \mathbb{N}, x_1, \ldots, x_n \in \mathcal{X}$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Thus, the "Gram Matrix" $K = ((K_{ij}))_{n \times n}$ defined by $K_{ij} = k(x_i, x_j)$ is positive semi-definite.

[9] A functional $\lambda : \mathcal{H} \to \mathbb{R}$ is *bounded* if there is a finite real constant B so that, for all $f \in \mathcal{H}$, $|\lambda(f)| \leq B\|f\|_{\mathcal{H}}$. It can be shown that the continuity of the functional λ is equivalent to boundedness.

Example 5.7 (Linear kernel)**.** Let $\mathcal{X} = \mathbb{R}^d$ and let $K(x, y) := x^\top y$, for any $x, y \in \mathbb{R}^d$, be the usual inner product in $\mathbb{R}^d$. Then the linear kernel K is symmetric and positive definite.

Example 5.8 (RKHS of the linear kernel)**.** Let $\mathcal{X} = \mathbb{R}^d$. Consider the space $\mathcal{H}$ of all linear forms on $\mathbb{R}^d$: $\mathcal{H} := \{f(x) = w^\top x : w \in \mathbb{R}^d\}$. Define the inner product by $\langle f, g\rangle_{\mathcal{H}} = v^\top w$ for $f(x) = v^\top x$ and $g(x) = w^\top x$. Then, the linear kernel $K(x, y) := x^\top y$ is a reproducing kernel for $\mathcal{H}$.

Example 5.9 (Gaussian and Laplace kernels)**.** When $\mathcal{X} = \mathbb{R}^d$, the *Gaussian* and *Laplace* kernels are defined as

$$K(x, y) := \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right), \quad K(x, y) := \exp\left(-\frac{\|x - y\|_2}{2\sigma^2}\right),$$

respectively, where $x, y \in \mathbb{R}^d$, $\sigma^2 > 0$. Both kernels are positive definite, but the proof of this fact is more involved than for the linear kernel.

The Moore-Aronszajn theorem (see below) is a sort of converse to (60): if a function K satisfies these conditions (symmetric and positive definite) then there is a Hilbert space of functions on $\mathcal{X}$ for which it is a reproducing kernel.

Proposition 5.10 (Moore-Aronszajn theorem)**.** Suppose that K is a symmetric, positive definite kernel on a set $\mathcal{X}$. Then there is a unique Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ for which K is a reproducing kernel.

Proof. The complete proof of this result is rather long. We give a sketch of the proof here. For all x in $\mathcal{X}$, define $K_x := K(x, \cdot)$. Let $\mathcal{H}_0$ be the linear span of $\{K_x : x \in \mathcal{X}\}$. Define an inner product on $\mathcal{H}_0$ by

$$\left\langle \sum_{j=1}^n \beta_j K_{y_j}, \sum_{i=1}^m \alpha_i K_{x_i} \right\rangle_{\mathcal{H}_0} := \sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j K(y_j, x_i),$$

where $\{\alpha_i\}_{i=1}^m, \{\beta_j\}_{j=1}^n \subset \mathbb{R}$ and $\{x_i\}_{i=1}^m, \{y_j\}_{j=1}^n \subset \mathcal{X}$. The symmetry of this inner product follows from the symmetry of K and the non-degeneracy follows from the fact that K is positive definite. We can show that

1. the point evaluation functionals L_x are continuous on $\mathcal{H}_0$,
2. any Cauchy sequence f_n in $\mathcal{H}_0$ which converges pointwise to 0 also converges in in $\mathcal{H}_0$-norm to 0.

Let $\mathcal{H}$ be the completion of $\mathcal{H}_0$ with respect to this inner product. We define an inner product in $\mathcal{H}$ as: suppose that $\{f_n\}_{n\geq 1}$ and $\{g_n\}_{n\geq 1}$ are sequences in $\mathcal{H}_0$ converging

to f and g respectively. Then $\{\langle f_n, g_n\rangle_{\mathcal{H}_0}\}_{n\geq 1}$ is convergent and its limit depends only on f and g (see [2, Lemma 5] for a proof of the above). Thus we define

$$\langle f, g\rangle_{\mathcal{H}} := \lim_{n\to\infty} \langle f_n, g_n\rangle_{\mathcal{H}_0}.$$

Next we have to show that $\mathcal{H}$ is indeed a Hilbert space with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ (see [2, Lemma 6] for a proof of this; we will have to further show that $\mathcal{H}$ is complete). Further we can show that $\mathcal{H}_0$ is dense in $\mathcal{H}$ (see see [2, Lemma 7] for a proof of this) and that the point evaluation map is continuous on $\mathcal{H}$ (see see [2, Lemma 8] for a proof of this).

Now we can check the reproducing property (59), i.e., $\langle f, K_x\rangle_{\mathcal{H}} = f(x)$, for all $f \in \mathcal{H}$, for all $x \in \mathcal{X}$. To prove uniqueness, let G be another Hilbert space of functions for which K is a reproducing kernel. For any x and y in $\mathcal{X}$, (59) implies that

$$\langle K_x, K_y\rangle_{\mathcal{H}} = K(x, y) = \langle K_x, K_y\rangle_G.$$

By linearity, $\langle\cdot,\cdot\rangle_{\mathcal{H}} = \langle\cdot,\cdot\rangle_G$ on the span of $\{K_x : x \in \mathcal{X}\}$. Then $G = \mathcal{H}$ by the uniqueness of the completion. See `http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/RKHS_Notes1.pdf` for a more detailed discussion on this proof. □

This proposition allows one to construct reproducing kernels on complicated spaces $\mathcal{X}$ (such as graphs, images) only by checking that the proposed kernel is positive definite and without explicitly defining the Hilbert space $\mathcal{H}$.

5.2.1 The Representer theorem

The representer theorem [7] shows that solutions of a large class of optimization problems can be expressed as kernel expansions over the sample points. We present a slightly more general version of the theorem with a simple proof [10].

Let $\mathcal{X}$ be an arbitrary set and let $\mathcal{H}_K$ be a RKHS of real valued functions on $\mathcal{X}$ with reproducing kernel $K(\cdot,\cdot)$. Let $\{(Y_i, X_i) : i = 1, \ldots, n\}$ be given data (the "training set") with $X_i \in \mathcal{X}$ (the "attribute vector"), and $Y_i \in \mathcal{Y}$ being the "response".

Theorem 5.11. Denote by $\Omega : [0, \infty) \to \mathbb{R}$ a strictly increasing function. Let $\ell : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Y})^n \to \mathbb{R} \cup \{\infty\}$ be an arbitrary loss function. Then each minimizer $f \in \mathcal{H}_K$ of the regularized risk functional

$$\ell(\{(X_i, Y_i, f(X_i)\}_{i=1}^n) + \Omega(\|f\|^2_{\mathcal{H}_K}) \tag{61}$$

admits a representation of the form

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(X_i, x), \qquad \forall\, x \in \mathcal{X}, \quad \text{where } \alpha_1, \ldots, \alpha_n \in \mathbb{R}. \tag{62}$$

Proof. Take any $f \in \mathcal{H}_K$. As $\mathcal{H}_K$ is a Hilbert space there is a unique decomposition of f as the sum of $f_n \in S := \overline{\text{span}}(\{K(X_1, \cdot), \ldots, K(X_n, \cdot)\})$ and $f_\perp \in S^\perp \subset \mathcal{H}_K$, the orthogonal complement of S (in $\mathcal{H}_K$):

$$f(x) = f_n(x) + f_\perp(x) := \sum_{i=1}^{n} \alpha_i K(X_i, x) + f_\perp(x).$$

Here $\alpha_i \in \mathbb{R}$ and $\langle f_\perp, K(X_i, \cdot)\rangle = 0$ for all $i = 1, \ldots, n$. By the reproducing property,

$$f(X_i) = \langle f, K(X_i, \cdot)\rangle = \langle f_n, K(X_i, \cdot)\rangle + \langle f_\perp, K(X_i, \cdot)\rangle = \langle f_n, K(X_i, \cdot)\rangle = f_n(X_i),$$

which implies that

$$\ell(\{(X_i, Y_i, f(X_i)\}_{i=1}^n) = \ell(\{(X_i, Y_i, f_n(X_i)\}_{i=1}^n).$$

Secondly, for all $f_\perp \in S^\perp$,

$$\Omega(\|f\|_{\mathcal{H}_K}^2) = \Omega\left(\left\|\sum_{i=1}^{n} \alpha_i K(X_i, \cdot)\right\|^2 + \|f_\perp\|^2\right) \geq \Omega\left(\left\|\sum_{i=1}^{n} \alpha_i K(X_i, \cdot)\right\|^2\right).$$

Hence, $\ell(\cdots)$ depends only on the component of f lying in the subspace S and $\Omega(\cdot)$ is minimized if f lies in that subspace. Hence, the criterion function is minimized if f lies in that subspace, and we can express the minimizer as in (62).

Note that as $\Omega(\cdot)$ is strictly non-decreasing, $\|f_\perp\|$ must necessarily be zero for f to be the minimizer of (61), implying that $\hat{f}$ must necessarily lie in the subspace S. □

Monotonicity of Ω does not prevent the regularized loss functional (61) from having multiple local minima. To ensure a global minimum, we would need to require convexity. If we discard the strictness of the monotonicity, then it no longer follows that each minimizer of the regularized loss admits an expansion (62); it still follows, however, that there is always another solution that is as good, and that does admit the expansion.

The significance of the representer theorem is that although we might be trying to solve an optimization problem in an infinite-dimensional space $\mathcal{H}_K$, containing linear combinations of kernels centered on arbitrary points of $\mathcal{X}$, it states that the solution lies in the span of n particular kernels — those centered on the training points. For suitable choices of loss functions, many of the α_i's often equal 0.

5.2.2 Feature map and kernels

A kernel can be thought of as a notion of similarity measure between two points in the "input points" in $\mathcal{X}$. For example, if $\mathcal{X} = \mathbb{R}^d$, then the canonical dot product

$$K(x, x') = \langle x, x' \rangle = \sum_{i=1}^{d} x_i x_i' ; \qquad x, x' \in \mathbb{R}^d.$$

can be taken as the kernel.

If $\mathcal{X}$ is a more complicated space, then we can still define a kernel as follows.

Definition 5.12 (Kernel)**.** Let $\mathcal{X}$ be a non-empty set. The function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a *kernel* if there exists a real Hilbert space $\mathcal{E}$ (not necessarily a RKHS), and a map $\varphi : \mathcal{X} \to \mathcal{E}$ such that for all $x, y \in \mathcal{X}$,

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{E}}. \tag{63}$$

Such map $\varphi : \mathcal{X} \to \mathcal{E}$ is referred to as the *feature map*, and space $\mathcal{E}$ as the *feature space*. Thus kernels are functions that can be written as an inner product in a feature space.

Exercise (HW3): Show that $K(\cdot, \cdot)$ defined in (63) is a positive definite function.

Thus, we can think of the patterns as $\varphi(x)$, $\varphi(x')$, and carry out geometric algorithms in the Hilbert space (feature space) $\mathcal{E}$. Usually, $\dim(\mathcal{E}) \gg \dim(\mathcal{X})$ (if $\dim(\mathcal{X})$ is defined).

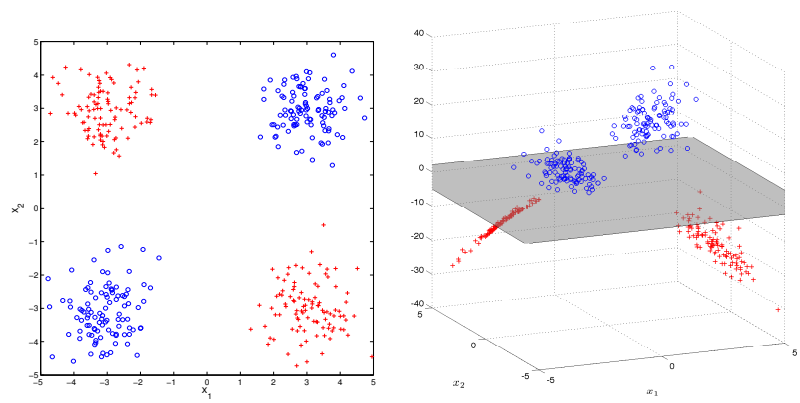
Note that for a given kernel, there may be more than one feature map, as demonstrated by the following example: take $\mathcal{X} = \mathbb{R}$, and $K(x, y) = xy = [\frac{x}{\sqrt{2}} \frac{x}{\sqrt{2}}][\frac{y}{\sqrt{2}} \frac{y}{\sqrt{2}}]^\top$, where we defined the feature maps $\varphi(x) = x$ and $\tilde{\varphi}(x) = [\frac{x}{\sqrt{2}} \frac{x}{\sqrt{2}}]$, and where the feature spaces are respectively, $\mathcal{E} = \mathbb{R}$ and $\tilde{\mathcal{E}} = \mathbb{R}^2$.

Exercise (HW3): For every $x \in \mathcal{X}$, assume that the sequence $\{f_n(x)\}_{n \geq 1} \in \ell_2(\mathbb{N})$, where $f_n : \mathcal{X} \to \mathbb{R}$, for all $n \in \mathbb{N}$. Then $K(x_1, x_2) := \sum_{n=1}^{\infty} f_n(x_1) f_n(x_2)$ is a kernel.

As $k(\cdot, \cdot)$ defined in (63) is symmetric and positive definite it induces a unique RKHS. Thus, to construct reproducing kernels on complicated spaces we only need to find a feature map φ.

Another way to characterize a symmetric positive definite kernel K is via the Mercer's Theorem.

Figure 4: (Feature space and feature map) On the left, the points are plotted in the original space. There is no linear classifier that can separate the red crosses from the blue circles. Mapping the points to a higher dimensional feature space ($x \mapsto \varphi(x) := (x_1, x_2, x_1x_2) \in \mathbb{R}^3$), we obtain linearly separable classes. A possible decision boundary is shown as a gray plane.



Definition 5.13 (Integral operator)**.** Let K be a continuous kernel on compact metric space $\mathcal{X}$, and let ν be a finite Borel measure on $\mathcal{X}$. Let $T_K : L_2(\mathcal{X}, \nu) \to \mathcal{C}(\mathcal{X})$ ($\mathcal{C}(\mathcal{X})$ being the space of all continuous real-valued functions on $\mathcal{X}$ thought of as a subset of $L_2(\mathcal{X}, \nu)$) be the linear map defined as:

$$(T_K f)(\cdot) = \int_{\mathcal{X}} K(x, \cdot) f(x)\, d\nu(x), \qquad f \in L_2(\mathcal{X}, \nu).$$

Such a T_K is called an integral operator.

Exercise (HW3): Show that T_K is a continuous function for all $f \in L_2(\mathcal{X}, \nu)$.

Theorem 5.14 (Mercer's Theorem)**.** Suppose that K is a continuous positive definite kernel on a compact set $\mathcal{X}$, and let ν be a finite Borel measure on $\mathcal{X}$ with $\mathrm{supp}(\nu) = \mathcal{X}$. Then there is an orthonormal basis $\{\psi_i\}_{i \in J}$ of $L_2(\mathcal{X}, \nu)$ consisting of eigenfunctions of T_K such that the corresponding sequence of eigenvalues $\{\lambda_i\}$ are non-negative. The eigenfunctions corresponding to non-zero eigenvalues are continuous on $\mathcal{X}$ and $K(\cdot, \cdot)$ has the representation

$$K(u, v) = \sum_{i \in J} \lambda_i \psi_i(u) \psi_i(v), \qquad u, v \in \mathcal{X},$$

where the convergence is absolute and uniform, i.e.,

$$\lim_{n \to \infty} \sup_{u, v \in \mathcal{X}} \left| K(u, v) - \sum_{i \in J : i = 1}^{n} \lambda_i \psi_i(u) \psi_i(v) \right| = 0.$$

Example 5.15. To take an analogue in the finite case, let $\mathcal{X} = \{x_1, \ldots, x_n\}$. Let $K_{ij} = K(x_i, x_j)$, and $f : \mathcal{X} \to \mathbb{R}^n$ with $f_i = f(x_i)$ and let ν be the counting measure. Then,

$$T_K f = \sum_{i=1}^{n} K(x_i, \cdot) f_i$$

and

$$\forall\, f, f^\top K f \geq 0 \;\Rightarrow\; K \text{ is p.s.d.} \;\Rightarrow\; K = \sum_{i=1}^{n} \lambda_i v_i v_i^\top.$$

Hence,

$$K(x_i, x_j) = K_{ij} \quad = \quad (V \Lambda V^\top)_{ij} = \sum_{k=1}^{n} \lambda_k v_{ki} v_{kj} = \sum_{k=1}^{n} \lambda_k v_{ki} v_{kj}.$$

Note that Mercer's theorem gives us another feature map for the kernel K, since:

$$K(u, v) = \sum_{i \in J} \lambda_i \psi_i(u) \psi_i(v) = \langle \varphi(u), \varphi(v) \rangle_{\ell_2(J)},$$

so we can take $\ell_2(J)$ as a *feature space*, and the corresponding *feature map* is $\varphi : \mathcal{X} \to \ell_2(J)$ where

$$\varphi : x \mapsto \left\{ \sqrt{\lambda_i} \psi_i(x) \right\}_{\ell_2(J)}.$$

This map is well defined as $\sum_{i \in J} |\sqrt{\lambda_i} \psi_i(x)|^2 = K(x, x) < \infty$.

Apart from the representation of the kernel function, Mercer's theorem also leads to a construction of RKHS using the eigenfunctions of the integral operator T_K.

5.3 Smoothing Splines

Let us consider again our nonparametric regression model

$$Y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are mean zero, uncorrelated random variables with a common variance σ^2. As with the kernel approach, there is a presumption that f is smooth. The smoothing spline approach tries to take direct advantage of this smoothness by augmenting the usual least squares criteria with a penalty for roughness. For instance, if the x_i's lie in $[0, 1]$, the estimator $\hat{f}$ might be chosen to minimize (over g)

$$\sum_{i=1}^{n} (Y_i - g(x_i))^2 + \lambda \|g^{(m)}\|_2^2,$$

where $\|\cdot\|^2$ is the L_2-norm of functions on $[0,1]$ under the Lebesgue measure, i.e.,

$$\|g\|_2^2 = \int_0^1 g^2(x)dx.$$

The constant λ is called the *smoothing parameter*. Larger values for λ will lead to a smoother $\hat{f}$, smaller values will lead to an estimate $\hat{f}$ that follows the observed data more closely (i.e.,$\hat{f}(x_i)$ will be closer to Y_i).

We can use the RKHS approach to solve the above optimization problem using the representer theorem. Please read Chapter 18.3 from [6] for the details (this was done in class).

Exercise (HW3): (Semiparametric models — partially linear regression model.) Consider a regression model with two explanatory variables x and w in which

$$Y_i = f(x_i) + \beta w_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

with $0 < x_1 < \ldots < x_n < 1$, $f \in W_m[0,1], \beta \in \mathbb{R}$, and the ϵ_i's are i.i.d. from $N(0, \sigma^2)$. This might be called a semiparametric model because the dependence on w is modeled parametrically, but the dependence on x is nonparametric. Following a penalized least squares approach, consider choosing $\hat{f}$ and $\hat{\beta}$ to minimize

$$\sum_{i=1}^{n} (Y_i - g(x_i) - \alpha w_i)^2 + \lambda \|g^{(m)}\|_2^2.$$

(a) Show that the estimator $\hat{f}$ will still be a natural spline of order $2m$.

(b) Derive explicit formulas based on linear algebra to compute $\hat{\beta}$ and $\hat{f}$.

5.4 Classification and Support Vector Machines

5.4.1 The problem of classification

We observe the data

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

where (X_i, Y_i)'s are i.i.d. random pairs, X_i takes values in the measurable space $(\mathcal{X}, \mathcal{A})$ and $Y_i \in \{-1, +1\}$ is a label. A new observation X arrives and the aim of classification is to predict the corresponding Y. We can interpret this task as a classification of X into one of the two groups labeled with -1 or $+1$.

Example 5.16 (Spam-filter)**.** We have a sample of n e-mail messages. For each message i, we count the percentages of 50 selected words characteristic for spam, such as the words *money*, *credit*, *Viagra* and so on. This constitutes the vectors of measurements $X_i \in \mathbb{R}^{50}$. Then, an expert provides the values $Y_i = +1$ if e-mail i is spam and $Y_i = -1$ otherwise. When a new message arrives, we would like to decide whether it is spam or not. For this purpose, we measure the corresponding percentages $X \in \mathbb{R}^{50}$ in this message, and based on X and on the training data $\mathcal{D}_n$, we have find a decision Y. The problem is usually solved by separating $\mathbb{R}^{50}$ in two parts (corresponding to spam and non-spam) the via a hyperplane depending on the training data $\mathcal{D}_n$. This is called a *linear classifier*.

At first sight, the observations are of the same form as in the problem of regression with random design. However, the important feature is that Y_i's are now binary. Even more important, in the classification context our final aim is different. We are not interested in estimation of the regression function $f^*(x) := \mathbb{E}(Y|X = x)$ but rather in predicting the value of the label Y. Note that the regression function has now the form

$$f^*(x) = \mathbb{P}(Y = 1|X = x) - \mathbb{P}(Y = -1|X = x) = 2\eta(x) - 1$$

where

$$\eta(x) := \mathbb{P}(Y = 1|X = x).$$

We define a *classifier* h as any measurable function from $\mathcal{X}$ to $\{-1, 1\}$. We predict the label for an observed X as $h(X)$. In practice, h depends on the observed data $\mathcal{D}_n$ but, in this section, we will assume that the observed data is fixed and thus h is just a function of X.

The performance of a classifier is characterized by the *probability of error* (also called the risk of classification), which is defined as:

$$R(h) := \mathbb{P}(Y \neq h(X)).$$

Our aim is to find the *best* classifier, i.e., a classifier which minimizes this risk:

$$h^* = \operatorname*{argmin}_h R(h).$$

We will call h^* the Bayes classifier and we call the minimal possible risk R^* the *Bayes risk*, i.e.,

$$R^* := \min_h R(h) = R(h^*).$$

The next theorem shows that such a classifier always exists.

Theorem 5.17. (i) The Bayes classifier has the form

$$h^*(x) = \begin{cases} 1, & \text{if } \eta(x) > 1/2, \\ -1, & \text{if } \eta(x) \le 1/2. \end{cases}$$

(ii) For any classifier h we have

$$R(h) - R(h^*) = \int_{x:h(x)\neq h^*(x)} |2\eta(x) - 1| dP_X(x),$$

where P_X is the probability distribution of X.

(iii) The Bayes risk is bounded by 1/2:

$$R^* = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}] \le \frac{1}{2}.$$

Example 5.18. Let $X \in \mathbb{R}^d$ admit a density $p(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}^d$. Then show that (Exercise (HW3))

$$h^*(x) = \begin{cases} 1, & \text{if } \pi p_1(x) > (1-\pi)p_{-1}(x), \\ -1, & \text{otherwise.} \end{cases}$$

where $\pi = \mathbb{P}(Y = 1)$ and $p_i(x) = p(x|Y = i)$ are conditional densities of X given $Y = i$, for $i = -1, 1$. This is the maximum likelihood classifier if and only if $\pi = 1/2$.

Parametric approach to classification: Assume that p_{-1}, p_1 in the example above are known, up to some parameters in $\mathbb{R}^k$. If we estimate these parameters then we can use the "plug-in classifier":

$$\begin{aligned} \hat{p}_1(X)\hat{\pi} > \hat{p}_{-1}(X)(1-\hat{\pi}) \quad &\Rightarrow \quad X \text{ is classified with } Y = 1 \\ \hat{p}_1(X)\hat{\pi} \le \hat{p}_{-1}(X)(1-\hat{\pi}) \quad &\Rightarrow \quad X \text{ is classified with } Y = -1 \end{aligned}$$

where $\hat{p}_1, \hat{p}_{-1}$ are parametric estimators of p_{-1}, p_1, and π. If p_i's are Gaussian densities $N(\theta_i, \Sigma), i = -1, 1$, then the decision rule is *linear*, which means that X is labeled 1 if and only if $X^\top a + b > 0$ for some $a \in \mathbb{R}^d, b \in \mathbb{R}$. Show this (Exercise (HW3)).

Nonparametric plug-in approach: We can also estimate the regression function $f^?*$ and then calculate $\hat{\eta}_n(x)$ as estimators of $\eta(x) = (f^*(x) + 1)/2$. Using this as the plug-in estimator, we derive the classifier

$$\hat{h}_n(x) = \begin{cases} 1, & \text{if } \hat{\eta}_n(x) > 1/2, \\ -1, & \text{otherwise.} \end{cases}$$

However, for this method to work we need $\hat{\eta}_n$ to be close to η, which is typically guaranteed if the function η has some smoothness properties. This is not always reasonable to assume.

Machine learning approach: This is also a fully nonparametric approach. Except for assuming that $Y \in \{-1, 1\}$, we do not make any other assumption on the joint distribution of (X, Y). The aim is to mimic the oracle h^* based on the data $\mathcal{D}_n$. But this is typically not possible. A more modest and achievable task is to mimic the oracle $h_{\mathcal{H}}$ within some reasonable restricted collection $\mathcal{H}$ of classifiers (also called the dictionary),

$$h_{\mathcal{H}} = \operatorname*{argmin}_{h \in \mathcal{H}} R(h).$$

An important example is given by the class of all linear classifiers:

$$\mathcal{H} = \{h : \mathbb{R}^d \to \{-1, 1\} : h(x) = 2I(x^\top a + b > 0) - 1, a \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

5.4.2 Minimum empirical risk classifiers

How to construct good classifiers based on the data? A first idea is to use the principle of *unbiased risk estimation*. We need to find an unbiased estimator for the risk $R(h) = \mathbb{P}(Y \neq h(X))$ and then to minimize this estimator in h over a given class $\mathcal{H}$. Note that the *empirical risk* is

$$R_n(h) = \frac{1}{n} \sum_{i=1}^{n} I\{Y_i \neq h(X_i)\}$$

is an unbiased estimator for $R(h)$ for all h. Minimizing $R_n(h)$ can be used to obtain a classifier.

Definition 5.19. Let $\mathcal{H}$ be a fixed collection of classifiers. The *empirical risk minimization* (ERM) classifier on $\mathcal{H}$ is defined by

$$\hat{h}_n := \operatorname*{argmin}_{h \in \mathcal{H}} R_n(h).$$

The ERM classifier always exists since the function R_n takes only a finite number of values, whatever is the class $\mathcal{H}$. Note that $I\{Y_i \neq h(X_i)\} = (Y_i - h(X_i))^2/4$ and thus

$$R_n(h) = \frac{1}{4n} \sum_{i=1}^{n} (Y_i - h(X_i))^2.$$

Therefore, $\hat{h}_n$ is the least squares estimator based on binary variables.

We expect the ERM classifier to have the risk close to that of the oracle classifier $h_{\mathcal{H}}$. Let us emphasize that we are not interested in accurate estimation of $h_{\mathcal{H}}$ and moreover there is no guarantee that $h_{\mathcal{H}}$ is unique. Mimicking the oracle means constructing a classifier $\hat{h}_n$ such that its risk $R(\hat{h}_n)$ is close to the oracle risk $\min_{h\in\mathcal{H}} R(h)$.

Computational considerations: To find $\hat{h}_n$, we should minimize on $\mathcal{H}$ the non-convex function

$$\frac{1}{n}\sum_{i=1}^{n} I\{Y_i \neq h(X_i)\},$$

and $\mathcal{H}$ is not a convex set because a convex combination of classifiers is not necessarily a classifier. Thus, the only possibility is to use combinatorial search. Even in the case where $\mathcal{H}$ is the class of linear rules, the computational complexity of combinatorial search on

$$A(\mathcal{D}_n) := \{b = (b_1, \ldots, b_n) : b_i = I\{h(X_i) \neq Y_i\}, h \in \mathcal{H}\}$$

will be of order $O(n^{d+1})$ where d is the dimension of the X_i's. This is prohibitive already for moderately large d.

A remedy is *convexification*: we replace the indicator function by a convex function and the class $\mathcal{H}$ by a convex class of functions, then solve a convex minimization problem and classify according to the sign of the solution. This approach was probably first used by [16] to define the method initially called the generalized portrait and renamed in the 1990's as the *support vector machine*.

5.4.3 Convexifying the ERM classifier

Let us first rewrite $R(h)$ in another form:

$$R(h) = \mathbb{P}(Y \neq h(X)) = \mathbb{P}(-Yh(X) \geq 0) = \mathbb{E}\Big[I\{-Yh(X) \geq 0\}\Big] = \mathbb{E}[\varphi_0(-Yh(X))],$$

where $\varphi_0(u) := I(u \geq 0)$. We now replace φ_0 by a convex function $\varphi : \mathbb{R} \to \mathbb{R}$ (sometimes called a *convex surrogate loss*) and define

$$\begin{aligned} R_\varphi(h) &:= \mathbb{E}[\varphi(-Yh(X))], \\ f^*_\varphi &:= \operatorname*{argmin}_{f:\mathcal{X}\to\mathbb{R}} R_\varphi(f), \\ R_{n,\varphi}(h) &:= \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i h(X_i)), \\ \hat{f}_{n,\varphi} &:= \operatorname*{argmin}_{f\in\mathcal{F}} R_{n,\varphi}(f), \end{aligned} \tag{64}$$

where $\mathcal{F}$ is a convex class of functions $f : \mathcal{X} \to \mathbb{R}$. The question is whether there are convex functions φ such that $h^* = \text{sign}(f^*_\varphi)$, where h^* is defined in Theorem 5.17?

Natural requirements to φ are: (i) convexity, (ii) φ should penalize more for wrong classification than for correct classification. Note that φ_0 does not penalize at all for correct classification, because $\varphi_0(-1) = 0$, but it penalizes for wrong classification since $\varphi_0(1) = 1$. However φ_0 is not convex. The first historical example of convex surrogate loss φ is the *hinge loss*:

$$\varphi_H(x) := (1 + x)_+.$$

It satisfies both the requirements (i) and (ii) above. The corresponding risk and its minimizer are

$$R_{\varphi_H}(f) := \mathbb{E}[(1 - Yf(X))_+], \qquad f^*_{\varphi_H} := \operatorname*{argmin}_{f:\mathcal{X}\to\mathbb{R}} R_{\varphi_H}(f).$$

Proposition 5.20. Let h^* be the Bayes classifier, i.e., $h^*(x) := \text{sign}(\eta(x) - 1/2)$. Then $f^*_{\varphi_H} = h^*$.

Proof. Recall that $\eta(x) = \mathbb{P}(Y = 1|X = x)$ and $h^*(x) = \text{sign}(f^*(x))$ with $f^*(x) = \mathbb{E}(Y|X = x) = 2\eta(x) - 1$. We can write

$$R_{\varphi_H}(f) = \int_{\mathcal{X}} \mathbb{E}[(1 - Yf(X))_+|X = x]dP_X(x)$$

where

$$\begin{aligned}\mathbb{E}[(1 - Yf(X))_+|X = x] &= \mathbb{P}(Y = 1|X = x)(1 - f(x))_+ + \mathbb{P}(Y = -1|X = x)(1 + f(x))_+ \\ &= \eta(1 - f(x))_+ + (1 - \eta(x))(1 + f(x))_+.\end{aligned}$$

Fix an arbitrary x and define

$$g(u) := \eta(x)(1 - u)_+ + (1 - \eta(x))(1 + u)_+.$$

We claim that

$$f^*_{\varphi_H}(x) = \operatorname*{argmin}_{u\in\mathbb{R}} g(u).$$

Next, observe that g is a piecewise affine function. Let $u^* = \text{argmin}_{u\in\mathbb{R}} g(u)$. We can see that:

$$g(u) = \begin{cases} \eta(x)(1 - u) + (1 - \eta(x))(1 + u) = 1 + (1 - 2\eta(x))u, & \text{if } |u| \le 1; \\ (1 - \eta(x))(1 + u), & \text{if } u > 1; \\ \eta(x)(1 - u), & \text{if } u < -1. \end{cases}$$

As g is affine for $u > 1$ (and $u < -1$) with nonnegative slope we see that u^* must belong to $[-1, 1]$. However, for $u \in [-1, 1]$, g is minimized at

$$u^* = \begin{cases} -1, & \text{if } \eta(x) \leq 1/2; \\ +1, & \text{if } \eta(x) > 1/2. \end{cases}$$

Thus, $f^*_{\varphi_H}(x) = u^* = \text{sign}(\eta(x) - 1/2) = h^*(x)$ for all $x \in \mathcal{X}$.

□

Classical examples of functions φ are the following: (i) $\varphi(x) = (1 + x)_+$ (hinge loss); (ii) $\varphi(x) = \exp(x)$ (exponential loss); (iii) $\varphi(x) = \log_2(1 + \exp(x))$ (logistic loss).

Proposition 5.21. Let φ' be positive and strictly increasing. Then $h^* = \text{sign}(f^*_\varphi)$.

Proof. Exercise (HW3). □

Given a solution $\hat{f}_{n,\varphi}$ to the minimization problem (64), we define the classifier $\hat{h}_{n,\varphi} := \text{sign}(\hat{f}_{n,\varphi})$. A popular choice for the set $\mathcal{F}$ is

$$\mathcal{F} = \left\{ \sum_{j=1}^{M} \theta_j h_j : \theta \in \Theta \right\}$$

where $\{h_1, \ldots, h_M\}$ is a dictionary of classifiers independent of the data $\{(X_i, Y_i)\}_{i=1}^n$ (the h_j's are often called the "weak learners"), and $\Theta \subset \mathbb{R}^M$ is a set of coefficients where $\Theta = \mathbb{R}^M$ or Θ is an ℓ_1-body or an ℓ_2-body as defined as follows.

- An ℓ_2-body is a set of the form

  $$\Theta = \left\{ \theta \in \mathbb{R}^M : \theta^\top K \theta \leq r \right\}$$

  for some symmetric positive semi-definite matrix K and a positive scalar r.

- An ℓ_1-body is either an ℓ_1-ball $\Theta = \{\theta \in \mathbb{R}^M : |\theta|_1 \leq r\}$, or the simplex

  $$\Theta = \Lambda^M = \left\{ \theta \in \mathbb{R}^M : \sum_{j=1}^{M} \theta_j = 1, \theta_j \geq 0 \right\}.$$

The hinge loss with an ℓ_2-body yields *support vector machines (SVM)*. The exponential and logit loss with an ℓ_1-body leads to *boosting*.

5.4.4 Support vector machine (SVM): definition

Suppose that $\mathcal{H}$ is a RKHS of functions on $\mathcal{X}$ with kernel K.

Consider the classification problem described in the previous subsections. A popular example of a convex set of functions $\mathcal{F}$ used in (64) is a ball in the RKHS $\mathcal{H}$:

$$\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq r\}, \quad r > 0.$$

Then (64) becomes

$$\min_{f \in \mathcal{H}: \|f\|_{\mathcal{H}} \leq r} R_{n,\varphi}(f).$$

The *support vector machine* is, by definition, a classifier obtained from solving this problem when $\varphi(x) = (1+x)_+$ (the Hinge loss):

$$\min_{f \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \lambda \|f\|_{\mathcal{H}}^2 \right). \tag{65}$$

Thus, by the representer theorem (see Theorem 5.11), it is enough to look for a solution of (65) in the finite dimensional space S (see the proof of Theorem 5.11) of dimension less than or equal to n. Solving the problem reduces to finding the coefficients θ_j in the representation (62).

Let $K_{ij} = K(X_i, X_j)$ and denote by K the symmetric matrix $(K_{ij})_{i,j=1,\ldots,n}$. Then for any $f \in S \subset \mathcal{H}$,

$$\|f\|_{\mathcal{H}}^2 = \sum_{i,j=1,\ldots,n} \theta_i \theta_j K_{ij} = \theta^\top K \theta.$$

Thus, the SVM minimization problem (65) reduces to

$$\min_{\theta \in \mathbb{R}^n} \left[\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i (K\theta)_i)_+ + \lambda \theta^\top K \theta \right], \tag{66}$$

where $(K\theta)_i$ is the i'th component of $K\theta$. Given the solution $\hat{\theta}$ of (66), the SVM classifier $\hat{h}_{n,\varphi}$ is determined as:

$$\hat{h}_{n,\varphi} = \text{sign}(\hat{f}_{n,\varphi}(x)), \qquad \text{where} \qquad \hat{f}_{n,\varphi}(x) = \sum_{i=1}^{n} \hat{\theta}_i K(X_i, x). \tag{67}$$

5.4.5 Analysis of the SVM minimization problem

Traditionally, the SVM minimization problem (66) is solved by reduction to a quadratic program after introducing some additional slack variables. Here, we choose to treat

the problem differently, using subdifferential calculus. For any convex objective function G, we have the equivalence

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^M} G(\theta) \Longleftrightarrow 0 \in \partial G(\hat{\theta}) \tag{68}$$

where $\partial G(\theta)$ is the subdifferential of G at θ. (In the particular case where G is differentiable at θ, the subdifferential is reduced to the gradient of G at θ: $\partial G(\hat{\theta}) = \{\nabla G(\theta)\}$.)

Proposition 5.22. The solution of the SVM optimization problem (66) has the form

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\theta}_i K(X_i, x),$$

where the coefficients $\hat{\theta}_i$ satisfy

$$\begin{aligned}
\hat{\theta}_i &= 0, && \text{if } Y_i \hat{f}(X_i) > 1, \\
\hat{\theta}_i &= \frac{Y_i}{2\lambda n}, && \text{if } Y_i \hat{f}(X_i) < 1, \\
\hat{\theta}_i &= \frac{\alpha_i Y_i}{2\lambda n}, && \text{with } \alpha_i \in [0,1], \qquad \text{if } Y_i \hat{f}(X_i) = 1.
\end{aligned}$$

The points X_i with $\hat{\theta}_i \neq 0$ are called the support vectors.

In practice, there are often not too many support vectors since only the points X_i that are misclassified or close to the decision boundary satisfy the condition $\hat{\theta}_i \neq 0$.

Proof. We will derive the expression for the subdifferential of the objective function in (66) by analyzing each term in the sum separately. Fix some index i and consider the function

$$\theta \mapsto \left(1 - Y_i \sum_{j=1}^{n} K_{ij}\theta_j\right)_+ = (1 - Y_i (K\theta)_i)_+.$$

Let $g_i(\theta)$ be a subgradient of this function and denote by $g_{ij}(\theta)$ its j'th component. There are three cases that follow immediately from the form of the subdifferential of the function $(1-x)_+$:

- if $Y_i(K\theta)_i > 1$ then $g_{ij}(\theta) = 0$,
- if $Y_i(K\theta)_i < 1$ then $g_{ij}(\theta) = -Y_i K_{ij}$,
- if $Y_i(K\theta)_i = 1$ then $g_{ij}(\theta) = -\alpha_i Y_i K_{ij}$, for some $\alpha_i \in [0,1]$.

We can wrap these three cases as $g_{ij}(\theta) = -\alpha_i Y_i K_{ij}$, with

(i) $\alpha_i = 0$ if $Y_i(K\theta)_i > 1$,

(ii) $\alpha_i = 1$ if $Y_i(K\theta)_i < 1$,

(iii) $\alpha_i \in [0,1]$ if $Y_i(K\theta)_i = 1$.

Consequently, the subdifferential $\partial\Big(\frac{1}{n}\sum_{i=1}^n (1 - Y_i(K\theta)_i)_+\Big)$ is composed of vectors of the form $-K\beta$ here $\beta = \Big(\frac{1}{n}\alpha_i Y_i\Big)_{i=1}^n$ with α_i satisfying the above conditions (i)–(iii). Next, the function $\lambda\theta^\top K\theta$ is differentiable and its gradient is $2\lambda K\theta$. Thus, the subdifferential of the objective function in (66) is composed of vectors of the form

$$-K\beta + 2\lambda K\theta.$$

Now, by (68), a vector $\hat\theta$ is a solution of (66) if and only if 0 belongs to the subdifferential of the objective function at $\hat\theta$, which can be written as $2\lambda\hat\theta - \beta = \epsilon$ for some ϵ satisfying $K\epsilon = 0$. It remains to note that we can always take $\epsilon = 0$ since the choice of ϵ in the null space of K does not modify the value of the objective function. This completes the proof. □

Observe that the SVM solution can be written as (67). Thus, if we consider the functions $\varphi_i(\cdot) = K(X_i, \cdot)$, we have

$$\hat f = \sum_{i=1}^n \hat\theta_i \varphi_j$$

so that $\hat f$ can be viewed as a linear classifier in "transformed coordinates". The functions $\varphi_i(\cdot) = K(X_i, \cdot)$ can be interpreted as "weak learners" but in this case they are not classifiers.

The strength of the RKHS approach is that the space $\mathcal{X}$ can be any arbitrary space (such as a graph or a semi-group, for example) but we transform each point $X_i \in \mathcal{X}$ into an finite-dimensional vector $Z_i = (\varphi_1(X_i), \ldots, \varphi_n(X_i))^\top \in \mathbb{R}^n$, and then use a linear classifier $\hat f(X) = \theta^\top Z$ in the finite-dimensional space $\mathbb{R}^n$ where $Z := (\varphi_1(X), \ldots, \varphi_n(X))^\top \in \mathbb{R}^n$. The classification rule for a new point Z is

$$\hat Y := \begin{cases} 1, & \text{if } \hat\theta^\top Z > 0, \\ -1, & \text{otherwise.} \end{cases}$$

For any learning point Z_i, if Z_i is correctly classified we have $Y_i\hat\theta^\top Z_i > 0$, and if Z_i is wrongly classified we have $Y_i\hat\theta^\top Z_i \le 0$. By Proposition 5.22 a solution $\hat\theta$ of the SVM minimization problem has the coordinates $\hat\theta_i$, $i = 1, \ldots, n$, satisfying:

- $\hat{\theta}_i = 0$ if $Y_i \hat{\theta}^\top Z_i > 1$. Interpretation: The point (Z_i, Y_i) does not affect the classification rule if Z_i is correctly classified with high margin (larger than 1), where the *margin* of the i'th observation is defined as $Y_i \hat{\theta}^\top Z_i = Y_i \hat{f}(X_i)$.
- $\hat{\theta}_i \neq 0$ if $Y_i \hat{\theta}^\top Z_i \leq 1$. The last inequality means that the point Z_i is wrongly classified or correctly classified with small margin (smaller than 1). If $\hat{\theta}_i \neq 0$, the point Z_i is called a *support vector*.

5.5 Kernel ridge regression

Consider the regression model

$$Y_i = f^*(x_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where f^* is the true regression function, x_i's take values in $\mathcal{X}$ (an arbitrary metric space), $\epsilon_1, \ldots, \epsilon_n$ are mean zero, uncorrelated random variables. We want to estimate f^* by minimizing the criterion function

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{H}} \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2. \tag{69}$$

By the representer theorem, we can claim that any solution to (69) is of the form $\hat{f}(\cdot) = \sum_{i=1}^n \alpha_i K(x_i, \cdot)$ for some weight vector $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. Thus, the above optimization problem can be equivalently expressed as:

$$\hat{\alpha} := \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \|Y - K\alpha\|^2 + \lambda \alpha^\top K \alpha,$$

where $K = ((K_{ij}))$ with $K_{ij} = K(x_i, x_j)$, $Y = (Y_1, \ldots, Y_n)$. Here we have used that for $\hat{f}$ in the span of $\{K(x_i, \cdot)\}_{i=1}^n$ and thus $f(x_i) = (K\alpha)_i$, $\|f\|_{\mathcal{H}}^2 = \alpha^\top K \alpha$. We can solve the above finite dimensional optimization problem to yield

$$K(K + \lambda I)\alpha = KY,$$

which shows that $\hat{\alpha} = (K + \lambda I)^{-1} Y$ is a solution.

5.6 Kernel principal component analysis (PCA)

Given a random vector $X \sim P$ in $\mathbb{R}^d$, PCA solves the following eigenvector (eigenvalue) problem:

$$\Sigma v = \lambda v,$$

where Σ is the covariance matrix of X and the eigenvector corresponding to the i'th largest eigenvalue is the i'th principal component, for $i = 1, \ldots, n$. Another way to view the PCA problem is to first consider the first principal component, which is a solution to the following optimization problem:

$$v_1 = \operatorname*{argmax}_{v \in \mathbb{R}^d : \|v\| \le 1} \operatorname{Var}(v^\top X) = \operatorname*{argmax}_{v \in \mathbb{R}^d : \|v\| \le 1} v^\top \Sigma v.$$

The second principal component is defined as the unit vector that maximizes $\operatorname{Var}(v^\top X)$ over all vectors v that are orthogonal to v_1, and so on.

Given i.i.d. samples $\{x_i\}_{i=1}^n$ from P, the sample principal components are obtained by solving the corresponding sample analogue:

$$\hat{\Sigma} v = \lambda v,$$

where $\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$ (here $\bar{x} = \sum_{i=1}^n x_i / n$ is the sample mean) is the sample covariance matrix of X. Similarly,

$$\hat{v}_1 = \operatorname*{argmax}_{v \in \mathbb{R}^d : \|v\| \le 1} v^\top \hat{\Sigma} v = \operatorname*{argmax}_{v \in \mathbb{R}^d : \|v\| \le 1} \frac{1}{n} \sum_{i=1}^n \left[v^\top (x_i - \bar{x}) \right]^2. \tag{70}$$

Now suppose that $X \sim P$ takes values in an arbitrary metric space $\mathcal{X}$. Suppose that $\mathcal{H}$ is a RKHS (of functions) on $\mathcal{X}$ with reproducing kernel K. We can use the kernel method to extent classical PCA to capture non-linear principal components. The first principal component can now be defined as

$$f_1 := \operatorname*{argmax}_{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le 1} \operatorname{Var}(f(X)) = \operatorname*{argmax}_{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le 1} \operatorname{Var}(\langle f, K(X, \cdot) \rangle_{\mathcal{H}}).$$

Let $\varphi(x) := K(x, \cdot)$ for all $x \in \mathcal{X}$ (note that here φ is not exactly the feature map, as $\varphi : \mathcal{X} \to \mathcal{H}$). Given a sample $\{x_i\}_{i=1}^n$ from P, the sample first principal component (function) can be defined analogously (as in (70)) as

$$\hat{f}_1 = \operatorname*{argmax}_{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le 1} \widehat{\operatorname{Var}}(\langle f, K(X, \cdot) \rangle_{\mathcal{H}}) = \operatorname*{argmax}_{\|f\|_{\mathcal{H}} \le 1} \frac{1}{n} \sum_{i=1}^n \left[\left\langle f, \varphi(x_i) - \frac{1}{n} \sum_{j=1}^n \varphi(x_j) \right\rangle_{\mathcal{H}} \right]^2.$$

We define the *empirical covariance operator* as

$$\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^n \tilde{\varphi}(x_i) \otimes \tilde{\varphi}(x_i), \qquad \text{where } \tilde{\varphi}(x_i) := \varphi(x_i) - \frac{1}{n} \sum_{j=1}^n \varphi(x_j).$$

We would like to find eigenfunctions $\hat{f}$ (the principal components) (Why? Exercise (HW3)) such that

$$\hat{\Sigma}(\hat{f}) = \lambda \hat{f}. \tag{71}$$

The question now is, how do we express the above equation in terms of kernels, i.e., how do we "kernelize" it? Towards this end, we make the following claim.

Claim: Any solution to (71) is of the form $\hat{f} = \sum_{i=1}^{n} \alpha_i \tilde{\varphi}(x_i)$ for some weight vector $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$.

Proof: First, we observe that any solution to (71) lies in Range($\hat{\Sigma}$). Linearity, and the nature of $\tilde{\varphi}(x_i) \otimes \tilde{\varphi}(x_i)$ (by definition $(a \otimes b)(c) := \langle b, c \rangle_{\mathcal{H}} a$) tell us that

$$\hat{\Sigma}(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \langle \tilde{\varphi}(x_i), \hat{f} \rangle_{\mathcal{H}} \, \tilde{\varphi}(x_i).$$

Therefore, (71) is equivalent to the following system of equations in $\alpha \in \mathbb{R}^n$:

$$\hat{\Sigma}\left(\sum_{i=1}^{n} \alpha_i \tilde{\varphi}(x_i) \right) = \lambda \sum_{i=1}^{n} \alpha_i \tilde{\varphi}(x_i).$$

For the above set of equations, we left-hand side equals

$$\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{H}} \, \tilde{\varphi}(x_j).$$

Using the fact that $\langle \tilde{\varphi}(x_j), \tilde{\varphi}(x_j) \rangle_{\mathcal{H}} = \tilde{K}(x_i, x_j)$, where $\tilde{K} = HKH$ (show this; Exercise (HW3); here $H = I_n - \mathbf{1}_{n \times n}/n$ and K is the Gram matrix, i.e., $K_{ij} = K(x_i, x_j)$) the above system of equations may be written as

$$\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_i \tilde{K}_{ij} \, \tilde{\varphi}(x_j) = \lambda \sum_{i=1}^{n} \alpha_i \tilde{\varphi}(x_i).$$

Taking inner products with $\varphi(x_l)$, for $l = 1, \ldots, n$, we get

$$\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_i \tilde{K}_{ij} \tilde{K}_{jl} = \lambda \sum_{i=1}^{n} \alpha_i \tilde{K}_{il}.$$

We now have a set of n linear equations in the vector $\alpha \in \mathbb{R}^n$. In matrix-vector form, it can be written very simply as

$$\tilde{K}^2 \alpha = \lambda n \tilde{K} \alpha.$$

The only solutions of this equation that are of interest to us are those that satisfy

$$\tilde{K} \alpha = \lambda n \alpha.$$

This is simply an eigenvalue/eigenvector problem in the matrix K.

6 Bootstrap

Suppose that we have data $\mathbf{X} \sim P$, and $\theta \equiv \theta(P)$ is a parameter of interest. Let $\hat{\theta} \equiv \hat{\theta}(\mathbf{X})$ be an estimator of θ. Suppose that we would want to construct a level-$(1-2\alpha)$ *confidence interval* for θ, i.e., find κ_α and $\kappa_{1-\alpha}$ such that

$$\mathbb{P}(\hat{\theta} - \kappa_\alpha \leq \theta \leq \hat{\theta} + \kappa_{1-\alpha}) \geq 1 - 2\alpha. \tag{72}$$

How do we find (estimate) κ_α and $\kappa_{1-\alpha}$ in such a general setting?

Problem: The distribution of $\hat{\theta} - \theta$ depends on P and might be unknown. Even if we know the asymptotics (e.g., asymptotically normal), we may want more accurate quantiles for a fixed sample size. In some situations, the asymptotic limiting distribution can depend on *nuisance* parameters that can be hard to estimate.

In these situations we can use the *bootstrap*.

To motivate the bootstrap method, let us consider the following simple scenario. Suppose that we model our data $\mathbf{X} = (X_1, \ldots, X_n)$ as a random sample from some distribution $P \in \mathcal{P}$, where $\mathcal{P}$ is a class of probability distributions. Let $\eta(\mathbf{X}, P)$ be a *root*, i.e., a random variable that possibly depends on both the distribution P and the sample $\mathbf{X}$ drawn from P (e.g., think of $\eta(\mathbf{X}, P)$ as $\sqrt{n}(\bar{X}_n - \mu)$, where $\bar{X}_n = \sum_{i=1}^n X_i/n$ and $\mu = \mathbb{E}(X_1)$). In fact, $\hat{\theta} - \theta$ (as described above) is a root.

In general, we may wish to estimate the mean or a quantile or some other probabilistic feature or the entire *distribution* of $\eta(\mathbf{X}, P)$. As mentioned above, the distribution of $\hat{\theta} - \theta$ depends on P and is thus unknown. Let $H_n(x, P)$ denote the c.d.f. of $\eta(\mathbf{X}, P)$, i.e.,

$$H_n(x, P) := \mathbb{P}_P(\eta(\mathbf{X}, P) \leq x). \tag{73}$$

Of course, if we can estimate $H_n(\cdot, P)$ then we can use this to construct CIs, test hypotheses; e.g., if $\eta(\mathbf{X}, P) = (\hat{\theta} - \theta)$ then being able to estimate $H_n(\cdot, P)$ immediately yields estimates of κ_α and $\kappa_{1-\alpha}$ as defined in (72).

Idea: What if we knew P and could draw unlimited replicated samples from P?

In that case we could approximate $H_n(x, P)$ as follows: draw repeated samples from P resulting in a series of values for the root $\eta(\mathbf{X}, P)$, then we could form an estimate of $H_n(x, P)$ by counting how many of the $\eta(\mathbf{X}, P)$'s are $\leq x$.

But, of course, we do not know P. However we can *estimate* P by $\hat{P}_n$ and use the above idea. This is the notion of bootstrap.

Definition 6.1 (Bootstrap)**.** The bootstrap is a method of replacing (plug-in) the unknown distribution P with a known distribution $\hat{P}_n$ (estimated from the data) in probability/expectation calculations.

The bootstrap approximation of $H_n(\cdot, P)$ is $\hat{H}_n(\cdot, \hat{P}_n)$, where $\hat{P}_n$ is an estimator of P obtained from the observed data (that we think is close to P), i.e.,

$$\hat{H}_n(x, \hat{P}_n) := \mathbb{P}^*_{\hat{P}_n}(\eta(\mathbf{X}^*, \hat{P}_n) \leq x | \mathbf{X}). \tag{74}$$

where $\mathbb{P}^*_{\hat{P}_n}(\cdot|\mathbf{X})$ is the conditional probability given the observed data $\mathbf{X}$ (under the estimated $\hat{P}_n$). Thus, bootstrap estimates the distribution of $\eta(\mathbf{X}, P)$ by that of $\eta(\mathbf{X}^*, \hat{P}_n)$, where $\mathbf{X}^*$ is a random sample (conditional on the data) drawn from the distribution $\hat{P}_n$. The idea is that

$$\text{if } \hat{P}_n \approx P, \qquad \text{then} \quad \hat{H}_n(\cdot, \hat{P}_n) \approx H_n(\cdot, P).$$

Question: How do we find $\hat{H}_n(\cdot, \hat{P}_n)$, the distribution of $\eta(\mathbf{X}^*, \hat{P}_n)$?

Answer: In most cases, the distribution of $\eta(\mathbf{X}^*, \hat{P}_n)$ is difficult to analytically compute, but we can *always* be approximated easily by Monte Carlo *simulations*.

The bootstrap can be broken down in the following simple steps:

- Find a "good" estimator $\hat{P}_n$ of P.
- Draw a large number (say, B) of random samples $\mathbf{X}^{*(1)}, \ldots, \mathbf{X}^{*(B)}$ from the distribution $\hat{P}_n$ and then compute $T^{*(j)} := \eta(\mathbf{X}^{*(j)}, \hat{P}_n)$, for $j = 1, \ldots, B$.
- Finally, compute the desired feature of $\eta(\mathbf{X}^*, \hat{P}_n)$ using the empirical c.d.f. $\tilde{H}^B_n(\cdot, \hat{P}_n)$ of the values $T^{*(1)}, \ldots, T^{*(B)}$, i.e.,

$$\tilde{H}^B_n(x, \hat{P}_n) := \frac{1}{B} \sum_{j=1}^{B} I\{T^{*(j)} \leq x\}, \qquad \text{for } \ x \in \mathbb{R}.$$

Intuitively,

$$\tilde{H}^B_n(\cdot, \hat{P}_n) \ \approx \ \hat{H}_n(\cdot, \hat{P}_n) \ \approx \ H_n(\cdot, P),$$

where the first approximation is from Monte Carlo error (and can be as small as we would like, by taking B as large as we want) and the second approximation is due to the bootstrap method. If $\hat{P}_n$ is a good approximation of P, then bootstrap can be successful.

6.1 Parametric bootstrap

In parametric models it is more natural to take $\hat{P}_n$ as the fitted parametric model.

Example 6.2 (Estimating the standard deviation of a statistic)**.** Suppose that $X_1, \ldots, X_n$ is random sample from $N(\mu, \sigma^2)$. Suppose that we are interested in the parameter

$$\theta = \mathbb{P}(X \leq c) = \Phi\left(\frac{c-\mu}{\sigma}\right),$$

where c is a given known constant. A natural estimator of θ is its MLE $\hat{\theta}$:

$$\hat{\theta} = \Phi\left(\frac{c-\bar{X}}{\hat{\sigma}}\right),$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

Question: How do we estimate the standard deviation of $\hat{\theta}$? There is no easy closed form expression for this.

Solution: We can bootstrap!

Draw many (say B) bootstrap samples of size n from

$$N(\bar{X}, \hat{\sigma}^2) \equiv \hat{P}_n.$$

For the j'th *bootstrap sample* we compute a sample average $\bar{X}^{*(j)}$, a sample standard deviation $\hat{\sigma}^{*(j)}$. Finally, we compute

$$\hat{\theta}^{*(j)} = \Phi\left(\frac{c-\bar{X}^{*(j)}}{\hat{\sigma}^{*(j)}}\right).$$

We can estimate the mean of $\hat{\theta}$ by $\bar{\theta}^* = \frac{1}{B}\sum_{j=1}^B \hat{\theta}^{*(j)}$. The standard deviation of $\hat{\theta}$ can then be estimated by the bootstrap standard deviation of the $\hat{\theta}^{*(j)}$ values, i.e.,

$$\left[\frac{1}{B}\sum_{j=1}^B (\hat{\theta}^{*(j)} - \bar{\theta}^*)^2\right]^{1/2}.$$

Example 6.3 (Comparing means when variances are unequal)**.** Suppose that we have two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from two possibly different normal populations. Suppose that

$$X_1, \ldots, X_m \text{ are i.i.d. } N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad Y_1, \ldots, Y_n \text{ are i.i.d. } N(\mu_2, \sigma_2^2).$$

Suppose that we want to test

$$H_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 \neq \mu_2.$$

We can use the test statistic

$$U = \frac{(m+n-2)^{1/2}(\bar{X}_m - \bar{Y}_n)}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}(S_X^2 + S_Y^2)^{1/2}},$$

where $\bar{X}_m = \frac{1}{m}\sum_{i=1}^m X_i$, $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$, $S_X^2 = \sum_{i=1}^m (X_i - \bar{X}_m)^2$ and $S_Y^2 = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$. Note that as $\sigma_1^2 \neq \sigma_2^2$, U does not necessarily follow a t-distribution.

Question: How do we find the critical value of this test?

The parametric bootstrap can proceed as follows:

First choose a large number B, and for $j = 1, \ldots, B$, simulate $(\bar{X}_m^{*(j)}, \bar{Y}_n^{*(j)}, S_X^{2*(j)}, S_Y^{2*(j)})$, where all four random variables are independent with the following distributions:

- $\bar{X}_m^{*(j)} \sim N(0, \hat{\sigma}_X^2/m)$,
- $\bar{Y}_n^{*(j)} \sim N(0, \hat{\sigma}_Y^2/n)$,
- $S_X^{2*(j)} \sim \hat{\sigma}_X^2\, \chi_{m-1}^2$,
- $S_Y^{2*(j)} \sim \hat{\sigma}_Y^2\, \chi_{n-1}^2$,

where $\hat{\sigma}_X^2 = S_X^2/(m-1)$ and $\hat{\sigma}_Y^2 = S_Y^2/(n-1)$. Then we compute

$$U^{*(j)} = \frac{(m+n-2)^{1/2}(\bar{X}_m^{*(j)} - \bar{Y}_n^{*(j)})}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}(S_X^{2*(j)} + S_Y^{2*(j)})^{1/2}}$$

for each j. We approximate the null distribution of U by the empirical distribution of the $\{U^{*(j)}\}_{j=1}^B$. Let c_n^* be the $\left(1-\frac{\alpha}{2}\right)$-quantile of the empirical distribution of $\{U^{*(j)}\}_{j=1}^B$. Then, we can reject H_0 if

$$|U| > c_n^*.$$

6.2 The nonparametric bootstrap

In problems where the distribution P is not indexed by a parametric family, a natural estimator of P is the empirical distribution $\hat{P}_n$ given by the distribution that puts $1/n$-mass at each of the observed data points.

Example 6.4. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample from a distribution F on $\mathbb{R}$. Suppose that we want a CI for the median θ of F. We can base a CI on the sample median M.

We want to estimate the distribution of $M - \theta$. Let $\eta(\mathbf{X}, F) := M - \theta$. We may choose $\hat{F} = F_n$, the empirical distribution function of the observed data. Thus, our method can be broken in the following steps:

- Choose a large number B and simulate many samples $\mathbf{X}^{*(j)}$, for $j = 1, \ldots, B$, (conditionally i.i.d. given the data) from F_n. This reduces to drawing *with replacement sampling* from $\mathbf{X}$.
- For each bootstrap sample we compute the sample median $M^{*(j)}$ and then find the appropriate sample quantiles of $\{M^{*(j)} - M\}_{i=1}^{B}$. Observe that $\eta(\mathbf{X}^*, F_n) = M^* - M$.

6.3 Consistency of the bootstrap

Suppose that $\hat{F}_n$ and F are the corresponding c.d.f.'s for $\hat{P}_n$ and P respectively. Suppose that $\hat{P}_n$ is a consistent estimator of P. This means that at each x in the support of X_1 where $F(x)$ is continuous, $\hat{F}_n(x) \to F(x)$ in probability or a.s. as $n \to \infty$[10]. If, in addition, $\hat{H}_n(x, P)$, considered as a functional of P, is *continuous* in an appropriate sense, it can be expected that $\hat{H}_n(x, \hat{P}_n)$ will be close to $H_n(x, P)$, when n is large.

Observe that $\hat{H}_n(x, \hat{P}_n)$ is a *random distribution function* (as it depends on the observed data). Let ρ be any notion of distance between two probability distributions that metrizes weak convergence, i.e., for any sequence of c.d.f.'s $\{G_n\}_{n \geq 1}$, we have

$$G_n \xrightarrow{d} G \qquad \text{if and only if} \qquad \rho(G_n, G) \to 0 \ \text{ as } \ n \to \infty.$$

In particular, we can take ρ to be the Prohorov metric[11] or the Levy metric[12]. For simplicity, we can also use the uniform distance (Kolmogorov metric) between G_n and G (which metrizes weak convergence if G is a continuous c.d.f.).

Definition 6.5. We say that the bootstrap is *weakly consistent* under ρ for $\eta(\mathbf{X}_n, P)$ if

$$\rho(H_n, \hat{H}_n) \xrightarrow{p} 0 \qquad \text{as } \ n \to \infty,$$

where H_n and $\hat{H}_n$ are defined in (73) and (74) respectively. We say that the bootstrap

[10] If F is a continuous c.d.f., then it follows from Polya's theorem that $\hat{F}_n \to F$ in probability or a.s. uniformly over x. Thus, $\hat{F}_n$ and F are uniformly close to one another if n is large.

[11]

[12]

is *strongly consistent* under ρ for $\eta(\mathbf{X}_n, P)$ if

$$\rho(H_n, \hat{H}_n) \overset{a.s.}{\to} 0 \qquad \text{as} \;\; n \to \infty.$$

In many problems, it can be shown that $H_n(\cdot, P)$ converges in distribution to a limit $H(\cdot, P)$. In such situations, it is much easier to prove that the bootstrap is consistent by showing that

$$\rho(\hat{H}_n, H) \overset{a.s./p}{\to} 0 \qquad \text{as} \;\; n \to \infty.$$

In applications, e.g., for construction of CIs, we are quite often interested in approximating the quantiles of H_n by that of $\hat{H}_n$ (as opposed to the actual c.d.f.). The following simple result shows that weak convergence, under some mild conditions, implies the convergence of the quantiles.

Exercise (HW4): Let $\{G_n\}_{n\geq 1}$ be a sequence of distribution functions on the real line converging weakly to a distribution function G, i.e., $G_n(x) \to G(x)$ at all continuity points x of G. Assume that G is continuous and strictly increasing at $y = G^{-1}(1-\alpha)$. Then,

$$G_n^{-1}(1-\alpha) := \inf\{x \in \mathbb{R} : G_n(x) \geq 1-\alpha\} \to y = G^{-1}(1-\alpha).$$

The following theorem, although quite obvious, gives us a general strategy to prove the consistency of the bootstrap in many problems.

Theorem 6.6. Let $\mathcal{C}_{\mathcal{P}}$ be a set of sequences $\{P_n \in \mathcal{P}\}_{n\geq 1}$ containing the sequence $\{P, P, \ldots\}$. Suppose that, for every sequence $\{P_n\} \in \mathcal{C}_{\mathcal{P}}$, $H_n(\cdot, P_n)$ converges weakly to a common limit $H(\cdot, P)$. Let $\mathbf{X}_n$ be a sample of size n from P. Assume that $\hat{P}_n$ is an estimator of P based on $\mathbf{X}_n$ such that $\{\hat{P}_n\}$ falls in $\mathcal{C}_{\mathcal{P}}$ with probability one. Then,

$$\rho(H_n(\cdot, P), \hat{H}_n(\cdot, \hat{P}_n)) \overset{a.s.}{\to} 0 \qquad \text{as} \;\; n \to \infty.$$

If $H(\cdot, P)$ is continuous and strictly increasing at $H^{-1}(1-\alpha, P)$ $(0 < \alpha < 1)$, then

$$\hat{H}_n^{-1}(1-\alpha, \hat{P}_n) \overset{a.s.}{\to} H(1-\alpha, P) \qquad \text{as} \;\; n \to \infty.$$

Further, if $H(x, P)$ is continuous in x, then

$$K(\hat{H}_n, H_n) := \sup_{x\in\mathbb{R}} |\hat{H}_n(x, \hat{P}_n) - H_n(x, P)| \overset{a.s.}{\to} 0 \qquad \text{as} \;\; n \to \infty.$$

The proof of the above theorem is also left as an exercise (HW4).

Remark 6.1. Often, the set of sequences $\mathcal{C}_{\mathcal{P}}$ can be described as the set of sequences $\{P_n\}_{n\geq 1}$ such that $d(P_n, P) \to 0$, where d is an appropriate "metric" on the space of probabilities. Indeed, one should think of $\mathcal{C}_{\mathcal{P}}$ as a set of sequences $\{P_n\}$ that are converging to P in an appropriate sense. Thus, the convergence of $H_n(\cdot, P_n)$ to $H(\cdot, P)$ is locally uniform in a specified sense. Unfortunately, the appropriate metric d will depend on the precise nature of the problem and the choice of the root.

Theorem 6.6 essentially says that to prove the consistency of the bootstrap it is enough to try to understand the limiting behavior of $H_n(\cdot, P_n)$, where P_n is any sequence of distributions "converging" (in some appropriate sense) to P. Thus, quite often, showing the consistency of the bootstrap boils down to showing the weak convergence of $\eta(\mathbf{X}_n, P_n)$ under a triangular array setup, as $\mathbf{X}_n$ is now an i.i.d. sample from P_n. For example, if the CLT plays a crucial role in proving that $H_n(\cdot, P)$ converges weakly to a limit $H(\cdot, P)$, the Lindeberg-Feller CLT theorem can be used to show that $H_n(\cdot, P_n)$ converges weakly to $H(\cdot, P)$.

Theorem 6.7 (Bootstrapping the sample mean)**.** Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. F and that $\sigma^2 := \mathrm{Var}_F(X_1) < \infty$. Let $\eta(\mathbf{X}, F) := \sqrt{n}(\bar{X}_n - \mu)$, where $\mu := \mathbb{E}_F(X_1)$ and $\bar{X}_n := \sum_{i=1}^n X_i/n$. Then,

$$K(\hat{H}_n, H_n) = \sup_{x\in\mathbb{R}} |H_n(x) - \hat{H}_n(x)| \xrightarrow{p} 0 \qquad \text{as } \ n \to \infty,$$

where $\hat{H}_n(x) \equiv \hat{H}_n(x, F_n)$ and F_n is the empirical c.d.f. of the sample $X_1, X_2, \ldots, X_n$.

Exercise (HW4): Show that foror almost all sequences $\mathbf{X} = \{X_1, X_2, \ldots\}$, the conditional distribution of $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$, given $\mathbf{X}$, converges in law to $N(0, \sigma^2)$ by the triangular array CLT (Lindeberg CLT).

Exercise (HW4): Show the following joint (unconditional) asymptotic distribution:

$$\left(\sqrt{n}(\bar{X}_n - \mu), \sqrt{n}(\bar{X}_n^* - \bar{X}_n)\right) \xrightarrow{d} (Z_1, Z_2),$$

where Z_1, Z_2 are i.i.d. $N(0, \sigma^2)$. In fact, a more general version of the result is true. Suppose that (U_n, V_n) is a sequence of random vectors such that $U_n \xrightarrow{d} Z \sim H$ (some Z) and $V_n|U_n \xrightarrow{d} Z$ (the same Z) a.s. Then $(U_n, V_n) \xrightarrow{d} (Z_1, Z_2)$, where Z_1, Z_2 are i.i.d. H.

Exercise (HW4): What do you think would be the limiting behavior of $\sqrt{n}(\bar{X}_n^* - \mu)$, conditional on the data $\mathbf{X}$?

6.4 Second-order accuracy of the bootstrap

One philosophical question about the use of the bootstrap is whether the bootstrap has any advantages at all when a CLT is already available. To be specific, suppose that $\eta(\mathbf{X}, F) = \sqrt{n}(\bar{X}_n - \mu)$. If $\sigma^2 := \mathrm{Var}_F(X_1) < \infty$, then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \qquad \text{and} \qquad K(\hat{H}_n, H_n) \xrightarrow{p} 0 \qquad \text{as } n \to \infty.$$

So two competitive approximations to $H_n(x)$ are $\Phi(x/\hat{\sigma}_n)$ (where $\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$) and $\hat{H}_n(x, F_n)$. It turns out that, for certain types of statistics, the bootstrap approximation is (theoretically) more *accurate* than the approximation provided by the CLT. Because any normal distribution is symmetric, the CLT cannot capture information about the skewness in the finite sample distribution of $\eta(\mathbf{X}, F)$. The bootstrap approximation does so. So the bootstrap succeeds in correcting for skewness, just as an Edgeworth expansion[13] would do. This is called Edgeworth correction by the bootstrap, and the property is called *second-order accuracy of the bootstrap.*

Theorem 6.8 (Second-order accuracy)**.** Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. F and that $\sigma^2 := \mathrm{Var}_F(X_1) < \infty$. Let $\eta(\mathbf{X}, F) := \sqrt{n}(\bar{X}_n - \mu)/\sigma$, where $\mu := \mathbb{E}_F(X_1)$ and $\bar{X}_n := \sum_{i=1}^n X_i/n$. If $\mathbb{E}_F|X_1|^3 < \infty$ and F is continuous, then,

$$K(\hat{H}_n, H_n) = o_p(n^{-1/2}) \qquad \text{as } n \to \infty,$$

where $\hat{H}_n(x) \equiv \hat{H}_n(x; F_n)$ is the c.d.f. of $\eta(\mathbf{X}^*, F_n) := \sqrt{n}(\bar{X}_n^* - \bar{X}_n)/\hat{\sigma}$ ($\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$) and F_n is the empirical c.d.f. of the sample $X_1, X_2, \ldots, X_n$.

Remark 6.2 (Rule of thumb)**.** Let $X_1, X_2, \ldots, X_n$ are i.i.d. F and $\eta(\mathbf{X}, F)$ be a root. If $\eta(\mathbf{X}, F) \xrightarrow{d} N(0, \tau^2)$, where τ does not dependent of F, then second-order accuracy is likely. Proving it will depend on the availability of an Edgeworth expansion for $\eta(\mathbf{X}, F)$. If τ depends on F (i.e., $\tau = \tau(F)$), then the bootstrap should be just first-order accurate.

6.5 Failure of the bootstrap

In spite of the many consistency theorems in the previous sections, there are instances where the ordinary bootstrap based on sampling with replacement from F_n actually

[13] We note that $T := \sqrt{n}(\bar{X}_n - \mu)/\sigma$ admits the following Edgeworth expansion:

$$\mathbb{P}(T \le x) = \Phi(x) + \frac{p_1(x|F)}{\sqrt{n}}\phi(x) + \frac{p_2(x|F)}{n}\phi(x) + \text{ smaller order terms},$$

where $p_1(x|F)$ and $p_2(x|F)$ are polynomials in x with coefficients depending on F.

does not work. Typically, these are instances where the root $\eta(\mathbf{X}, F)$ fails to admit a CLT. Before seeing a few examples, we list a few situations where the ordinary bootstrap fails to estimate the c.d.f. of $\eta(\mathbf{X}, F)$ consistently:

(a) $\eta(\mathbf{X}, F) = \sqrt{n}(\bar{X}_n - \mu)$ when $\mathrm{Var}_F(X_1) = \infty$.

(b) $\eta(\mathbf{X}, F) = \sqrt{n}(g(\bar{X}_n) - g(\mu))$ and $\nabla g(\mu) = 0$.

(c) $\eta(\mathbf{X}, F) = \sqrt{n}(g(\bar{X}_n) - g(\mu))$ and g is not differentiable at μ.

(d) The underlying population F_θ is indexed by a parameter θ, and the support of F_θ depends on the value of θ.

(e) The underlying population F_θ is indexed by a parameter θ, and the true value θ_0 belongs to the *boundary* of the parameter space Θ.

Exercise (HW4): Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be an i.i.d. sample from F and $\sigma^2 = \mathrm{Var}_F(X_1) = 1$. Let $g(x) = |x|$ and let $\eta(\mathbf{X}, F) = \sqrt{n}(g(\bar{X}_n) - g(\mu))$. If the true value of μ is 0, then by the CLT for $\bar{X}_n$ and the continuous mapping theorem, $\eta(\mathbf{X}, F) \xrightarrow{d} |Z|$ with $Z \sim N(0, \sigma^2)$. Show that the bootstrap does not work in this case.

6.6 Subsampling: a remedy to the bootstrap

The basic idea of *subsampling* is to approximate the sampling distribution of a statistic based on the values of the statistic computed over *smaller subsets* of the data. For example, in the case where the data are n observations that are i.i.d., a statistic is computed based on the entire data set and is recomputed over all $\binom{n}{b}$ data sets of size b. These recomputed values of the statistic are suitably normalized to approximate the true sampling distribution.

Suppose that $X_1, \ldots, X_n$ is a sample of n i.i.d. random variables having a common probability measure denoted by P. Suppose that the goal is to construct a confidence region for some parameter $\theta(P) \in \mathbb{R}$.

Let $\hat{\theta}_n \equiv \theta_n(X_1, \ldots, X_n)$ be an estimator of $\theta(P)$. It is desired to estimate or approximate the true sampling distribution of $\hat{\theta}_n$ in order to make inferences about $\theta(P)$.

Let $H_n(\cdot, P)$ be the sampling c.d.f. of $\tau_n(\hat{\theta}_n - \theta)$ based on a sample of size n from P, where τ_n is a normalizing constant. Essentially, the only assumption that we will need to construct asymptotically valid confidence intervals for $\theta(P)$ is the following: there

exists a limiting non-degenerate c.d.f. $H(\cdot, P)$ such that $H_n(\cdot, P)$ converges weakly to $H(\cdot, P)$ as $n \to \infty$.

To describe the method let $Y_1, \ldots, Y_{N_n}$ be equal to the $N_n := \binom{N}{b}$ subsets of size b of $\{X_1, \ldots, X_n\}$, ordered in any fashion. Of course, the Y_i's depend on b and n, but this notation has been suppressed. Only a very weak assumption on b will be required. In typical situations, it will be assumed that $b/n \to 0$ and $b \to \infty$ as $n \to \infty$.

Now, let $\hat{\theta}_{n,b,j}$ be equal to the statistic $\hat{\theta}_b$ evaluated at the data set Y_j. The approximation to $H_n(x, P)$ we study is defined by

$$L_{n,b}(x) = \frac{1}{N_n} \sum_{i=1}^{N_n} I\{\tau_b(\hat{\theta}_{n,b,j} - \hat{\theta}_n) \leq x\}.$$

The motivation behind the method is the following. For any j, Y_j is a random sample of size b from P. Hence, the exact distribution of $\tau_b(\hat{\theta}_{n,b,i} - \theta(P))$ is $H_b(\cdot, P)$. The empirical distribution of the N_n values of $\tau_b(\hat{\theta}_{n,b,j} - \hat{\theta}_n)$ should then serve as a good approximation to $H_b(P) \approx H_n(P)$. Of course, $\theta(P)$ is unknown, so we replace $\theta(P)$ by $\hat{\theta}_n$, which is asymptotically permissible because $\tau_b(\hat{\theta}_n - \theta(P))$ is of order $\tau_b/\tau_n \to 0$.

Theorem 6.9. Assume that there exists a limiting non-degenerate c.d.f. $H(\cdot, P)$ such that $H_n(\cdot, P)$ converges weakly to $H(\cdot, P)$ as $n \to \infty$. Also assume $\tau_b/\tau_n \to 0$, $b \to \infty$, and $b/n \to 0$ as $n \to \infty$.

(i) If x is a continuity point of $H(.\cdot, P)$, then $L_{n,b}(x) \xrightarrow{p} H(x, P)$.

(ii) If $H(\cdot, P)$ is continuous, then $\sup_x |L_{n,b}(x) - H_n(x, P)| \xrightarrow{p} 0$.

(iii) Assume $\tau_b(\hat{\theta}_n - \theta(P)) \to 0$ almost surely and, for every $d > 0$,

$$\sum_n \exp\{-d(n/b)\} < \infty.$$

Then, the convergences in (i) and (ii) hold with probability one.

Proof. See the proof of Theorem 2.2.1 in [8]. □

6.7 Bootstrapping regression models

Regression models are among the key ones that differ from the i.i.d. setup and are also among the most widely used. Bootstrap for regression cannot be model-free; the particular choice of the bootstrap scheme depends on whether the errors are i.i.d. or

not. We will only talk about the linear model with deterministic x's and i.i.d. errors. Additional moment conditions will be necessary depending on the specific problem to which the bootstrap will be applied; see e.g., [4]. First let us introduce some notation.

We consider the model

$$y_i = \beta^\top x_i + \epsilon_i,$$

where β is a $p \times 1$ ($p < n$) vector and so is x_i, and ϵ_i's are i.i.d. F with mean 0 and variance $\sigma^2 < \infty$.

Let X be the $n \times p$ design matrix with the i'th row equal to x_i and let $Y := (y_1, \ldots, y_n) \in \mathbb{R}^n$. The least squares estimator of β is defined as

$$\hat{\beta}_n := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - x_i^\top \beta)^2 = (X^\top X)^{-1} X^\top Y,$$

where we assume that $(X^\top X)^{-1}$ is nonsingular.

We may be interested in the sampling distribution of

$$(X^\top X)^{-1}(\hat{\beta}_n - \beta) \sim H_n(F).$$

First observe that H_n only depends on F. The *residual bootstrap* scheme is described below.

Compute the *residual* vector

$$\hat{\epsilon} = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^\top := Y - X\hat{\beta}_n.$$

We consider the centered residuals:

$$\tilde{\epsilon}_i = y_i - x_i^\top \hat{\beta}_n - \frac{1}{n} \sum_{j=1}^n \hat{\epsilon}_j, \qquad \text{for} \quad i = 1, \ldots, n.$$

The bootstrap estimator of the distribution $H_n(F)$ is $H_n(\tilde{F}_n)$, where $\tilde{F}_n$ is the empirical c.d.f. of $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n$.

We proved in class that an application of the Lindeberg-Feller CLT shows that the above bootstrap scheme is consistent, under the conditions:

(i) p is fixed (as n grows);

(ii) $\frac{1}{n} X_n^\top X_n \to \Sigma$, where Σ is positive definite;

(iii) $\frac{1}{\sqrt{n}} |x_{ij,n}| \to 0$ as $n \to \infty$, where $X_n = (x_{ij,n})$.

6.8 Bootstrapping a nonparametric function: the Grenander estimator

Consider $X_1, \ldots, X_n$ i.i.d. from a nonincreasing density f_0 on $[0, \infty)$. The goal is to estimate f_0 nonparametrically. In particular, we consider the nonparametric maximum likelihood estimator (NPMLE) of f_0, defined as

$$\tilde{f}_n := \arg\max_{f\downarrow} \prod_{i=1}^{n} f(X_i),$$

where the maximization if over all nonincreasing densities on $[0, \infty)$. It can see shown that

$$\tilde{f}_n = \mathrm{LCM}'[F_n],$$

where F_n is the empirical c.d.f. of the data, and $\mathrm{LCM}'[F_n]$ denotes the right-hand slope of the least concave majorant of F_n; see e.g.,
http://www.math.yorku.ca/∼hkj/Teaching/Bristol/notes.pdf
for the characterization, computation and theoretical properties of $\tilde{f}_n$.

In class we considered bootstrapping the Grenander estimator $\tilde{f}_n$, the NPMLE of f_0, at a fixed point $t_0 > 0$, in the interior of the support of f_0. We sketched a proof of the inconsistency of bootstrapping from F_n or $\mathrm{LCM}'[F_n]$; see [11] for the details. We also derived sufficient conditions for the consistency of any bootstrap scheme in this problem. Furthermore, we showed that we can consistently bootstrap from a smoothed version of $\tilde{f}_n$.

7 Multiple hypothesis testing

In the *multiple hypothesis testing*[14,15] problem we wish to test many hypotheses simultaneously. The null hypotheses are denoted by $H_{0,i}, i = 1, \ldots, n$, where n denotes the total number of hypotheses.

Consider a prototypical example: we test $n = 1000$ null hypotheses at level 0.05 (say). Suppose that everything is null (i.e., all the null hypotheses are true) — even then on an average we expect 50 rejections.

In general, the problem is how do we detect the true non-null effects (hypotheses where the null is not true) when a majority of the null hypotheses are true? This question has received a lot of attention in the statistical literature, particularly in genomic experiments. Consider the following example.

Example 7.1 (Prostate cancer study)**.** DNA microarrays measure expression levels of tens of thousands of genes. The data consist of levels of mRNA, which are thought to measure how much of a protein the gene produces. A larger number implies a more active gene.

Suppose that we have n genes and data on the expression levels for each gene among healthy individuals and those with prostate cancer. In the example considered in [3], $n = 6033$ genes were measured on 50 control patients and 52 patients with prostate cancer. The data obtained are (X_{ij}) where

$$X_{ij} = \text{ gene expression level on gene } i \text{ for the } j\text{'th individual.}$$

We want to test the effect of the i'th gene. For the i'th gene, we use the following test statistic:

$$\frac{\bar{X}_{i\cdot}^P - \bar{X}_{i\cdot}^C}{\mathrm{sd}(\ldots)} \sim t_{100}, \qquad \text{under } H_{0,i},$$

where $\bar{X}_{i\cdot}^P$ denotes the average expression level for the i'th gene for the 52 cancer patients and $\bar{X}_{i\cdot}^C$ denotes the corresponding value for the control patients and $\mathrm{sd}(\ldots)$ denotes the standard error of the difference. We reject the null $H_{0,i}$ for gene i if the test statistic exceeds the critical value $t_{100}^{-1}(1-\alpha)$, for $\alpha \in (0,1)$.

There are two main questions that we will address on this topic:

[14] Many thanks to Jimmy K Duong for scribing the lecture notes based on which this section is adapted.

[15] Most of the material here can be found in the lecture notes by Emmanuel Candes; see http://statweb.stanford.edu/~candes/stats300c/lectures.html.

- *Global testing.* In global testing, our primary interest is *not* on the n hypotheses $H_{0,i}$, but instead on the global hypothesis $H_0 : \cap_{i=1}^n H_{0,i}$, the intersection of the $H_{0,i}$'s.
- *Multiple testing.* In this scenario we are concerned with the individual hypotheses $H_{0,i}$ and want to say something about each hypothesis.

7.1 Global testing

Consider the following prototypical (Gaussian sequence model) example:

$$y_i = \mu_i + z_i, \quad \text{for } \ i = 1, \ldots, n, \tag{75}$$

where z_i's are i.i.d. $N(0,1)$, the μ_i's are unknown constants and we only observe the y_i's. We want to test

$$H_{0,i} : \mu_i = 0 \qquad \text{versus} \qquad H_{1,i} : \mu_i \neq 0 \text{ (or } \mu_i > 0).$$

In global testing, the goal is to test the hypothesis:

$$H_0 : \mu_i = 0, \text{ for all } i(\text{no signal}), \qquad \text{versus} \qquad H_1 : \text{ at least one } \mu_i \text{ is non-zero.}$$

The complication is that if we do each of these tests $H_{0,i}$ at level α, and then want to combine them, the global null hypothesis H_0 might not have level α. This is the first hurdle.

Data: $p_1, p_2, \ldots, p_n$: p-values for the n hypotheses.

We will assume that under $H_{0,i}$, $p_i \sim \text{Unif}(0,1)$. (we are not assuming independence among the p_i's yet.)

7.2 Bonferroni procedure

Suppose that $\alpha \in (0,1)$ is given. The *Bonferroni procedure* can be described as:

- Test $H_{0,i}$ at level α/n, for all $i = 1, \ldots, n$.
- Reject the global null hypothesis H_0 if we reject $H_{0,i}$ for some i.

This can be succinctly expressed as looking at the minimum of the p-values, i.e.,

$$\text{Reject } H_0 \qquad \text{if} \qquad \min_{i=1,\ldots,n} p_i \leq \frac{\alpha}{n}.$$

Question: Is this a valid level-α test, i.e., is $P_{H_0}(\text{Type I error}) \overset{?}{\leq} \alpha$? **Answer**: Yes. Observe that

$$\begin{aligned}
\mathbb{P}_{H_0}(\text{Rejecting } H_0) &= \mathbb{P}_{H_0}\left(\min_{i=1,\ldots,n} p_i \leq \alpha/n\right) \\
&= \mathbb{P}_{H_0}(\cup_{i=1}^n \{p_i \leq \alpha/n\}) \\
&\leq \sum_{i=1}^n P_{H_{0,i}}(p_i \leq \alpha/n), \quad \text{(crude upper bound)} \\
&= n \cdot \alpha/n, \quad \text{since } p_i \sim \text{Unif}([0,1]) \text{ under null} \\
&= \alpha.
\end{aligned}$$

So this is a valid level-α test, whatever the p_i's are (the p_i's could be dependent).

Question: Are we being too conservative (the above is an upper bound)? As we are testing each hypothesis using a very small level α/n most of the p-values would fail to be significant. The feeling is that we need a *very strong signal* for some i to detect the global null using the Bonferroni method.

Answer: We are not doing something very crude, if all the p-values are independent.

Question: What is the exact level of the test?

Answer: If the p_i's are independent, then observe that

$$\begin{aligned}
\mathbb{P}_{H_0}\left(\min_i p_i \leq \alpha/n\right) &= 1 - \mathbb{P}_{H_0}\left(\cap_{i=1}^n \{p_i > \alpha/n\}\right) \\
&= 1 - \prod_{i=1}^n P_{H_{0,i}}(p_i > \alpha/n) \quad \text{(using independence)} \\
&= 1 - \left(1 - \frac{\alpha}{n}\right)^n \\
&\xrightarrow{\text{as } n\to\infty} 1 - e^{-\alpha} \\
&\approx \alpha \qquad \text{(for } \alpha \text{ small)}.
\end{aligned}$$

Thus, the Bonferroni approach is not a bad thing to do, especially when we have independent p-values.

7.2.1 Power of the Bonferroni procedure

Let us now focus on the power of the Bonferroni method. To discuss power we need a model for the alternative.

Question: Consider the example of the Gaussian sequence model mentioned previously. Under what scenario for the μ_i's do we expect the Bonferroni test to do well?

Answer: If we have (a few) strong signals, then the Bonferroni procedure is good. We will try to formalize this now.

In the Gaussian sequence model the Bonferroni procedure reduces to: Reject $H_{0,i}$ ($H_{0,i} : \mu_i = 0$ vs. $H_{1,i} : \mu_i > 0$) if

$$y_i > z_{\alpha/n},$$

where $z_{\alpha/n}$ is the $(1 - \alpha/n)$'th quantile of the standard normal distribution.

Question: How does $z_{\alpha/n}$ behave? Do we know its order (when α is fixed and n is large)?

Answer: As first approximation, $z_{\alpha/n}$ is like $\sqrt{2 \log n}$ (an important number for Gaussian random variables).

Fact 1. Here is a fact from extreme value theory about the order of the maximum of the z_i's, i.e., $\max_{i=1,\ldots,n} z_i$:

$$\frac{\max_{i=1,\ldots,n} z_i}{\sqrt{2 \log n}} \xrightarrow{\text{a.s.}} 1,$$

i.e., if we have a bunch of n independent standard normals, the maximum is like $\sqrt{2 \log n}$ (Exercise: show this).

Fact 2. Bound on $1 - \Phi(t)$:

$$\frac{\phi(t)}{t}\left(1 - \frac{1}{t^2}\right) \leq 1 - \Phi(t) \leq \frac{\phi(t)}{t},$$

which implies that

$$1 - \Phi(t) \approx \frac{\phi(t)}{t} \qquad \text{for } t \text{ large.}$$

Here is a heuristics proof of the fact that $z_{\alpha/n} \approx \sqrt{2 \log n}$:

$$\begin{aligned}
1 - \Phi(t) \approx \frac{\phi(t)}{t} &= \frac{\alpha}{n} \\
\Leftrightarrow \frac{e^{-t^2/2}}{\sqrt{2\pi} t} &= \frac{\alpha}{n} \\
\Leftrightarrow -\frac{t^2}{2} &= \cancel{\log(\sqrt{2\pi} t)} + \log(\alpha/n) \quad (\text{as } \log(\sqrt{2\pi} t) \text{ is a smaller order term}) \\
\approx t^2 &= -2\log(\alpha/n) = 2\log n - 2\log\alpha \approx \sqrt{2\log n}.
\end{aligned}$$

The mean of $\max_{i=1,\ldots,n} z_i$ is like $\sqrt{2\log n}$ and the fluctuations around the mean is of order $O_p(1)$.

Exercise: Use the Gaussian concentration inequality to derive this result. Note that the maximum is a Lipschitz function.

To study the power of the Bonferroni procedure, we consider the following stylistic regimes (in the following the superscript (n) is to allow the variables to vary with n):

(i) $\mu_1^{(n)} = (1+\epsilon)\sqrt{2\log n}$ and $\mu_2 = \ldots = \mu_n = 0$,

(ii) $\mu_1^{(n)} = (1-\epsilon)\sqrt{2\log n}$ and $\mu_2 = \ldots = \mu_n = 0$,

where $\epsilon > 0$. So, in both settings, we have a one strong signal, and everything else is 0.

In case (i), the signal is slightly stronger than $\sqrt{2\log n}$; and in case (ii), the signal is slightly weaker than $\sqrt{2\log n}$. We will show that Bonferroni actually works for case (i) (by that we mean the power of the test actually goes to 1). Meanwhile, the Bonferroni procedure fails for case (ii) — the power of the test converges to α.

This is not only a problem with the Bonferroni procedure — it can be shown that no test can detect the signal in case (ii).

Case (i):

$$\begin{aligned}\mathbb{P}(\max y_i > z_{\alpha/n}) &= \mathbb{P}\left(\{y_1 > z_{\alpha/n}\} \cup \left\{\max_{i=2,\ldots,n} y_i > z_{\alpha/n}\right\}\right) \\ &\geq \mathbb{P}(\{y_1 > z_{\alpha/n}\}) \\ &\approx \mathbb{P}\left(z_1 > \sqrt{2\log n} - (1+\epsilon)\sqrt{2\log n}\right) \to 1.\end{aligned}$$

In this regime, just by looking at y_1, we will be able to detect that H_0 is not true.

Case (ii):

$$\mathbb{P}(\max y_i > z_{\alpha/n}) \leq \mathbb{P}(y_1 > z_{\alpha/n}) + \mathbb{P}\left(\max_{i=2,\ldots,n} y_i > z_{\alpha/n}\right).$$

Note that the first term is equal to $\mathbb{P}(z_1 > \epsilon\sqrt{2\log n}) \to 0$ as $n \to \infty$; whereas the second term converges to $1 - e^{-\alpha}$. Hence, we have shown that in this case the power of the test is less than or equal to the level of the test. So the test does as well as just plain guesswork.

This shows the dichotomy in the Bonferroni procedure; that by just changing the signal strength you can always recover or you can fail $(1-\alpha)$ of the time.

Whenever we have a hypothesis testing procedure, there has to be an effort in trying to understand the power of the procedure. And it is quite often the case that different tests (using different test statistics) are usually geared towards detecting different kinds of departures from the null. Here, the Bonferroni procedure is geared towards detecting sparse, strong signals.

7.3 Chi-squared test

Consider the Gaussian sequence model described in (75) and suppose that we want to test the global null hypothesis:

$$H_0 : \mu_i = 0, \text{ for all } i, \text{(no signal)} \qquad \text{versus} \qquad H_1 : \text{ at least one } \mu_i \text{ is non-zero.}$$

Letting $Y = (y_1, \ldots, y_n)$, the *chi-squared test* can be expressed as:

$$\text{Reject } H_0 \text{ if} \quad T := \|Y\|^2 > \chi^2_n(1-\alpha).$$

Note that under H_0,

$$T \sim \chi^2_n,$$

and under H_1,

$$T \sim \chi^2_n(\|\mu\|^2),$$

where $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ and $\chi^2_n(\|\mu\|^2)$ denotes the non-central χ^2_n distribution with non-centrality parameter $\|\mu\|^2$.

This test is going to have high power when $\|\mu\|^2$ is large. So, this test would have high power when there are many weak signals (even if each μ_i is slightly different from zero as we square it and add these up we can get a substantially large $\|\mu\|^2$). The Bonferroni procedure may not be able to detect a scenario like this — given α/n to each hypothesis if the signal strengths are weak all of the p-values (for the different hypotheses) might be considerably large.

7.4 Fisher's combination test

Suppose that $p_1, \ldots, p_n$ are the n p-values obtained from the n hypotheses tests. We assume that the p_i's are independent. The *Fisher's combination test* rejects the global null hypothesis if

$$T := \sum_{i=1}^{n} -2 \log p_i$$

is large. Observe that, under H_0,

$$T := -2\sum_{i=1}^{n} \log p_i \sim \chi^2_{2n}.$$

This follows from the fact that under $H_{0,i}$,

$$-\log p_i \sim \text{Exp}(1) \equiv \text{Gamma}(1,1).$$

Again, as this test is aggregating the p-values, it will hopefully be able to detect the presence of many weak signals.

7.5 Multiple testing/comparison problem: false discovery rate

Until now, we have been considering tests of the global null $H_0 = \cap_i H_{0,i}$. For some testing problems, however, our goal is to accept or reject each individual $H_{0,i}$. Given n hypotheses, we have four types of outcomes in multiple testing:

	Accept $H_{0,i}$	Reject $H_{0,i}$	
$H_{0,i}$ true	U	V	n_0
$H_{0,i}$ false	T	S	$n - n_0$
	$n - R$	R	n

where $R =$ number of rejections is an observed random variable; U, V, S, T are unobserved random variables. Note that

$$V = \text{ number of false discoveries.}$$

Suppose that the hypotheses indexed by $I_0 \subseteq \{1, \ldots, n\}$ are truly null with $|I_0| = n_0$ and the remaining hypotheses are non-null.

Ideally, we would not like to make false discoveries. But if you are not willing to make any false discoveries, which basically translates to our threshold/cutoff being really large for each test, then we will not be able make any discoveries at all.

Traditionally, statisticians want to control the *family-wise error rate* (FWER) :

$$\text{FWER } = \mathbb{P}(V \geq 1).$$

It is very easy to design a test whose FWER is controlled by a predetermined level α: reject or accept each hypothesis H_{0_i} according to a test whose type I error is at most α/n. Indeed, this is the Bonferroni method. By the union bound, one then has

$$\text{FWER} = \mathbb{P}\left(\cup_{i \in I_0} \{\text{Reject } H_{0,i}\}\right) \leq \sum_{i \in I_0} \mathbb{P}\left(\text{Reject } H_{0,i}\right) \leq \frac{\alpha n_0}{n} \leq \alpha.$$

In modern theory of hypothesis testing, control of the FWER is considered too stringent mainly because it leads to tests that fail to reject many non-null hypotheses as well.

The *false discovery rate* (FDR) is an error control criterion developed in the 1990's as an alternative to the FWER. When the number of tests is in the tens of thousands or even higher, FWER control is so stringent a criterion that individual departures from the null have little chance of being detected. In such cases, it may be unreasonable to control the probability of having any false rejections. Attempting to do so would leave us with virtually no power to reject individual non-nulls. Sometimes, control of FWER is even not quite needed.

A new point of view advanced by [1] proposes controlling the expected proportion of errors among the rejected hypotheses. The *false discovery proportion* (FDP) is defined as

$$\text{FDP} := \frac{V}{\max(R, 1)}.$$

FDP is an unobserved random variable, so the criterion we propose to control is its expectation, which we refer to as the *false discovery rate*:

$$\text{FDR} := \mathbb{E}(\text{FDP}).$$

The Benjamini-Hochberg (BH) procedure controls FDR at any desired level (e.g., suppose we take $q = 0.2$), i.e.,

$$\text{FDR} \ \leq q = 0.2;$$

thus out of all of the rejections we make we are willing to have 20% of them be false, on an average.

The BH procedure can be described as: suppose that $p_1, \ldots, p_n$ are the p-values from the n hypotheses tests. Let

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$$

be the sorted p-values. Let

$$i_0 := \max\left\{i \leq n : p_{(i)} \leq q\frac{i}{n}\right\}, \qquad 0 < q < 1.$$

We reject all the hypotheses $H_{0,(i)}$ for $1 \leq i \leq i_0$ (reject those hypotheses with p-values from $p_{(1)}$ to $p_{(i_0)}$). Pictorially this can be easily expressed as: draw the line

with slope q passing through the origin and plot the ordered p-values, and reject all the hypotheses whose p-values lie above the line after the last time it was below the line.

Another way to view the BH procedure is via the following sequential description: start with $\{i = n\}$ and keep accepting the hypothesis corresponding to $p_{(i)}$ as long as $p_{(i)} > qi/n$. As soon as $p_{(i)} \leq iq/n$, stop and reject all the hypotheses corresponding to $p_{(j)}$ for $j \leq i$.

Theorem 7.2. Suppose that the p-values $p_1, \ldots, p_n$ are independent. Then

$$\text{FDR } = \mathbb{E}\left(\frac{V}{\max(R,1)}\right) \leq q.$$

Remark 7.1. Note that the above result states that the BH procedure controls FDR for all configurations of $\{H_{0,i}\}_{i=1}^n$.

Proof. Without loss of generality suppose that $H_{0,1}, \ldots, H_{0,n_0}$ are true. Observe that

$$\{R = r\} = \left\{p_{(r)} \leq \frac{r}{n}q, p_{(s)} > \frac{s}{n}q, \forall s > r\right\}.$$

Further, under $\{R = r\}$, $V = \sum_{i=1}^{n_0} \mathbb{1}\{p_i \leq \frac{r}{n}q\}$. Thus,

$$\begin{aligned}
& \left\{p_1 \leq \frac{r}{n}q, R = r\right\} \\
= & \left\{p_1 \leq \frac{r}{n}q, p_{(r)} \leq \frac{r}{n}q, p_{(s)} > \frac{s}{n}q, \forall s > r\right\} \\
= & \left\{p_1 \leq \frac{r}{n}q, p_{r-1}^{(-1)} \leq \frac{r}{n}q, p_s^{(-1)} > \frac{s+1}{n}q, \forall s \geq r\right\} \\
= & \left\{p_1 \leq \frac{r}{n}q, \tilde{R}(p^{(-1)}) = r - 1\right\},
\end{aligned}$$

where $p^{(-1)} = (p_2, \ldots, p_n)$ and $\tilde{R} = \sup\{1 \leq i \leq n-1 : p_{(i)}^{(-1)} \leq \frac{i+1}{n}q\}$. Finally we can

show that

$$
\begin{aligned}
\text{FDR} &= \mathbb{E}\left(\frac{V}{R}\mathbb{1}\{R \neq 0\}\right) \\
&= \mathbb{E}(\sum_{r=1}^{n} \frac{V}{r}\mathbb{1}\{R = r\}) \\
&= \sum_{r=1}^{n} \frac{1}{r}\mathbb{E}(V\mathbb{1}\{R = r\}) \\
&= \sum_{r=1}^{n} \frac{1}{r}\sum_{i=1}^{n_0} P\left(p_i \leq \frac{r}{n}q, R = r\right) \\
&= \sum_{r=1}^{n} \frac{1}{r}n_0\mathbb{P}\left(p_1 \leq \frac{r}{n}q, R = r\right) \qquad \text{(by exchangeability)} \\
&= \sum_{r=1}^{n} \frac{n_0}{r}\mathbb{P}\left(p_1 \leq \frac{r}{n}q)\mathbb{P}(\tilde{R}(p^{(-1)}) = r - 1\right) \qquad \text{(by independence)} \\
&= \sum_{r=1}^{n} \frac{n_0}{r}\frac{r}{n}q\mathbb{P}(\tilde{R}(p^{(-1)}) = r - 1) \\
&= \frac{n_0}{n}q \leq q.
\end{aligned}
$$

□

7.6 The Bayesian approach: connection to empirical Bayes

By formulating the multiple testing problem in a simple Bayesian framework, we are able to construct procedures that control a quantity closely related to the FDR as we have previously defined.

We assume that we have n hypotheses, which are null ($H = 0$) with probability π_0 and non-null ($H = 1$) with probability $1 - \pi_0$. Our observations $\{X_i\}_{i=1}^n$ (p-values/z-values) are thus assumed to come from the mixture distribution

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0)f_1(x)$$

where f_0 is the density of X_i if null is true (with c.d.f. F_0; e.g., $U[0, 1]$ or $N(0, 1)$) and f_1 is the density of X_i otherwise (with c.d.f. F_1). Let H denote the unobserved variable that takes the value 0 or 1 depending on whether the null hypothesis is true or not.

In this setup, we observe $X \in A$ and wonder whether it is null or not. By Bayes'

rule, we can evaluate this probability to be

$$\begin{aligned}\phi(A) &:= \mathbb{P}(H=0|X\in A) \qquad \text{(posterior probability of the null hypothesis)}\\ &= \frac{\mathbb{P}(X\in A|H=0)}{\mathbb{P}(X\in A)}\\ &= \frac{\pi_0 P_0(A)}{P(A)} = \frac{\pi_0 \int_A f_0(x)dx}{P(A)},\end{aligned}$$

where $P_0(A)$ denotes the probability of a set A under the null distribution.

We can call the quantity $\phi(A)$ the *Bayes false discovery rate* (BFDR). If we report $x \in A$ as non-null, $\phi(A)$ is the probability that we have made a false discovery. What should be A? If we reject $H_{0,i}$ if $X_i > x_c$ (e.g., if we are testing $H_{0,i} : \mu_i = 0$ vs. $H_{1,i} : \mu_i > 0$) then $A = [x_c, \infty)$. In practice, we will have some critical value x_c and A will take one of the following forms:

$$[x_c, \infty) \qquad (-\infty, x_c] \qquad (-\infty, -x_c] \cup [x_c, \infty). \tag{76}$$

In order to make use of the above machinery, we need to have knowledge of π_0, f_0 and f_1. It is extremely unlikely that we would know these quantities in practice. By using empirical Bayes techniques, we are able to accurately estimate these quantities based on our data, as explained below.

We proceed by assuming the following: (i) usually f_0 is known (assumed $N(0,1)$ or Unif$(0,1)$); (ii) π_0 is 'almost known', in the sense that it's a fraction close to 1 in many applications; (iii) f_1 is unknown.

Without knowing $P(A)$, the BFDR cannot be computed. However, we can estimate this quantity by

$$\widehat{P(A)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_A(X_i).$$

This yields the BFDR estimate:

$$\widehat{\text{BFDR}} = \widehat{\phi(A)} = \frac{\widehat{\pi}_0 P_0(A)}{\widehat{P(A)}}.$$

If n is large, then $\widehat{P(A)}$ will be close to $P(A)$, and thus $\widehat{\text{BFDR}}$ may be a good estimate of BFDR.

7.6.1 Global versus local FDR

Classical BH theory only lets us discuss false discovery rates for tail sets of the form (76). An advantage of the Bayesian theory is that we can now compute and

bound the FDR for generic measurable sets A. [3] likes to distinguish between the "local" and "global" FDR rates:

$$\text{Global FDR} : \text{FDR}(x_c) = \phi([x_c, \infty)), \qquad \text{Local FDR} : \text{FDR}(x_c) = \phi(\{x_c\}),$$

where $\text{FDR}(x_c)$ will in general be well-defined provided all distributions have continuous densities, i.e.,

$$\phi(\{x_0\}) = \frac{\pi_0 f_0(x_0)}{f(x_0)}.$$

These two quantities can be very different.

Example 7.3. Suppose that $F_0 = N(0,1)$ and $F_1 = \text{Unif}(-10, 10)$, $\pi_0 = 1/2$. In other words, under the null hypotheses the test statistics are standard Gaussian, whereas under the alternatives they have a uniform distribution over a medium-size interval, and on average half the hypotheses are null. In this case:

$$\phi([2,\infty)) = \frac{1 - \Phi(2)}{8/20 + (1 - \Phi(2))} \approx 0.054, \qquad \phi(\{2\}) = \frac{\phi(2)}{1/20 + \phi(2)} \approx 0.52.$$

Thus, a global FDR analysis suggests that $x \geq 2$ is strong evidence for the alternative, whereas a local FDR analysis tells us that in fact $x = 2$ is mild evidence for the null. (There is no contradiction here — under the data generating distribution, given that $x \geq 2$ you would expect that $x >> 2$, and so the expected global FDR is small.)

The beauty of local FDR theory is that it can tell us the probability that any given hypothesis is null, instead of just giving us the expected proportion of nulls among all rejections. It's down side, of course, is that it relies on more complex Bayesian machinery. Standard BH theory (which is what people mostly use in practice) gives us weaker global FDR type results, but requires much less assumptions to go through. For more on this topic see [3, Chapter 5].

7.6.2 Empirical Bayes interpretation of BH(q)

How does the BH procedure relate to the empirical Bayes procedures we are discussing? First, we note that z-values map to p-values using the relation

$$p_i = F_0(X_i), \qquad (X_i \text{ is the test statistic}).$$

Using this we observe that

$$p_{(i)} = F_0(X_{(i)}), \qquad \text{and} \qquad \frac{i}{n} = \widehat{F}_n(X_{(i)}) \approx F(X_{(i)}).$$

Thus,

$$i : p_{(i)} \leq \frac{i}{n} q \quad \Leftrightarrow \quad \frac{F_0(X_{(i)})}{\widehat{F}_n(X_{(i)})} \leq q \quad \approx \quad \frac{F_0(X_{(i)})}{F(X_{(i)})} \leq q.$$

Thus, assuming that $\widehat{\text{BFDR}}$ was computed with $\pi_0 = 1$, we observe that

$$\frac{F_0(X_{(i)})}{\widehat{F}_n(X_{(i)})} \leq q \quad \approx \quad \widehat{\phi}((-\infty, X_{(i)}]) \leq q.$$

The claim below then follows.

Claim: The empirical Bayes formulation of BH(q) is to reject $H_{0,(i)}$ for all $i \leq i_0$ where i_0 is the largest index such that

$$\widehat{\text{BFDR}}((-\infty, x_{(i_0)}]) \leq q.$$

Assuming independence of the test statistics, the FDR is at most q.

Note that π_0 is usually unknown. However, usually we set $\pi_0 = 1$, which results in a conservative estimate of the FDR.

8 High dimensional linear regression

Consider the standard linear regression model

$$\mathbf{y} = \mathbf{X}\beta^* + \mathbf{w},$$

where $\mathbf{X} \in \mathbb{R}^{N \times p}$ is the design matrix, $\mathbf{w} \in \mathbb{R}^N$ is the vector of noise variables (i.e., $\mathbb{E}(\mathbf{w}) = \mathbf{0}$), and $\beta^* \in \mathbb{R}^p$ is the unknown coefficient vector. We are interested in estimating β^* from the observed response $\mathbf{y}$. In this section we consider the situation where $p \gg N$ (or p is comparable to N) and study the performance of the *lasso* estimator[16] (least absolute shrinkage and selection operator; see e.g., [13]):

$$\hat{\beta} := \operatorname*{argmin}_{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq R} \|\mathbf{y} - \mathbf{X}\beta\|_2^2, \tag{77}$$

where $R > 0$ is a tuning parameter. The above is sometimes called as the *constrained* form of the lasso solution. An equivalent form (due to Lagrangian duality) is the penalized version

$$\min_{\beta \in \mathbb{R}^p} \left[\frac{1}{2N} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_N \|\beta\|_1 \right], \tag{78}$$

where $\lambda_N > 0$ is the Lagrange multiplier associated with the constraint $\|\beta\|_1 \leq R$.

The lasso estimator performs both variable selection and regularization simultaneously; it has good prediction accuracy and offers interpretability to the statistical model it produces. Figure 5 shows a simple illustration of the performance of the constrained lasso estimator (and ridge regression[17]) and gives some intuition as to why it can also perform variable selection.

Given a lasso estimate $\hat{\beta} \in \mathbb{R}^p$, we can assess its quality in various ways. In some settings, we are interested in the predictive performance of $\hat{\beta}$, so that we might compute a prediction loss function of the form

$$\mathcal{L}(\hat{\beta}, \beta^*) := \frac{1}{N} \|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\|_2^2,$$

corresponding to the mean-squared error of $\hat{\beta}$ over the given samples of $\mathbf{X}$. If the unknown vector β^* is of primary interest then a more appropriate loss function to consider would be the ℓ_2-error

$$\mathcal{L}_2(\hat{\beta}, \beta^*) := \|\hat{\beta} - \beta^*\|_2^2.$$

[16] This material is mostly taken from [5].

[17] In ridge regression we consider the problem: $\min_{\beta \in \mathbb{R}^p : \|\beta\|_2^2 \leq R^2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2$.

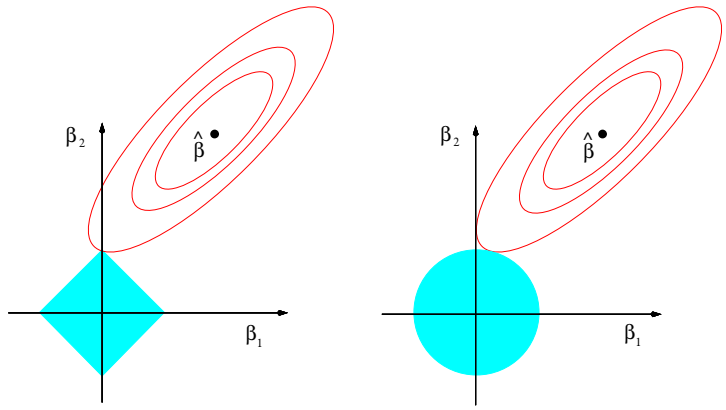


Figure 5: Estimation picture for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the residual-sum-of-squares function. The point $\hat{\beta}$ depicts the usual (unconstrained) least-squares estimate.

8.1 Strong convexity

The lasso minimizes the least-squares loss $f_N(\beta) := \frac{1}{2N}\|\mathbf{y} - \mathbf{X}\beta\|_2^2$ subject to an ℓ_1-constraint. Let us suppose that the difference in function values $\Delta f_N = |f_N(\hat{\beta}) - f_N(\beta^*)|$ converges to zero as the sample size N increases. The key question is the following: what additional conditions are needed to ensure that the ℓ_2-norm of the parameter vector difference $\Delta\beta = \|\hat{\beta} - \beta^*\|_2$ also converges to zero? Figure 6 illustrates two scenarios that suggest that the function f_N has to be suitably "curved".

A natural way to specify that a function is suitably "curved" is via the notion of strong convexity. More specifically, given a differentiable function $f : \mathbb{R}^p \to \mathbb{R}$, we say that it is *strongly convex* with parameter $\gamma > 0$ at $\theta^* \in \mathbb{R}^p$ if the inequality

$$f(\theta) - f(\theta^*) \geq \nabla f(\theta^*)^\top(\theta - \theta^*) + \frac{\gamma}{2}\|\theta - \theta^*\|_2^2$$

holds for all $\theta \in \mathbb{R}^p$. Note that this notion is a strengthening of ordinary convexity, which corresponds to the case $\gamma = 0$. When the function f is twice continuously differentiable, an alternative characterization of strong convexity is in terms of the Hessian $\nabla^2 f$: in particular, the function f is strongly convex with parameter γ around $\theta^* \in \mathbb{R}^p$ if and only if the minimum eigenvalue of the Hessian matrix $\nabla^2 f(\theta)$ is at least γ for all vectors θ in a neighborhood of θ^*.

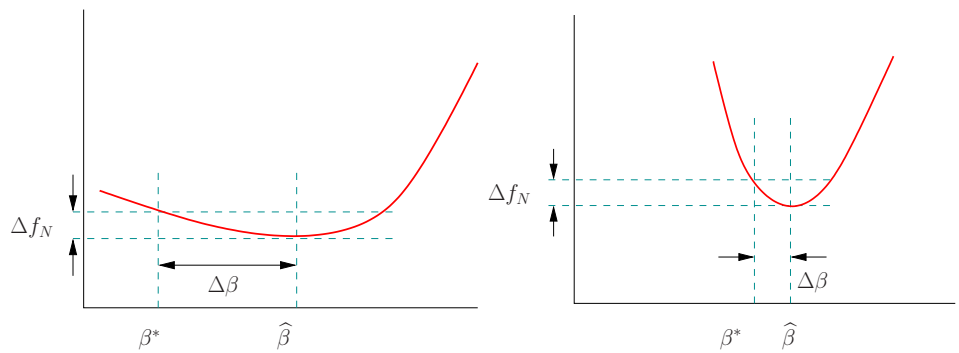


Figure 6: Relation between differences in objective function values and differences in parameter values. Left: the function f_N is relatively "flat" around its optimum $\hat{\beta}$, so that a small function difference $\Delta f_N = |f_N(\hat{\beta}) - f_N(\beta^*)|$ does not imply that $\Delta\beta = \|\hat{\beta} - \beta^*\|_2$ is small. Right: the function f_N is strongly curved around its optimum, so that a small difference Δf_N in function values translates into a small difference in parameter values.

8.2 Restricted strong convexity and ℓ_2-error $\|\hat{\beta} - \beta^*\|_2$

Let us now return to the high-dimensional setting, in which the number of parameters p might be larger than N. It is clear that the least-squares objective function $f_N(\beta)$ is always convex; under what additional conditions is it also strongly convex? A straightforward calculation yields that $\nabla^2 f(\beta) = \mathbf{X}^\top\mathbf{X}/N$ for all $\beta \in \mathbb{R}^p$. Thus, the least-squares loss is strongly convex if and only if the eigenvalues of the $p \times p$ positive semidefinite matrix $\mathbf{X}^\top\mathbf{X}$ are uniformly bounded away from zero. However, it is easy to see that any matrix of the form $\mathbf{X}^\top\mathbf{X}$ has rank at most $\min\{N, p\}$, so it is always rank-deficient — and hence not strongly convex — whenever $N < p$. Figure 7 illustrates the situation.

For this reason, we need to relax our notion of strong convexity. It turns out, as will be clarified by the analysis below, that it is only necessary to impose a type of strong convexity condition for some subset $\mathcal{C} \subset \mathbb{R}^p$ of possible perturbation vectors $\nu \in \mathbb{R}^p$.

Definition 8.1 (Restricted strong convexity)**.** We say that a function $f : \mathbb{R}^p \to \mathbb{R}$ satisfies *restricted strong convexity* at $\theta^* \in \mathbb{R}^p$ with respect to $\mathcal{C} \subset \mathbb{R}^p$ if there is a constant $\gamma > 0$ such that

$$\frac{\nu^\top \nabla^2 f(\theta)\nu}{\|\nu\|_2^2} \geq \gamma \qquad \text{for all nonzero } \nu \in \mathcal{C},$$

and for all $\theta \in \mathbb{R}^p$ in a neighborhood of θ^*.

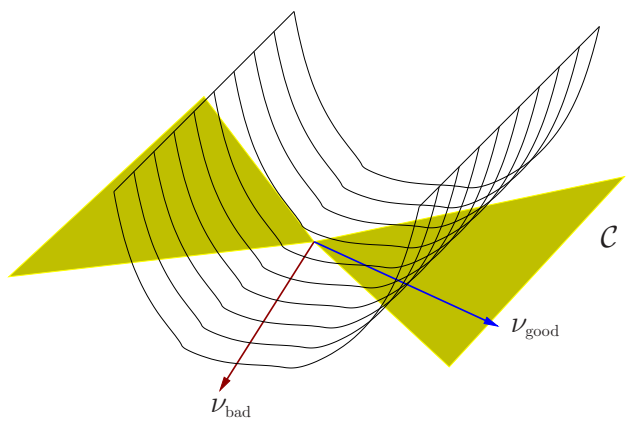


Figure 7: A convex loss function in high-dimensional settings (with $p \gg N$) cannot be strongly convex; rather, it will be curved in some directions but flat in others. As will be shown in later, the lasso error $\hat{\nu} = \hat{\beta} - \beta^*$ must lie in a restricted subset $\mathcal{C}$ of $\mathbb{R}^p$. For this reason, it is only necessary that the loss function be curved in certain directions of space.

In the specific case of linear regression, this notion is equivalent to lower bounding the *restricted eigenvalues* of the design matrix — in particular, requiring that

$$\frac{\frac{1}{N}\nu^\top \nabla^2 \mathbf{X}^\top \mathbf{X} \nu}{\|\nu\|_2^2} \geq \gamma \qquad \text{for all nonzero } \nu \in \mathcal{C}. \tag{79}$$

This is referred to as the γ-*RE condition*.

So, what constraint sets $\mathcal{C}$ are relevant? Suppose that the parameter vector β^* is sparse — say supported on the subset $S = S(\beta^*)$. Defining the lasso error $\hat{\nu} = \hat{\beta} - \beta^*$, let $\hat{\nu}_S \in \mathbb{R}^{|S|}$ denote the subvector indexed by elements of S, with $\hat{\nu}_{S^c}$ defined in an analogous manner. For appropriate choices of the ℓ_1-ball radius — or equivalently, of the regularization parameter λ_N — it turns out that the lasso error satisfies a *cone constraint* of the form

$$\|\hat{\nu}_{S^c}\|_1 \leq \alpha \|\hat{\nu}_S\|_1,$$

for some constant $\alpha \geq 1$. Thus, we consider a restricted set of the form

$$\mathcal{C}(S, \alpha) := \{\nu \in \mathbb{R}^p : \|\nu_{S^c}\|_1 \leq \alpha \|\nu_S\|_1\},$$

for some parameter $\alpha \geq 1$.

Theorem 8.2. Suppose that the design matrix $\mathbf{X}$ satisfies the restricted eigenvalue bound (79) with parameter $\gamma > 0$ over $\mathcal{C}(S, 1)$. Then any estimate $\hat{\beta}$ based on the constrained lasso (77) with $R = \|\beta^*\|_1$ satisfies the bound

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{4}{\gamma} \sqrt{\frac{k}{N}} \frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{\sqrt{N}}.$$

Before proving this result, let us discuss the different factors in the above bound. First, it is important to note that this result is deterministic, and apply to any set of linear regression equations with a given observed noise vector $\mathbf{w}$. Based on our earlier discussion of the role of strong convexity, it is natural that lasso ℓ_2-error is inversely proportional to the restricted eigenvalue constant $\gamma > 0$. The second term k/N is also to be expected, since we are trying to estimate an unknown regression vector with k unknown entries based on N samples. As we have discussed, the final term in both bounds, involving either $\|\mathbf{X}^\top \mathbf{w}\|_\infty$, reflects the interaction of the observation noise $\mathbf{w}$ with the design matrix $\mathbf{X}$.

Example 8.3 (Classical linear Gaussian model)**.** We begin with the classical linear Gaussian model for which the noise $\mathbf{w} \in \mathbb{R}^N$ is Gaussian with i.i.d. $N(0, \sigma^2)$ entries. Let us view the design matrix $\mathbf{X}$ as fixed, with columns $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$. For any given column $j \in \{1, \ldots, p\}$, a simple calculation shows that the random variable $x_j^\top \mathbf{w}/N$ is distributed as $N(0, \frac{\sigma^2}{N}\frac{\|\mathbf{x}_j\|_2^2}{N})$. Consequently, if the columns of the design matrix $\mathbf{X}$ are normalized (meaning $\|\mathbf{x}_j\|_2^2/N = 1$ for all $j = 1, \ldots, p$), then this variable has $N(0, \frac{\sigma^2}{N})$ distribution, so that we have the Gaussian tail bound

$$\mathbb{P}\left(\frac{|\mathbf{x}_j^\top \mathbf{w}|}{N} \geq t\right) \leq 2e^{-\frac{Nt^2}{2\sigma^2}} \qquad \text{for } \ t > 0.$$

Since $\|\mathbf{X}^\top \mathbf{w}\|_\infty/N$ corresponds to the maximum over p such variables, the union bound yields

$$\mathbb{P}\left(\frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N} \geq t\right) \leq 2e^{-\frac{Nt^2}{2\sigma^2}+\log p} = 2e^{-\frac{1}{2}(\tau-2)\log p},$$

where the second equality follows by setting $t = \sigma\sqrt{\frac{\tau \log p}{N}}$ for some $\tau > 2$. Consequently, we conclude that the lasso error satisfies the bound

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{4\sigma}{\gamma}\sqrt{\frac{\tau k \log p}{N}}$$

with probability at least $1 - 2e^{-\frac{1}{2}(\tau-2)\log p}$.

Proof of Theorem 8.2. In this case, since β^* is feasible and $\hat{\beta}$ is optimal, we have the inequality $\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 \leq \|\mathbf{y} - \mathbf{X}\beta^*\|_2^2$. Defining the error vector $\hat{\nu} := \hat{\beta} - \beta^*$, substituting in the relation $\mathbf{y} = \mathbf{X}\beta^* + \mathbf{w}$, and performing some algebra yields the *basic inequality*

$$\frac{\|\mathbf{X}\hat{\nu}\|_2^2}{2N} \leq \frac{\mathbf{w}^\top \mathbf{X}\hat{\nu}}{N}. \tag{80}$$

Applying a version of Hölder's inequality to the right-hand side yields the upper bound $\frac{1}{N}|\mathbf{w}^\top \mathbf{X}\hat{\nu}| \leq \frac{1}{N}\|\mathbf{X}^\top \mathbf{w}\|_\infty \|\hat{\nu}\|_1$.

Next, we claim that the inequality $\|\hat{\beta}\|_1 \leq R = \|\beta^*\|_1$ implies that $\hat{\nu} \in \mathcal{C}(S, 1)$. Observe that

$$\begin{aligned} R = \|\beta^*_S\|_1 &\geq \|\beta^* + \hat{\nu}\|_1 \\ &= \|\beta^*_S + \hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1 \\ &\geq \|\beta^*_S\| - \|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1. \end{aligned}$$

Rearranging this inequality, we see that $\|\hat{\nu}_{S^c}\|_1 \leq \|\hat{\nu}_S\|_1$, which shows that $\hat{\nu} \in \mathcal{C}(S, 1)$.

Thus, we have

$$\|\hat{\nu}\|_1 = \|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1 \leq 2\|\hat{\nu}_S\|_1 \leq 2\sqrt{k}\|\hat{\nu}\|_2,$$

where we have used the Cauchy-Schwarz inequality in the last step.

On the other hand, applying the restricted eigenvalue condition to the left-hand side of the inequality (80) yields

$$\gamma \frac{\|\hat{\nu}\|_2^2}{2} \leq \frac{\|\mathbf{X}\hat{\nu}\|_2^2}{2N} \leq \frac{\mathbf{w}^\top \mathbf{X}\hat{\nu}}{N} \leq \frac{1}{N}\|\mathbf{X}^\top \mathbf{w}\|_\infty \|\hat{\nu}\|_1 \leq \frac{1}{N}\|\mathbf{X}^\top \mathbf{w}\|_\infty 2\sqrt{k}\|\hat{\nu}\|_2.$$

Putting together the pieces yields the claimed bound. □

Exercise (HW 4): Suppose that the design matrix $\mathbf{X}$ satisfies the restricted eigenvalue bound (79) with parameter $\gamma > 0$ over $\mathcal{C}(S, 3)$. Given a regularization parameter $\lambda_N \geq 2\|\mathbf{X}^\top \mathbf{w}\|_\infty / N > 0$, show that any estimate $\hat{\beta}$ from the regularized lasso (78) satisfies the bound

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{3}{\gamma}\sqrt{\frac{k}{N}}\sqrt{N}\lambda_N.$$

8.3 Bounds on prediction error

In this section we focus on the Lagrangian lasso (78) and develop some theoretical guarantees for the prediction error $\mathcal{L}(\hat{\beta}, \beta) := \frac{1}{N}\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\|_2^2$.

Theorem 8.4. Consider the Lagrangian lasso with a regularization parameter $\lambda_N \geq \frac{2}{N}\|\mathbf{X}^\top \mathbf{w}\|_\infty$.

(a) Any optimal solution $\hat{\beta}$ satisfies

$$\frac{1}{N}\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\|_2^2 \leq 6\|\beta^*\|_1 \lambda_N.$$

(b) If β^* is supported on a subset S, and the design matrix $\mathbf{X}$ satisfies the γ-RE condition (79) over $\mathcal{C}(S, 3)$, then any optimal solution $\hat{\beta}$ satisfies

$$\frac{1}{N}\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\|_2^2 \leq \frac{9}{\gamma}|S|\lambda_N^2.$$

As we have discussed, for various statistical models, the choice $\lambda_N = c\sigma\sqrt{\frac{\log p}{N}}$ is valid for Theorem 8.4 with high probability, so the two bounds take the form

$$\begin{aligned}\frac{1}{N}\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\|_2^2 &\leq c_1\sigma R_1\sqrt{\frac{\log p}{N}}, \quad \text{and} \\ \frac{1}{N}\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\|_2^2 &\leq c_2\frac{\sigma}{\gamma}\frac{|S|\log p}{N},\end{aligned}$$

for suitable constants c_1, c_2. The first bound, which depends on the ℓ_1-ball radius R_1, is known as the "slow rate" for the lasso, since the squared prediction error decays as $1/\sqrt{N}$. On the other hand, the second bound is known as the "fast rate" since it decays as $1/N$. Note that the latter is based on much stronger assumptions: namely, the hard sparsity condition that β^* is supported on a small subset S, and more disconcertingly, the γ-RE condition on the design matrix $\mathbf{X}$. In principle, prediction performance should not require an RE condition, so that one might suspect that this requirement is an artifact of our proof technique. However, this dependence turns out to be unavoidable for any polynomial-time method; see e.g., [18] where, under a standard assumption in complexity theory, the authors prove that no polynomial-time algorithm can achieve the fast rate without imposing an RE condition.

Proof of Theorem 8.4. Define the function

$$G(\nu) := \frac{1}{2N}\|\mathbf{y} - \mathbf{X}(\beta^* + \nu)\|_2^2 + \lambda_N\|\beta^* + \nu\|_1.$$

Noting that $\hat{\nu} := \hat{\beta} - \beta^*$ minimizes G by construction, we have $G(\hat{\nu}) \leq G(0)$. Some algebra yields the *modified basic inequality*:

$$\frac{\|\mathbf{X}\hat{\nu}\|_2^2}{2N} \leq \frac{\mathbf{w}^\top\mathbf{X}\hat{\nu}}{N} + \lambda_N\{\|\beta^*\|_1 - \|\beta^* + \hat{\nu}\|_1\}. \tag{81}$$

Thus,

$$\begin{aligned}0 &\leq \frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N}\|\hat{\nu}\|_1 + \lambda_N\{\|\beta^*\|_1 - \|\beta^* + \hat{\nu}\|_1\} \\ &\leq \left\{\frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N} - \lambda_N\right\}\|\hat{\nu}\|_1 + 2\lambda_N\|\beta^*\|_1 \\ &\leq \frac{1}{2}\lambda_N\left\{-\|\hat{\nu}\|_1 + 4\|\beta^*\|_1\right\},\end{aligned}$$

where the last step uses the fact that $\frac{1}{N}\|\mathbf{X}^\top\mathbf{w}\|_\infty \leq \lambda_N/2$ (by assumption). Therefore, $\|\hat{\nu}\|_1 \leq 4\|\beta^*\|_1$. Returning again to the modified basic inequality (81), we have

$$\frac{\|\mathbf{X}\hat{\nu}\|_2^2}{2N} \leq \frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N}\|\hat{\nu}\|_1 + \lambda_N\|\beta^*\|_1 \leq \frac{\lambda_N}{2}\cdot 4\|\beta^*\|_1 + \lambda_N\|\beta^*\|_1 \leq 3\lambda_N\|\beta^*\|_1,$$

which establishes (a).

To prove (b), observe that as $\beta^*_{S^c} = 0$, we have $\|\beta^*\|_1 = \|\beta^*_S\|_1$, and

$$\|\beta^* + \hat{\nu}\|_1 = \|\beta^*_S + \hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1 \geq \|\beta^*_S\|_1 - \|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1.$$

Substituting this relation into the modified basic inequality (81) yields

$$\begin{aligned} \frac{\|\mathbf{X}\hat{\nu}\|_2^2}{2N} &\leq \frac{\mathbf{w}^\top \mathbf{X}\hat{\nu}}{N} + \lambda_N\{\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S^c}\|_1\}. \\ &\leq \frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N}\|\hat{\nu}\|_1 + \lambda_N\{\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S^c}\|_1\}. \end{aligned} \tag{82}$$

Given the stated choice of λ_N, the above inequality yields

$$\begin{aligned} \frac{\|\mathbf{X}\hat{\nu}\|_2^2}{2N} &\leq \frac{\lambda_N}{2}\{\|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1\} + \lambda_N\{\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S^c}\|_1\} \\ &\leq \frac{3}{2}\lambda_N\|\hat{\nu}_S\|_1 \quad \leq \quad \frac{3}{2}\lambda_N\sqrt{k}\|\hat{\nu}\|_2, \end{aligned} \tag{83}$$

where $k := |S|$.

Next we claim that the error vector $\hat{\nu}$ associated with any lasso solution $\hat{\beta}$ belongs to the cone $\mathcal{C}(S, 3)$. Since $\frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N} \leq \frac{\lambda_N}{2}$, inequality (82) implies that

$$0 \leq \frac{\lambda_N}{2}\|\hat{\nu}\|_1 + \lambda_N\{\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S^c}\|_1\}.$$

Rearranging and then dividing out by $\lambda_N > 0$ yields that $\|\hat{\nu}_{S^c}\|_1 \leq 3\|\hat{\nu}_S\|_1$ as claimed.

As the error vector $\hat{\nu}$ belongs to the cone $\mathcal{C}(S, 3)$, the γ-RE condition guarantees that $\|\hat{\nu}\|_2^2 \leq \frac{1}{N\gamma}\|\mathbf{X}\hat{\nu}\|_2^2$. Therefore, using (82) gives

$$\frac{1}{N}\|\mathbf{X}\hat{\nu}\|_2^2 \leq 3\lambda_N\sqrt{k}\|\hat{\nu}\|_2 \leq 3\lambda_N\sqrt{\frac{k}{N\gamma}}\|\mathbf{X}\hat{\nu}\|_2 \quad \Rightarrow \quad \frac{1}{\sqrt{N}}\|\mathbf{X}\hat{\nu}\|_2 \leq 3\lambda_N\sqrt{\frac{k}{\gamma}}.$$

This completes the proof. □

Exercise (HW 4): State and prove the analogous theorem for the constrained form of the lasso (given in (77)) where you take $R = \|\beta^*\|_1$.

8.4 Equivalence between ℓ_0 and ℓ_1-recovery

As seen in Theorem 8.4(b), the ℓ_1-constraint yields a bound on the prediction error that is almost optimal — if we knew the set S then using linear regression would yield

a bound of the order $\sigma|S|/N$; using lasso, we just pay an additional multiplicative factor of $\log p$. As S is obviously unknown, we can think of fitting all possible linear regression models with $k := |S|$ predictors and then choosing the best one. This would be equivalent to solving the following ℓ_0-problem:

$$\min_{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq k} \|\mathbf{y} - \mathbf{X}\beta\|_2^2,$$

where $\|\beta\|_0$ denotes the number of non-zero components of β. Obviously, this procedure is computationally infeasible and possibly NP hard.

In this subsection we compare the ℓ_0 and ℓ_1-problems in the *noiseless* setup. This would shed light on when we can expect the ℓ_1 relaxation to perform as well as solving the ℓ_0-problem. More precisely, given an observation vector $\mathbf{y} \in \mathbb{R}^N$ and a design matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$, let us consider the two problems

$$\min_{\beta \in \mathbb{R}^p : \mathbf{X}\beta = \mathbf{y}} \|\beta\|_0 \tag{84}$$

and

$$\min_{\beta \in \mathbb{R}^p : \mathbf{X}\beta = \mathbf{y}} \|\beta\|_1. \tag{85}$$

The above linear program (LP) (85) is also known as the *basis pursuit LP*. Suppose that the ℓ_0-based problem (84) has a unique optimal solution, say $\beta^* \in \mathbb{R}^p$. Our interest is in understanding when β^* is also the unique optimal solution of the ℓ_1-based problem (85), in which case we say that the basis pursuit LP is equivalent to ℓ_0-recovery. Remarkably, there exists a very simple necessary and sufficient condition on the design matrix $\mathbf{X}$ for this equivalence to hold.

Definition 8.5 (Exact recovery property)**.** An $N \times p$ design matrix $\mathbf{X}$ is said to satisfy the *exact recovery property* for $S \subset \{1, \ldots, p\}$ (or S-ERP) if every $\beta^* \in \mathbb{R}^p$ supported on S uniquely minimizes $\|\beta\|_1$ subject to $\mathbf{X}\beta = \mathbf{X}\beta^*$.

For a given subset $S \subset \{1, 2, \ldots, p\}$, let us define the following set:

$$\mathcal{C}(S) := \{\beta \in \mathbb{R}^p : \|\beta_{S^c}\|_1 \leq \|\beta_S\|_1\}.$$

The set $\mathcal{C}(S)$ is a cone but is not convex (Exercise (HW4): show this), containing all vectors that are supported on S, and other vectors as well. Roughly, it corresponds to the cone of vectors that have most of their mass allocated to S. Recall that we have already seen the importance of the set $\mathcal{C}(S)$ in the recovery of β^* and $\mathbf{X}\beta^*$ using the lasso estimator.

Given a matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$, its *nullspace* is given by

$$\text{null}(\mathbf{X}) = \{\beta \in \mathbb{R}^p : \mathbf{X}\beta = \mathbf{0}\}.$$

Definition 8.6 (Restricted nullspace property)**.** For a given subset $S \subset \{1, 2, \ldots, p\}$, we say that the design matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$ satisfies the *restricted nullspace property* over S, denoted by $\mathrm{RN}(S)$, if

$$\mathrm{null}(\mathbf{X}) \cap \mathcal{C}(S) = \{0\}.$$

In words, the $\mathrm{RN}(S)$ property holds when the only element of the cone $\mathcal{C}(S)$ that lies within the nullspace of $\mathbf{X}$ is the all-zeroes vector. The following theorem highlights the connection between the exact recovery property and the restricted nullspace property.

Theorem 8.7. The matrix $\mathbf{X}$ is S-ERP if and only if it is $\mathrm{RN}(S)$.

Since the subset S is not known in advance — indeed, it is usually what we are trying to determine — it is natural to seek matrices that satisfy a uniform version of the restricted nullspace property. For instance, we say that the *uniform RN property* of order k holds if $\mathrm{RN}(S)$ holds for all subsets of $\{1, \ldots, p\}$ of size at most k. In this case, we are guaranteed that the ℓ_1-relaxation succeeds for any vector supported on any subset of size at most k.

Proof of Theorem 8.7. First, suppose that $\mathbf{X}$ satisfies the $\mathrm{RN}(S)$ property. Let $\beta^* \in \mathbb{R}^p$ be supported on S and let $\mathbf{y} = \mathbf{X}\beta^*$. Let $\hat{\beta} \in \mathbb{R}^p$ be any optimal solution to the basis pursuit LP (85), and define the error vector $\hat{\nu} := \hat{\beta} - \beta^*$. Our goal is to show that $\hat{\nu} = 0$, and in order to do so, it suffices to show that $\hat{\nu} \in \mathrm{null}(\mathbf{X}) \cap \mathcal{C}(S)$. On the one hand, since β^* and $\hat{\beta}$ are optimal (and hence feasible) solutions to the ℓ_0 and ℓ_1-problems, respectively, we are guaranteed that $\mathbf{X}\beta^* = \mathbf{y} = \mathbf{X}\hat{\beta}$, showing that $X\hat{\nu} = \mathbf{0}$. On the other hand, since β^* is also feasible for the ℓ_1-based problem (85), the optimality of $\hat{\beta}$ implies that $\|\hat{\beta}\|_1 \leq \|\beta^*\|_1 = \|\beta^*_S\|_1$. Writing $\hat{\beta} = \beta^* + \hat{\nu}$, we have

$$\|\beta^*_S\|_1 \geq \|\hat{\beta}\|_1 = \|\beta^*_S + \hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1 \geq \|\beta^*_S\|_1 - \|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1.$$

Rearranging terms, we find that $\hat{\nu} \in \mathcal{C}(S)$. Since $\mathbf{X}$ satisfies the $\mathrm{RN}(S)$ condition by assumption, we conclude that $\hat{\nu} = 0$, as required.

Suppose now that $\mathbf{X}$ is S-ERP. We will use the method of contradiction here to show that $\mathbf{X}$ is $\mathrm{RN}(S)$. Thus, assume that $\mathbf{X}$ is not $\mathrm{RN}(S)$. Then there exists $h \neq 0 \in \mathrm{null}(\mathbf{X})$ such that

$$\|h_S\|_1 \geq \|h_{S^c}\|_1. \tag{86}$$

Set $\beta^* \in \mathbb{R}^p$ such that $\beta^*_S = h_S$ and $\beta^*_{S^c} = 0$. Then β^* is supported on S. Thus, by the S-ERP β^* uniquely minimizes $\|\beta\|_1$ subject to $\mathbf{X}\beta = \mathbf{X}\beta^* := \mathbf{y}$.

Set $\beta^+ \in \mathbb{R}^p$ such that $\beta^+_S = 0$ and $\beta^+_{S^c} = -h_{S^c}$. Then observe that $\mathbf{X}\beta^* = \mathbf{X}\beta^+$ as

$$\mathbf{X}\beta^* = \mathbf{X}_S h_S = -\mathbf{X}_{S^c} h_{S^c} = \mathbf{X}\beta^+$$

(recall that $\mathbf{X}h = 0$). Thus, $\beta^+ \in \mathbb{R}^p$ is a feasible solution to the optimization problem: $\min \|\beta\|_1$ subject to $\mathbf{X}\beta = \mathbf{X}\beta^* = \mathbf{y}$. Thus, by the uniqueness of β^*, $\|\beta^*\|_1 < \|\beta^+\|_1$ which is equivalent to $\|h_S\|_1 < \|h_{S^c}\|_1$ — a contradiction to (86). This completes the proof. □

8.4.1 Sufficient conditions for restricted nullspace

Of course, in order for Theorem 8.7 to be useful in practice, we need to verify the RN property. A line of work has developed various conditions for certifying the uniform RN property. The simplest and historically earliest condition is based on the *pairwise incoherence*

$$r(\mathbf{X}) := \max_{j \neq k \in \{1,\ldots,p\}} \frac{|\langle \mathbf{x}_j, \mathbf{x}_k \rangle|}{\|\mathbf{x}_j\|_2 \|\mathbf{x}_k\|_2}.$$

For centered $\mathbf{x}_j$ this is the maximal absolute pairwise correlation. When $\mathbf{X}$ is rescaled to have unit-norm columns, an equivalent representation is given by $r(\mathbf{X}) = \max_{j \neq k} |\langle \mathbf{x}_j, \mathbf{x}_k \rangle|$, which illustrates that pairwise incoherence measures how close the Gram matrix $\mathbf{X}^\top \mathbf{X}$ is to the p-dimensional identity matrix in an element-wise sense.

The following result shows that having a low pairwise incoherence is sufficient to guarantee exactness of the basis pursuit LP.

Proposition 8.8 (Pairwise incoherence implies RN)**.** Suppose that for some integer $k \in \{1, 2, \ldots, p\}$, the pairwise incoherence satisfies the bound $r(\mathbf{X}) < \frac{1}{3k}$. Then $\mathbf{X}$ satisfies the uniform RN property of order k, and hence, the basis pursuit LP is exact for all vectors with support at most k.

Proof. See [5][Section 10.4.3] for a proof of this claim. □

An attractive feature of pairwise incoherence is that it is easily computed; in particular, in $O(Np^2)$ time. A disadvantage is that it provides very conservative bounds that do not always capture the actual performance of ℓ_1-relaxation in practice.

Definition 8.9 (Restricted isometry property)**.** For a tolerance $\delta \in (0, 1)$ and integer $k \in \{1, 2, \ldots, p\}$, we say that the restricted isometry property RIP(k, δ) holds if

$$\|\mathbf{X}^\top \mathbf{X} - I_k\|_{op} \leq \delta$$

for all subsets $S \subset \{1, 2, \ldots, p\}$ of cardinality k. We recall here that $\|\cdot\|_{op}$ denotes the operator norm, or maximal singular value of a matrix.

Thus, we see that RIP(k, δ) holds if and only if for all subsets S of cardinality k, we have

$$\frac{\|\mathbf{X}_S u\|_2^2}{\|u\|_2^2} \in [1-\delta, 1+\delta], \qquad \text{for all } \; u \neq 0 \in \mathbb{R}^k;$$

hence the terminology of restricted isometry. The following result, which we state without any proof, shows that the RIP is a sufficient condition for the RN property to hold.

Proposition 8.10 (RIP implies RNP)**.** If RIP$(2k, \delta)$ holds with $\delta < 1/3$, then the uniform RN property of order k holds, and hence the ℓ_1-relaxation is exact for all vectors supported on at most k elements.

References

[1] Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B 57*(1), 289–300.

[2] Berlinet, A. and C. Thomas-Agnan (2004). *Reproducing kernel Hilbert spaces in probability and statistics.* Kluwer Academic Publishers, Boston, MA. With a preface by Persi Diaconis.

[3] Efron, B. (2010). *Large-scale inference*, Volume 1 of *Institute of Mathematical Statistics (IMS) Monographs.* Cambridge University Press, Cambridge. Empirical Bayes methods for estimation, testing, and prediction.

[4] Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist. 9*(6), 1218–1228.

[5] Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity.*

[6] Keener, R. W. (2010). *Theoretical statistics.* Springer Texts in Statistics. Springer, New York. Topics for a core course.

[7] Kimeldorf, G. and G. Wahba (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl. 33*, 82–95.

[8] Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York.

[9] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist. 9*(2), 65–78.

[10] Schölkopf, B., R. Herbrich, and A. J. Smola (2001). A generalized representer theorem. In *Computational learning theory*, pp. 416–426. Springer.

[11] Sen, B., M. Banerjee, and M. Woodroofe (2010). Inconsistency of bootstrap: the Grenander estimator. *Ann. Statist. 38*(4), 1953–1977.

[12] Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist. 12*(4), 1285–1297.

[13] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*(1), 267–288.

[14] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[15] van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[16] Vapnik, V. and A. Lerner (1963). Pattern recognition using generalized portrait method. *Automation and remote control 24*, 774–780.

[17] Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist. 27*(5), 1564–1599.

[18] Zhang, Y., M. J. Wainwright, and M. I. Jordan (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *arXiv preprint arXiv:1402.1918*.