Homework 1

Advanced Topics in Statistical Learning, Spring 2023

Due Friday February 10 at 5pm

1 Mathematical statistics warm-up [15 points]

(a) Suppose that $X_n \geq 0$ and $\mathbb{E}[X_n] = O(r_n)$. Prove that $X_n = O_p(r_n)$. [1 pt]

(b) Suppose that $X_n \geq 0$ and $X_n = O_p(r_n)$. Give an example to show that in general, this does not imply that $\mathbb{E}[X_n] = O(r_n)$. [1 pt]

(c) Prove that for $X \geq 0$, it holds that

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\, dt.$$

You may assume that X is continuously distributed and hence has a probability density function. [2 pts]

(d) Suppose that $X_n \geq 0$ and $X_n = O_p(r_n)$, the latter bound holding "exponentially fast", meaning that there are constants $\gamma_0, n_0 > 0$ such that for all $\gamma \geq \gamma_0$ and $n \geq n_0$, we have

$$X_n \leq \gamma r_n, \quad \text{with probability at least } 1 - \exp(-\gamma).$$

Prove that $\mathbb{E}[X_n] = O(r_n)$. Hint: use the formulation for $\mathbb{E}[X_n]$ from the last question. [3 pts]

(e) Let $X_1, \dots, X_n \sim P$, i.i.d., with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}[X_i]$. Define

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

(i) Prove that $s_n^2 \xrightarrow{p} \sigma^2$. [1 pt]

(ii) Prove that $\sqrt{n}(\bar{X}_n - \mu)/s_n \xrightarrow{d} N(0,1)$. [1 pt]

(f) Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$.

(i) Prove that $\mathbb{E}[(Y - f(X))^2]$ is minimized at $f(x) = \mathbb{E}[Y|X = x]$. [1 pt]

(ii) Prove that $\mathbb{E}[(Y - X^T\beta)^2]$ is minimized at $\beta = \Sigma^{-1}\alpha$, where $\Sigma = \mathbb{E}[XX^T]$ and $\alpha = \mathbb{E}[YX]$. [1 pt]

(g) This part will involve a small bit of coding. Attach your code in an appendix.

(i) Simulate Brownian motion on $[0,1]$, and a Brownian bridge on $[0,1]$, and plot them. [1 pt]

(ii) Simulate the 95th percentile of the supremum of the Brownian bridge, i.e., the value q such that

$$\mathbb{P}\Big(\sup_{t \in [0,1]} B(t) \geq q \Big) = 0.05.$$

where $B(t)$, $t \in [0,1]$ is the Brownian bridge. [1 pt]

(iii) Draw $X_1, \dots, X_n \sim F$ from any distribution F of your liking (uniform, normal, etc.), calculate the Kolmogorov-Smirnov (KS) test statistic

$$T = \sqrt{n} \sup_x |F_n(x) - F(x)|,$$

where F_n is the empirical distribution of $X_1, \dots, X_n$, and calculate the proportion of times out of (say) 1000 repetitions that T exceeds the threshold q computed in part (ii). [2 pts]

2 Risk analysis for least squares [15 points]

In this exercise, we will work on risk calculations for least squares regression.

(a) First, we start with an algebraic fact. Suppose that $A, B \succeq 0$, which we write to mean that are positive semidefinite matrices (symmetric with nonnegative eigenvalues). Prove that $\mathrm{tr}(AB) \geq 0$. [2 pts]

Hint: there are many ways to prove this, but for one, take an eigendecomposition of B, and expand the trace as a sum of products involving its eigenvectors.

(b) For this part and the next, suppose that we observe i.i.d. $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$. We write $f(x) = \mathbb{E}[y_i | x_i = x]$, $\epsilon_i = y_i - f(x_i)$, and assume that each $x_i \perp\!\!\!\perp \epsilon_i$. We denote $\sigma^2 = \mathrm{Var}[\epsilon_i]$.

Let $Y \in \mathbb{R}^n$ be the response vector and $X \in \mathbb{R}^{n \times d}$ the predictor matrix (whose i^{th} row is x_i). Let $\hat{\beta} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y$ be the least squares solution of Y on X (where we assume that $X^T X$ is invertible, which requires $d \leq n$), and let $\hat{f}(x) = x^\mathsf{T} \hat{\beta}$.

Follow/reproduce the calculations in the review lecture to show that [3 pts]

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(\hat{f}(x_i) \,|\, X)\right] = \sigma^2 \frac{d}{n},$$

and that, for an independent draw x_0 from the predictor distribution, [3 pts]

$$\mathbb{E}[\mathrm{Var}(\hat{f}(x_0) \,|X, x_0)] = \frac{\sigma^2}{n}\mathrm{tr}\Big(\mathbb{E}[X^\mathsf{T} X]\, \mathbb{E}[(X^\mathsf{T} X)^{-1}]\Big).$$

Therefore, using part (a), argue that [1 pt]

$$\mathbb{E}[\mathrm{Var}(\hat{f}(x_0) \,|X, x_0)] \geq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(\hat{f}(x_i) \,|\, X)\right].$$

Hint: the calculations in lecture assumed the underlying model was linear and hence the bias (both in- and out-of-sample) was zero. But if you look back carefully, the variance calculations are unaffected by whether the true mean is linear or not.

(c) Follow/reproduce the calculations leading up to Theorem 1 in Rosset and Tibshirani (2020) to prove the inequality: [6 pts]

$$\mathbb{E}[\mathrm{Bias}^2(\hat{f}(x_0) \,|X, x_0)] \geq \mathbb{E}\left[\frac{1}{n}\sum_{\ell=1}^n \mathrm{Bias}^2(\hat{f}(x_i) \,|\, X)\right].$$

Note that you have shown that

$$\mathrm{Risk}(\hat{f}) \geq \mathbb{E}[\mathrm{Risk}(\hat{f}; x_{1:n})].$$

In other words, the out-of-sample risk of least squares is always at least as large as the in-sample risk (integrated over the feature values). To emphasize, this assumes nothing really at all (i.e., no underlying linear model) about the data model, except for the independence of x_i and ϵ_i.

(d) As a bonus, prove or disprove: there is a predictor distribution such that we get an equality in the last display, i.e., the out-of-sample and in-sample risks are equal. Note that we are still talking about standard least squares regression, so we are restricting attention to distributions such that $X^\mathsf{T} X$ is almost surely invertible.

3 Risk analysis for wavelet denoising [22 points]

In this exercise, we will analyze the risk of wavelet denoising.

(a) Assume for now that we observe data according to the normal sequence model

$$z_\ell = \theta_\ell + \delta_\ell, \quad \ell = 1, \dots, N, \tag{1}$$

where $\delta_\ell \sim N(0, \tau^2)$, independently, for $\ell = 1, \dots, N$. Consider the soft-thresholding estimator,

$$\hat{\theta}_\ell = S_\lambda(z_\ell) = \begin{cases} z_\ell - \lambda & \text{if } z_\ell > \lambda \\ 0 & \text{if } |z_\ell| \le \lambda \\ z_\ell + \lambda & \text{if } z_\ell < -\lambda \end{cases}, \quad \ell = 1, \dots, N.$$

Here $\lambda \ge 0$ is a tuning parameter. For arbitrary λ, prove that we have the exact risk expression: [3 pts]

$$\mathbb{E}\|\theta - \hat{\theta}\|_2^2 = \sum_{\ell=1}^N r(\theta_\ell, \lambda),$$

where

$$r(\mu, \lambda) = \mu^2 \int_{\frac{-\lambda-\mu}{\tau}}^{\frac{\lambda-\mu}{\tau}} \phi(z)\, dz + \int_{\frac{\lambda-\mu}{\tau}}^{\infty} (\tau z - \lambda)^2 \phi(z)\, dz + \int_{-\infty}^{\frac{-\lambda-\mu}{\tau}} (\tau z + \lambda)^2 \phi(z)\, dz,$$

and ϕ denotes the standard (univariate) normal density function.

(b) Prove that for $\lambda = \tau\sqrt{2 \log N}$, we have the risk upper bound: [5 pts]

$$\mathbb{E}\|\theta - \hat{\theta}\|_2^2 \le \tau^2 + (2 \log N + 1) \sum_{\ell=1}^N \min\{\theta_\ell^2, \tau^2\}.$$

Hint: start with $\tau^2 = 1$ for simplicity. Prove that, for any $\mu, \lambda \ge 0$, we have $0 \le \partial r(\mu, \lambda)/\partial \mu \le 2\mu$. From this, argue that $r(\mu, \lambda)$ is monotone increasing in μ, and further

$$r(\mu, \lambda) \le r(0, \lambda) + \min\{\mu^2, r(\infty, \lambda)\}.$$

Then, derive upper bounds on $r(0, \lambda)$ and $r(\infty, \lambda)$ (for the former you can use Mills' ratio, for the latter you can use direct arguments) to give

$$r(\mu, \lambda) \le e^{-\lambda^2/2} + \min\{\mu^2, 1 + \lambda^2\}.$$

Plug in $\lambda = \sqrt{2 \log N}$; show that an analogous bound holds for general $\tau^2 > 0$; and sum the bound over $\mu = \theta_\ell$, $\ell = 1, \dots, N$ to give the result.

(c) Now consider the nonparametric regression model

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \dots, n, \tag{2}$$

where $\epsilon_i \sim N(0, \sigma^2)$, independently, for $i = 1, \dots, n$, and $x_i \in [0, 1]$, $i = 1, \dots, n$ are fixed (more on them later). We are going to analyze the L^2 risk of a wavelet smoothing estimator $\hat{f}$,

$$\mathbb{E}\|f - \hat{f}\|_2^2 = \mathbb{E}\bigg[\int_0^1 (f(x) - \hat{f}(x))^2\, dx\bigg].$$

The estimator $\hat{f}$ will be defined by

$$\hat{f}(x) = \sum_{j,k} \tilde{\theta}_{jk}(y) \psi_{jk}, \tag{3}$$

where each ψ_{jk} is a Haar wavelet function, and each $\tilde{\theta}_{jk}(y)$ is a noisy empirical wavelet coefficient.

We begin with a simple Haar calculation. To recall the Haar basis on $[0, 1]$, first define $\psi(x) = 1\{x \in (0, 1/2]\} - 1\{x \in (1/2, 1]\}$. Then the Haar basis is given by the collection

$$1,\ \psi_{jk},\ \text{for } k = 0, \dots, 2^j - 1 \text{ and } j = 0, 1, 2, \dots,$$

where $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. (For notational convenience, we let $\psi_{-10} = 1$, and implicitly index all basis calculations starting from $j = -1$.) Verify that this collection is orthonormal in L^2: show that the functions are pairwise orthogonal and unit norm, with respect to the L^2 inner product on $[0, 1]$, [2 pts]

$$\langle g, h \rangle = \int_0^1 g(x)h(x)\, dx.$$

(Accordingly the L^2 norm is simply given by $\|g\|_2^2 = \langle g, g \rangle = \int_0^1 g(x)^2\, dx$.)

(d) Explain why it is that we can write [1 pt]

$$\|f - \hat{f}\|_2^2 = \sum_{j,k} (\theta_{jk}(f) - \tilde{\theta}_{jk}(y))^2,$$

where the wavelet coefficients of f are

$$\theta_{jk}(f) = \langle f, \psi_{jk} \rangle = \int_0^1 f(x)\psi_{jk}(x)\, dx,$$

and $\tilde{\theta}_{jk}(y)$ are the coefficients to define the estimator $\hat{f}$ in its Haar basis expansion (3).

Hint: by orthonormality, observe that $f = \sum_{j,k} \theta_{jk}(f)\psi_{jk}$. It suffices to just name the theorem that relates the L^2 norm of a function to the norm of its coefficients.

(e) We define the last few parts needed to understand $\hat{f}$ and analyze its risk. For each j, k, we define the empirical wavelet coefficient

$$\tilde{\theta}_{jk}(f) = \frac{1}{n}\sum_{i=1}^n f(x_i)\psi_{jk}(x_i).$$

We also define a noisy empirical wavelet coefficient

$$\tilde{\theta}_{jk}(y) = \begin{cases} S_\lambda\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^n y_i\psi_{jk}(x_i)\right) & \text{if } j \le j^* \\ 0 & \text{if } j > j^* \end{cases},$$

where S_λ is the soft-thresholding operator, as before, and j^* is a truncation level, to be chosen.

By part (d), and the inequality $(a + b)^2 \le 2a^2 + 2b^2$ (applied twice), we have

$$\mathbb{E}\|f - \hat{f}\|_2^2 \le \underbrace{2\sum_{j>j^*,k} \theta_{jk}^2(f)}_{e_1} + \underbrace{4\sum_{j\le j^*,k} (\theta_{jk}(f) - \tilde{\theta}_{jk}(f))^2}_{e_2} + \underbrace{4\,\mathbb{E}\left[\sum_{j\le j^*,k} (\tilde{\theta}_{jk}(f) - \tilde{\theta}_{jk}(y))^2\right]}_{e_3}.$$

We can interpret e_1 as the *truncation error*, e_2 as the *discretization error* (between population and empirical wavelet coefficients), and e_3 as the *estimation error* (in estimating the empirical wavelet coefficients from noisy data).

Denote by $\theta_{j\cdot}(f)$ the vector $(\theta_{jk}(f) : k = 0, \dots, 2^j - 1)$. Assume that $\mathrm{TV}(f) \le 1$, and assume that the design points $x_i = i/n$, $i = 1, \dots, n$ are evenly-spaced. It can be shown that

$$\|\theta_{j\cdot}(f)\|_1 \le c_1 2^{-j/2}, \quad \|\tilde{\theta}_{j\cdot}(f)\|_1 \le c_2 2^{-j/2}, \quad \text{and} \quad \|\theta_{j\cdot}(f) - \tilde{\theta}_{j\cdot}(f)\|_1 \le c_3\frac{2^{j/2}}{n}, \tag{4}$$

for constants $c_1, c_2, c_3 > 0$. Use the first and third inequalities to show that there is a truncation level j^* such that sum of truncation and discretization errors satisfy $e_1 + e_2 \le C/n$, for another constant $C > 0$. [2 pts]

(f) It remains to study the estimation error. Assume that n is a power of 2. Show that, starting from the nonparametric regression model (2), we may transform this to a model of the form [3 pts]

$$z_\ell = \tilde{\theta}_\ell(f) + \delta_\ell, \quad \ell = 1, \dots, n,$$

where $\delta_\ell \sim N(0, \sigma^2/n)$, independently, for $\ell = 1, \dots, n$. Note that here, in indexing wavelet coefficients, we collapse the pair j, k into a single index ℓ.

Hint: use the appropriate truncation level j^*, from part (e), and only consider $j \le j^*$. Then define a matrix Ψ with elements $[\Psi]_{i\ell} = \psi_\ell(x_i)/n$, where in indexing the Haar wavelets, we again collapse the pair j, k into a single index ℓ. Using the fact we have an evenly-spaced design $x_i = i/n$, $i = 1, \dots, n$, show that $\Psi\Psi^\mathsf{T} = \frac{1}{n} I$, where I is the $n \times n$ identity matrix.

(g) Finally, note that from the transformation in part (f) you have brought yourself back to the problem studied in parts (a), (b): soft-thresholding under the sequence model (1), with noise level $\tau^2 = \sigma^2/n$.

From the risk bound from part (b), note that we have

$$\mathbb{E}\left[\sum_{j \le j^*, k} (\tilde{\theta}_{jk}(f) - \tilde{\theta}_{jk}(y))^2\right] \le \frac{\sigma^2}{n} + (2\log n + 1) \sum_{j \le j^*, k} \min\left\{\tilde{\theta}_{jk}^2(f), \frac{\sigma^2}{n}\right\}.$$

Use the second inequality in (4), on the empirical wavelet coefficients, to establish that for each j, [4 pts]

$$\sum_k \min\left\{\tilde{\theta}_{jk}^2(f), \frac{\sigma^2}{n}\right\} \le C\frac{\sigma^2}{n} 2^j \min\left\{1, 2^{-3j/2}\frac{\sqrt{n}}{\sigma}\right\},$$

for a constant $C > 0$. Show that gives the estimation error bound, [2 pts]

$$e_3 \le C \log n \left(\frac{\sigma^2}{n}\right)^{2/3}.$$

for a constant $C > 0$, redefined as needed.

Hint: the first bound (second-to-last display) is a bit tricky, whereas the second (last display) is more of a straight algebraic calculation, summing the first bound over j. To prove the first, argue that

$$\sup_{\|\tilde{\theta}_{j\cdot}\|_1 \le c_j} \sum_k \min\left\{\tilde{\theta}_{jk}^2, \frac{\sigma^2}{n}\right\}$$

will be achieved at a vector $\tilde{\theta}_{j\cdot}$ for which each entry is equal to 0 or $\sigma/\sqrt{n}$, except for (possibly) one entry, which is defined so that we hit the constraint $\|\tilde{\theta}_{j\cdot}\|_1 = c_j$. For the current problem, note that we have $c_j = c_2 2^{-j/2}$.

Concluding note: the risk bound you have shown, redefining the constant $C > 0$ as needed, is

$$\mathbb{E}\|f - \hat{f}\|_2^2 \le C\left[\frac{1}{n} + \log n \left(\frac{\sigma^2}{n}\right)^{2/3}\right],$$

for estimating a function with $\mathrm{TV}(f) \le 1$ using Haar wavelet denoising. This is minimax rate optimal for the class of functions with bounded TV, ignoring log factors (which could be removed from the upper bound with a slightly finer analysis).

4 Bonus: minimax lower bounds for Hölder classes

This is an optional bonus problem. Consider the d-dimensional Hölder ball

$$C^k(L; [0,1]^d) = \left\{f : [0,1]^d \to \mathbb{R} \,:\, |D^\alpha f(x) - D^\alpha f(z)| \le L\|x - z\|_2 \text{ for all } |\alpha| < k, \text{ and } x, z \in [0,1]^d\right\},$$

where $k \geq 1$ is an integer and $L > 0$ is a constant (not growing with n), and consider the data model

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0,1), \quad i = 1, \dots, n.$$

You may either assume a Random-X model: (x_i, ϵ_i), $i = 1, \dots, n$ are i.i.d. with each $x_i \sim \mathrm{Unif}[0,1]^d$ and $x_i \perp \epsilon_i$, or a Fixed-X model: x_i, $i = 1, \dots, n$ are fixed and laid out on a regular lattice in $[0,1]^d$, and ϵ_i, $i = 1, \dots, n$ are i.i.d. Whatever you prefer.

Generalize the calculations given in the minimax theory lecture to show that there exists a constant $c > 0$ (which may depend on L) such that

$$\inf_{\hat{f}} \sup_{f \in C^k_d(L;[0,1]^d)} \mathbb{E}\|f - \hat{f}\|_2^2 \geq c n^{-2k/(2k+d)}.$$

To be clear, here $\|g\|_2^2 = \int_{[0,1]^d} g(x)^2 \, dx$. Make sure to write the full calculation from start to finish (even detailing the steps that may be analogous to those in lecture).

Explain why this certifies that the upper bound given in the last question for estimation over a univariate TV ball is rate optimal (ignoring log factors).

References

Saharon Rosset and Ryan J. Tibshirani. From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 15(529):138–151, 2020.