Exercises for Theory of Statistics (Stats300b)

John Duchi

Winter Quarter 2021

Contents

1 Background

Question 1.1: Let p_n and q_n be (a sequence of) densities with respect to some base measure μ. Define the *likelihood ratio* as

$$L_n(x) := \begin{cases} q_n(x)/p_n(x) & \text{if } p_n(x) > 0 \\ 1 & \text{if } p_n(x) = q_n(x) \\ +\infty & \text{otherwise.} \end{cases}$$

Let X_n be distributed according to the distribution with density p_n. Show that $L_n(X_n)$ is uniformly tight.

Question 1.2: Let X_n be uniformly distributed on the set $\{1/n, 2/n, \ldots, 1\}$ and X be uniformly distributed on $[0, 1]$. Show that $X_n \xrightarrow{d} X$ as $n \to \infty$. Does $X_n \xrightarrow{p} X$?

Question 1.3: Let $F_n : \mathbb{R} \to [0, 1]$ be a sequence of non-decreasing functions converging uniformly to some $F : \mathbb{R} \to [0, 1]$, a continuous and strictly increasing function that is onto $(0, 1)$. Show that for all $\epsilon \in (0, \frac{1}{2})$, we have

$$\sup_\alpha \left\{ \left| F_n^{-1}(\alpha) - F^{-1}(\alpha) \right| : \epsilon \le \alpha \le 1 - \epsilon \right\} \to 0.$$

Here we define $G^{-1}(\alpha) = \inf\{x \in \mathbb{R} : G(x) \ge \alpha\}$ for any non-decreasing function G.

Question 1.4: Let $X_i \in \mathbb{R}$ be i.i.d. according to a distribution with CDF F, which for simplicity we assume to be continuous. Let F_n be the empirical CDF given by $F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \le t\}$. Without appealing to the Glivenko-Cantelli theorem, show that

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{p} 0.$$

Hint: Use the fact that F and F_n are non-decreasing and consider subsets of $\mathbb{R}$.

Question 1.5: Let $X_1, \ldots, X_n$ be drawn i.i.d. $\mathsf{Beta}(\theta, 1)$ for some $\theta > 0$. Letting $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ denote the sample mean and $\widehat{\theta}_n = \frac{\overline{X}_n}{1 - \overline{X}_n}$, give the limiting distribution of the sequence

$$\sqrt{n} \left(\widehat{\theta}_n - \theta \right)$$

or demonstrate that it does not exist.

Question 1.6: Let $\{X_i^n\}$, $i = 1, \ldots, n$ and $n \in \mathbb{N}$ be a triangular array of random variables, where $X_i^n \overset{\text{iid}}{\sim} \mathsf{Bernoulli}(\theta_n)$ and $\theta_n = 1/\sqrt{n}$. Define $\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i^n$. Is $n^{3/4}(\widehat{\theta}_n - \theta_n)$ asymptotically normal? If so, give the limiting mean and variance, and if not, demonstrate why not.

Question 1.7 (Moment generating function background)**:** A mean zero random variable X is *σ^2-sub-Gaussian* if $\mathbb{E}[\exp(\lambda X)] \le \exp(\frac{\lambda^2 \sigma^2}{2})$ for all $\lambda \in \mathbb{R}$.

(a) Show that if Z is mean-zero Gaussian with variance σ^2, then $\mathbb{E}[\exp(\lambda Z)] = \exp(\frac{\lambda^2 \sigma^2}{2})$.

(b) Show that if X_i, $i = 1, \ldots, n$, are i.i.d. mean zero σ^2-sub-Gaussian random variables, then $\mathbb{E}[\max_{i \le n} X_i] \le \sqrt{2\sigma^2 \log n}$.

Question 1.8: Let $\|\cdot\|_{\rm TV}$ be the total variation distance, that is,

$$\|P - Q\|_{\rm TV} = \sup_A |P(A) - Q(A)|$$

for probability distributions P and Q. Let μ be any measure such that $P \ll \mu$ and $Q \ll \mu$, and let p and q be the densities of P and Q with respect to μ. Show the following equalities, where $\wedge$ denotes min and $\vee$ denotes max.

(a) $2\|P - Q\|_{\rm TV} = \int |p - q| d\mu$.

(b) $\sup_{\|f\|_\infty \le 1} \int f(x)(dP(x) - dQ(x)) = 2\|P - Q\|_{\rm TV}$.

(c) $2\|P - Q\|_{\rm TV} = \int (p - q)_+ \, d\mu + \int (q - p)_+ \, d\mu$.

(d) $\|P - Q\|_{\rm TV} = \int (p \vee q) d\mu - 1$.

(e) $\|P - Q\|_{\rm TV} = 1 - \int (p \wedge q) d\mu$.

Question 1.9: Let P, Q have densities p, q w.r.t. a measure μ. The Hellinger distance $d_{\rm hel}$ between P and Q is defined by (its square)

$$d_{\rm hel}^2(P, Q) := \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 d\mu.$$

Show that

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

Question 1.10 (Reproducing kernel Hilbert spaces)**:** A vector space $\mathcal{H}$ is a *Hilbert space* if it is a complete normed vector space, with norm $\|\cdot\|$, and there is an inner product $\langle \cdot, \cdot \rangle$ such that $\langle u, u \rangle = \|u\|^2$ for $u \in \mathcal{H}$. In this question, we will investigate the construction of one type of Hilbert space known as a *reproducing kernel Hilbert space* (RKHS).

An RKHS $\mathcal{H}$ is a collection of functions $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a measurable space, equipped with an inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{H}$. In addition to the inner product $\langle \cdot, \cdot \rangle$, such Hilbert spaces are equipped with what is known as the *representer of evaluation*, that is, a collection of functions r_x indexed by $x \in \mathcal{X}$ such that $r_x \in \mathcal{H}$ for each x, i.e. $r_x : \mathcal{X} \to \mathbb{R}$, and

$$\langle f, r_x \rangle = f(x)$$

for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$.

Let $\mathcal{X}$ be a (measurable) space. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *kernel* if it is positive definite, meaning that for all $n \in \mathbb{N}$ and distinct $x_i \in \mathcal{X}$, the kernel (or *Gram*) matrix

$$K := \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \ddots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix}$$

is symmetric and positive semidefinite (PSD) (so $k(x, x') = k(x', x)$ for all x, x'). That is, for all $\alpha \in \mathbb{R}^n$, we have $\alpha^\top K \alpha \ge 0$. Now, consider the class of functions $\mathcal{H}_0$, where $f \in \mathcal{H}_0$ maps $\mathcal{X} \to \mathbb{R}$, defined by the linear span of $\{k(x, \cdot) \mid x \in \mathcal{X}\}$. (That is, if $f \in \mathcal{H}_0$ then $f(x) = \sum_{i=1}^m \alpha_i k(x, x_i)$ for

some $m \in \mathbb{N}$, $\alpha \in \mathbb{R}^m$, $x_i \in \mathcal{X}$.) For $f(\cdot) = \sum_{i=1}^m \alpha_i k(\cdot, x_i)$ and $g(\cdot) = \sum_{j=1}^n \beta_j k(\cdot, x_j)$, define an inner product on $\mathcal{H}_0$ by

$$\langle f, g \rangle = \left\langle \sum_{i=1}^m \alpha_i k(\cdot, x_i), \sum_{j=1}^n \beta_j k(\cdot, z_j) \right\rangle := \sum_{i,j} \alpha_i \beta_j k(x_i, z_j).$$

Define $\mathcal{H}$ to be the completion of $\mathcal{H}_0$ for this inner product, that is, we define $f \in \mathcal{H}$ by

$$f(x) := \lim_{n \to \infty} f_n(x) \tag{1.1}$$

for Cauchy sequences $\{f_n\}_{n \in \mathbb{N}} \subset \mathcal{H}_0$ (which are Cauchy with respect to the inner product and norm on $\mathcal{H}_0$).

(a) Show that k has the *reproducing property* for $\mathcal{H}$, that is, for $f \in \mathcal{H}$ and $x \in \mathcal{X}$,

$$\langle f, k(\cdot, x) \rangle = f(x),$$

and that the limit (1.1) exists.

(b) Show that if $\mathcal{H}$ is an RKHS with representer of evaluation r_x, then

$$k(x, z) := \langle r_x, r_z \rangle$$

defines a valid kernel (i.e. it is positive definite and symmetric, and $\langle f, k(\cdot, x) \rangle = f(x)$ for all $x \in \mathcal{X}$).

Another view of RKHS's is in terms of *feature maps*. Let $\mathcal{F}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$, which we call the feature space. It is a theorem (known as Mercer's theorem) that if k is a positive definite kernel, there is a Hilbert space $\mathcal{F}$ and function $\varphi : \mathcal{X} \to \mathcal{F}$ such that $k(x, z) = \langle \varphi(x), \varphi(z) \rangle_{\mathcal{F}}$. Of course, by our construction above, given a PSD function (kernel) k and associated RKHS $\mathcal{H}$, we can always take $\varphi(x) = k(\cdot, x)$ and $\mathcal{F} = \mathcal{H}$ directly.

(c) Let $\varphi : \mathcal{X} \to \mathcal{F}$ for a Hilbert (feature) space $\mathcal{F}$. Show that $k(x, z) = \langle \varphi(x), \varphi(z) \rangle_{\mathcal{F}}$ is a valid kernel.

(d) Consider the Gaussian or Radial Basis Function (RBF), defined on $\mathbb{R}^d \times \mathbb{R}^d$ by

$$k(x, z) = \exp\left(-\frac{1}{2} \|x - z\|_2^2\right).$$

Exhibit a function $\phi : \mathbb{R} \to \mathbb{C}$ and distribution P on $\mathbb{R}^d$ such that

$$k(x, z) = \mathbb{E}_P[\phi(W^\top x)^* \phi(W^\top z)] \text{ for } W \sim P,$$

where * denotes the complex conjugate. Is k a valid kernel?

(e) Consider the min function, defined on $\mathbb{R}_+$ by

$$k(x, z) = \min\{x, z\}.$$

Exhibit a function $\phi : \mathbb{R}^2 \to \mathbb{R}$ such that

$$k(x, z) = \int_0^\infty \phi(x, t) \phi(z, t) dt.$$

Is k a valid kernel?

Question 1.11: Let X be a non-negative random variable. Show that for all $\theta \in [0,1]$, we have

$$\mathbb{P}(X \geq \theta \mathbb{E}[X]) \geq (1-\theta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

It may be easier to show the stronger inequality

$$\mathbb{P}(X \geq \theta \mathbb{E}[X]) \geq (1-\theta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2] - \theta(2-\theta)\mathbb{E}[X]^2}.$$

Question 1.12: Consider a triangular array of random vectors $\{X_i^n\}$, $i = 1, \ldots, n$, where for a dimension $d = d(n)$ we have $X_i^n \in \mathbb{R}^d$ and the dimension satisfies the limit $d/n = d(n)/n \to \gamma \in (0,1)$ as $n \to \infty$. For each $n \in \mathbb{N}$, let P_n be a distribution on $\mathbb{R}^d$, where for $X \sim P_n$ we have $\mathbb{E}[X] = \mathbf{0}_d$, the zeros vector in $\mathbb{R}^d$, $\mathbb{E}[XX^\top] = \mathrm{Cov}(X) = I_d$, and the coordinates of X are independent and satisfy $\mathbb{E}[(X)_j^4] = \tau^4$, that is, the coordinates have 4th moment $\tau^4 < \infty$. We assume that $X_i^n \overset{\text{iid}}{\sim} P_n$ for each n.

(a) Does

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i^n$$

have a limit distribution as $n \to \infty$? If so, what is it?

(b) Define the function

$$h(w,x,y,z) = \langle w, x \rangle \langle y, z \rangle = (w^\top x)(y^\top z).$$

For indices $i, j, k, l \in \mathbb{N}$ and $X_m \overset{\text{iid}}{\sim} P_n$ so that $X \in \mathbb{R}^d$, compute

$$\mathbb{E}[h(X_i, X_j, X_k, X_l)].$$

Hint: The only cases that matter are when $i = j = k = l$, $i = j$ and $k = l$, $i = k$ and $j = l$, or when at least one index is distinct from all the others.

(c) Define the statistic

$$V_n := \frac{1}{n^4} \sum_{i,j,k,l \leq n} h(X_i^n, X_j^n, X_k^n, X_l^n).$$

Show that

$$\mathbb{E}[V_n] = \gamma^2 + o(1)$$

as $n \to \infty$.

(d) Does

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i^n \right\|_2^2$$

have a limiting distribution as $n \to \infty$ (here, do not rescale any of the quantities above)? If so, what is it? *Hint:* One approach is to use the result of Question 1.11.

(e) Let $m \in \mathbb{N}$ be a fixed positive integer. Let $\{w_i^n\}_{i=1}^m \subset \mathbb{R}^d$ be a collection of m distinct vectors, defined for each n, with $\|w_i^n\|_2 = 1$ and $\langle w_i^n, w_j^n \rangle = 0$ if $i \neq j$.[1] Give the limiting distribution of

$$\max_{j \leq m} \left\langle w_j^n, \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i^n \right\rangle.$$

Hint: Use the Lindeberg central limit theorem.

[1] Assume n is large enough that this is possible.

2 Convex Analysis and Statistics

Question 2.1 (One-dimensional Jensen's inequality)**:** Let X be a real-valued random variable with $\mathbb{E}[|X|] < \infty$ and $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a convex function, meaning that

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y) \text{ for } x, y \in \mathbb{R}, \lambda \in [0,1].$$

Let $\operatorname{dom} f = \{x : f(x) < +\infty\}$, which is convex (it is an interval, a fact you may use to answer this question). Any convex function f is continuous on the interior of its domain.

(a) Let $\mathcal{I} \subset \mathbb{R}$ be a non-empty interval. Show that $f : \mathcal{I} \to \mathbb{R}$ is convex if and only if for any $x_0 \in \mathcal{I}$, the slope function

$$s(x) := \frac{f(x) - f(x_0)}{x - x_0}$$

is non-decreasing on $\mathcal{I} \setminus \{x_0\}$. [*Hint:* For $y \ge x > x_0$, write $x = \lambda y + (1-\lambda)x_0$]

(b) Define the left and right derivatives

$$f'_{\text{left}}(x) := \limsup_{t \downarrow 0} \frac{f(x) - f(x-t)}{t} \text{ and } f'_{\text{right}}(x) := \liminf_{t \downarrow 0} \frac{f(x+t) - f(x)}{t}.$$

Show that

$$f'_{\text{left}}(x) = \sup_{t > 0} \frac{f(x) - f(x-t)}{t} \text{ and } f'_{\text{right}}(x) = \inf_{t > 0} \frac{f(x+t) - f(x)}{t}.$$

(c) Show that $f'_{\text{left}}(x) \le f'_{\text{right}}(x)$.

(d) Show that if we define the subgradient set as the interval $\partial f(x) = [f'_{\text{left}}(x), f'_{\text{right}}(x)]$, then $f(y) \ge f(x) + g(y-x)$ for all $g \in \partial f(x)$.

We say that f is *strictly convex at the point* x if for all $x_0, x_1 \ne x$ and $\lambda \in (0,1)$ such that $\lambda x_0 + (1-\lambda)x_1 = x$, we have $f(x) < \lambda f(x_0) + (1-\lambda)f(x_1)$.

(e) Prove the following stronger version of Jensen's inequality: for any convex f with $\mathbb{E}[X] \in \operatorname{dom} f$, we have $\mathbb{E}[f(X)] \ge f(\mathbb{E}[X])$. If f is strictly convex at $\mathbb{E}[X]$, then $\mathbb{E}[f(X)] = f(\mathbb{E}[X])$ if and only if $X = \mathbb{E}[X]$ with probability 1.

Question 2.2: Let P and Q be distributions on a common measurable space $\mathcal{X}$, and let μ be a measure such that $P, Q \ll \mu$ (for example, $\mu = P + Q$ suffices). Let $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ be the densities of P and Q, respectively. The KL-divergence between P and Q is

$$D_{\text{kl}}(P \| Q) := \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x).$$

Show that $D_{\text{kl}}(P \| Q) \ge 0$, and $D_{\text{kl}}(P \| Q) = 0$ if and only if $P = Q$. [*Hint:* Jensen's inequality. You may use that if f is convex, then $f''(t) > 0$ for almost all t implies that f is strictly convex.]

Question 2.3 (Nice properties of exponential families)**:** Let p_θ be an exponential family density (with respect to some base measure μ on $\mathcal{X}$) the form

$$p_\theta(x) = \exp\left(\langle \theta, T(x) \rangle - A(\theta)\right)$$

where $A(\theta) = \log \int \exp(\langle \theta, T(x) \rangle) d\mu(x)$ and $T : \mathcal{X} \to \mathbb{R}^d$. You may assume that A is infinitely differentiable on $\Theta = \operatorname{dom} A := \{\theta \in \mathbb{R}^d : A(\theta) < \infty\}$, which is open and convex, and you may interchange integration and expectation without comment. (This is true generally for exponential family models.)

(a) Prove that $\theta \mapsto A(\theta)$ is a convex function. [*Hint:* Hölder's inequality.]

(b) Show that if $\nabla^2 A(\theta) \succ 0$, that is, $\nabla^2 A(\theta)$ is strictly positive definite for all θ, then the parameter θ is identifiable. [*Hint:* Use the KL-divergence.]

Question 2.4 (Fun with projections)**:** In this problem, you will (as I threatened in class) prove the existence of projections in Hilbert spaces. We will use real Hilbert spaces. A real Hilbert space is a vector space $\mathcal{V}$ with an inner product $\langle v, w \rangle$ that is linear in its first and second arguments, and we define the norm $\|v\|^2 = \langle v, v \rangle$, and $\mathcal{V}$ is complete, meaning that Cauchy sequences in $\mathcal{V}$ converge.

Let $C \subset \mathcal{V}$ be a closed *convex* set that does not contain 0. Define $M = \inf_{x \in C} \|x\|$. We will show that this infimum is uniquely attained at a point x_C satisfying

$$\langle x_C, y - x_C \rangle \geq 0 \quad \text{for all } y \in C.$$

(a) Prove the parallelogram identity, that is, that $\frac{1}{2}\|x - y\|^2 + \frac{1}{2}\|x + y\|^2 = \|x\|^2 + \|y\|^2$.

(b) Let $x_n \subset C$ be a sequence with $\|x_n\|^2 \to \inf_{x \in C} \|x\|^2$. Show that x_n is a Cauchy sequence.

(c) Argue (in one line) that the limit x_C of the sequence x_n from part (b) belongs to C.

(d) Show that x_C satisfies $\langle x_C, y - x_C \rangle \geq 0$ for all $y \in C$.

(e) Show that x_C minimizes $\|x\|^2$ over C if and only if $\langle x_C, y - x_C \rangle \geq 0$ for all $y \in C$.

(f) Now, consider a general point $x \neq 0$. Using the results of the previous parts, argue that there exists a unique point $\pi_C(x) := \operatorname{argmin}_{y \in C}\{\|x - y\|^2\}$, the projection of x onto C, which is characterized by

$$\langle \pi_C(x) - x, y - \pi_C(x) \rangle \geq 0 \quad \text{for all } y \in C.$$

Draw a picture of your result.

Question 2.5: Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and $\lambda \geq 0$. Recall that for symmetric matrices A, B, the notation $A \succeq B$ means that $A - B$ is positive semidefinite. Throughout this question, we assume $\lambda \geq 0$ and $c > 0$.

(a) Show that if

$$f(y) \geq f(x) + \frac{\lambda}{2}\|y - x\|_2^2$$

for all y such that $\|y - x\|_2 \leq c$, then

$$f(y) \geq f(x) + \frac{\lambda}{2}\min\left\{c, \|x - y\|_2\right\}\|x - y\|_2 \quad \text{for all } y.$$

(b) Assume f is twice continuously differentiable on the set $\{y : \|y - x\|_2 \leq c\}$ and that $\nabla^2 f(y) \succeq \lambda I$ for y such that $\|y - x\|_2 \leq c$. If $\nabla f(x) = 0$, show that

$$f(y) \geq f(x) + \frac{\lambda}{2}\min\left\{c, \|x - y\|_2\right\}\|x - y\|_2 \quad \text{for all } y.$$

(c) Assume that f is twice continuously differentiable on the set $\{y : \|y - x\|_2 \leq c\}$ and that $\nabla^2 f(y) \succeq \lambda I$ for y such that $\|y - x\|_2 \leq c$. Show that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\lambda}{2}\min\left\{c, \|x - y\|_2\right\}\|x - y\|_2 \quad \text{for all } y.$$

(d) What may we conclude about x in parts (a) and (b)? (Consider the cases $\lambda = 0$ and $\lambda > 0$ separately.)

Question 2.6: Let $\mathcal{X}$ be a measurable space and $X_i \stackrel{\text{iid}}{\sim} P$, where P is a probability distribution on $\mathcal{X}$. Let $\Theta \subset \mathbb{R}^d$ be an open set and let $\ell : \Theta \times \mathcal{X} \to \mathbb{R}_+$ be a loss function that is convex in its first argument, that is, $\theta \mapsto \ell(\theta, x)$ is convex. Define the risk functional $R(\theta) := \mathbb{E}_P[\ell(\theta, X)]$, which is the expected loss of a vector θ. Let $\theta^\star = \operatorname{argmin}_{\theta \in \Theta} R(\theta)$ and assume that the Hessian $\nabla^2 R(\theta^\star) \succ 0$, that is, the Hessian of the risk is positive definite at the point $\theta^\star$, and assume that $\theta^\star \in \operatorname{int} \Theta$. Make the following assumption:

(i) There is a function $H : \mathcal{X} \to \mathbb{R}_+$ such that $\mathbb{E}[H^2(X)] < \infty$ and the Hessian $\nabla^2 \ell(\theta, x)$ is $H(x)$ Lipschitz in θ, that is,

$$\left\|\nabla^2 \ell(\theta, x) - \nabla^2 \ell(\theta', x)\right\|_{\text{op}} \le H(x) \left\|\theta - \theta'\right\| \quad \text{for all } \theta, \theta' \in \Theta.$$

We will show that under these conditions, if we define the empirical risk

$$\widehat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\theta, X_i)$$

where $X_i \stackrel{\text{iid}}{\sim} P$, and $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \widehat{R}_n(\theta)$, then we have asymptotic normality of $\widehat{\theta}_n$. You may assume that gradients and Hessians can be passed through all expectations and integrals and as many moments of $\nabla \ell$ as you need.

(a) Argue that $R(\theta)$ and $\widehat{R}_n$ are convex in θ.

(b) Using the above assumptions, show that $\widehat{\theta}_n \stackrel{p}{\to} \theta^\star$. You may use the following result (see Question 2.5): if a function f is convex and satisfies $\nabla^2 f(\theta) \succeq \lambda I$ for all θ satisfying $\|\theta - \theta_0\| \le c$, then

$$f(\theta) \ge f(\theta_0) + \nabla f(\theta_0)^T (\theta - \theta_0) + \frac{\lambda}{2} \min\left\{\|\theta_0 - \theta\|^2, c\|\theta_0 - \theta\|\right\}.$$

(c) Assuming that $\widehat{\theta}_n \stackrel{p}{\to} \theta^\star$, use a Taylor expansion to show that

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \stackrel{d}{\to} \mathsf{N}\left(0, (\nabla^2 R(\theta^\star))^{-1} \Sigma (\nabla^2 R(\theta^\star))^{-1}\right)$$

where $\Sigma = \operatorname{Cov}(\nabla \ell(\theta^\star, X))$ is the covariance matrix of the gradient of the loss.

Question 2.7 (Log-concavity, Boyd & Vandenberghe Ex. 3.54)**:** Let $F : \mathbb{R} \to \mathbb{R}_+$ be a twice continuously differentiable function with $F(t) > 0$ for all $t \in (a, b)$. We say that F is log-concave (on (a, b)) if $t \mapsto \log F(t)$ is a concave function (on the interval (a, b)). This is equivalent to $\frac{d^2}{dt^2} \log F(t) \le 0$ for all $t \in (a, b)$.

(a) Show that F is log-concave on (a, b) if and only if $F(t)F''(t) \le F'(t)^2$ for all $t \in (a, b)$.

Define $F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t \exp(-\frac{1}{2}u^2) du$, the Gaussian CDF. We will verify that F is log-concave. You may interchange the order of differentiation and integration without comment in your arguments if needed. (It may not be needed.)

(b) Show that $F''(t)F(t) \le F'(t)^2$ for all $t \ge 0$.

(c) Show that for any pair $t, u \in \mathbb{R}$, we have $tu \leq \frac{1}{2}t^2 + \frac{1}{2}u^2$.

(d) Show that $\exp(-\frac{u^2}{2}) \leq \exp(\frac{t^2}{2} - tu)$, and conclude that

$$\int_{-\infty}^{t} e^{-\frac{1}{2}u^2} du \leq e^{\frac{1}{2}t^2} \int_{-\infty}^{t} e^{-ut} du.$$

(e) Verify that $F''(t)F(t) \leq F'(t)^2$ for $t < 0$.

Question 2.8 (Convexity of minimizers of convex functions)**:** A function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is jointly convex in its arguments if for $\lambda \in [0, 1]$,

$$f(\lambda x_0 + (1-\lambda)x_1, \lambda y_0 + (1-\lambda)y_1) \leq \lambda f(x_0, y_0)(1-\lambda)f(x_1, y_1)$$

for all x_0, x_1, y_0, y_1 (where if one of the arguments is not in dom f, then $f = +\infty$ and $+\infty \leq +\infty$).

(a) Show that if f is convex, then

$$g(x) := \begin{cases} 0 & \text{if } f(x) \leq 0 \\ +\infty & \text{otherwise} \end{cases}$$

is convex.

(b) Show that if $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is (jointly) convex, then for any convex set $\mathcal{Y} \subset \mathbb{R}^m$ the function $g(x) := \inf_{y \in \mathcal{Y}} f(x, y)$ is convex.

(c) Show that for any convex f_0, f_1, the value functional

$$v(x) := \inf_y \{f_0(x, y) \text{ s.t. } f_1(x, y) \leq 0\}$$

is convex in x.

Question 2.9 (Subgradients)**:** Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. The *subgradient set* of f at x is

$$\partial f(x) := \{g \in \mathbb{R}^d \mid f(y) \geq f(x) + \langle g, y - x \rangle \text{ for all } y\}.$$

If f is defined on all of $\mathbb{R}^d$, this set is always non-empty; otherwise it is non-empty on the relative interior of the domain of f. When f is differentiable, $\partial f(x) = \{\nabla f(x)\}$, so that $\partial f(x)$ is simply the gradient.

(a) Draw a picture of the subgradient of a convex function.

(b) Show that f is minimized at the point $x^\star$ if and only if $0 \in \partial f(x^\star)$.

(c) Let $f(x) = \|x\|_2$. Show that

$$\partial f(x) = \begin{cases} x / \|x\|_2 & \text{if } x \neq 0 \\ \{u \in \mathbb{R}^d \mid \|u\|_2 \leq 1\} & \text{if } x = 0. \end{cases}$$

(d) Let $f(x) = h(Ax)$ for some $A \in \mathbb{R}^{n \times d}$. Show that $\partial f(x) = A^T \, \partial h(v)|_{v=Ax}$.

(e) Let $f(x) = \|x\|_1$. Show that $\partial f(x)$ consists of vectors $v \in [-1, 1]^d$ satisfying

$$v_j \in \begin{cases} \{1\} & \text{if } x_j > 0 \\ [-1, 1] & \text{if } x_j = 0 \\ \{-1\} & \text{if } x_j < 0. \end{cases}$$

Question 2.10 (Growth of convex functions)**:** Let $h : \mathbb{R}^d \to \mathbb{R}$ be convex.

(a) Let $\theta_0, \theta_1 \in \mathbb{R}^d$, and for $t \in \mathbb{R}_+$ define $\theta_t = \theta_0 + t(\theta_1 - \theta_0)$. Show that $h(\theta_t) - h(\theta_0) \geq t[h(\theta_1) - h(\theta_0)]$ for all $t \geq 1$. *Hint:* For $t \geq 1$, you have $\theta_1 = \frac{1}{t}\theta_t + (1 - \frac{1}{t})\theta_0$.

(b) Suppose for some $r > 0$ that $h(\theta) > h(\theta_0)$ for all θ satisfying $\|\theta - \theta_0\|_2 = r$, that is, in a sphere around $h(\theta_0)$. Show that $h(\theta) > h(\theta_0)$ for all θ such that $\|\theta - \theta_0\|_2 > r$.

Question 2.11 (Growth of an absolute loss)**:** Consider the loss function $\ell(t) = |t|$. We will show that under various types of (random) smoothing, it still exhibits reasonable growth properties. In particular, we will consider functions of the form

$$\varphi(t) := \mathbb{E}[\ell(t + Z) - \ell(Z)],$$

where Z is a *symmetric* random variable with various distributions.

(a) Argue that $\varphi(t)$ is 1-Lipschitz and convex in t and that $\varphi(t) \in [0, |t|]$ for all $t \in \mathbb{R}$.

(b) Let Z have density π, where $\pi(z) \geq p_{\min}$ for all $z \in [-\tau, \tau]$, where $p_{\min}, \tau > 0$ are both positive. Define the Huber loss

$$h_\tau(t) := \begin{cases} \frac{1}{2}t^2 & \text{for } |t| \leq \tau \\ \tau|t| - \frac{1}{2}\tau^2 & \text{for } |t| > \tau. \end{cases}$$

Show that

$$\varphi(t) \geq 2p_{\min}h_\tau(t).$$

Hint. It may be useful to consider derivatives via Lemma 7.14.1.

(c) Assume Z has a point mass at $Z = 0$, that is, $\mathbb{P}(Z = 0) = p_0 > 0$. Show that $\varphi(t) \geq p_0|t|$.

3 Asymptotic Efficiency

Question 3.1: Consider estimating the cumulative distribution function $\mathbb{P}(X \leq x)$ at a fixed point x based on a sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$, the distribution of X. A standard non-parametric estimator is $T_n := n^{-1} \sum_{i=1}^n \mathbf{1}\{X_i \leq x\}$.

(a) What is the asymptotic distribution of T_n?

(b) Suppose we know that $X_i \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$ for some unknown θ. Letting $\Phi(x) = P(Z \leq x)$ be the standard normal CDF, another possible estimator is $G_n := \Phi(x - \overline{X}_n)$ where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. What is the asymptotic distribution of G_n?

(c) What is the asymptotic relative efficiency of G_n with respect to T_n?

(d) Suppose that the data are non-normal. Show that G_n is not consistent.

(e) Again, assume that the data are normal $\mathsf{N}(\theta, 1)$. Give a consistent estimator $\widehat{\theta}_n$ of θ based on T_n. What is the asymptotic distribution of your estimator? What is its efficiency relative to the mean $\overline{X}_n$?

Question 3.2 (One-step estimators)**:** Let $\{P_\theta\}_{\theta \in \Theta}$ be a family of models where $\Theta \subset \mathbb{R}^d$ is open and let $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$, where P_θ has density p_θ w.r.t. the measure μ as usual. Assume that $\ell_\theta(x) = \log p_\theta(x)$ is twice continuously differentiable in θ and $\nabla^2 \ell_\theta(x)$ is $M(x)$-Lipschitz, where $\mathbb{E}_\theta[M^2(X)] < \infty$ for all $\theta \in \Theta$. You may assume that the order of differentiation and expectation can be exchanged.

Suppose that $\widehat{\theta}_n$ is a $\sqrt{n}$-consistent estimator, that is,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = O_{P_{\theta_0}}(1).$$

Let $L_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$, where $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$. Consider the *one-step estimator* δ_n that solves the first-order approximation to $\nabla L_\theta(\theta) = 0$ given by

$$\nabla L_n(\widehat{\theta}_n) + \nabla^2 L_n(\widehat{\theta}_n)(\delta_n - \widehat{\theta}_n) = 0.$$

(a) What is the asymptotic distribution of δ_n?

Suppose that the family $\{P_\theta\}_{\theta \in \mathbb{R}}$ is the Cauchy family, with densities

$$p_\theta(x) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}.$$

Let $X_i \overset{\text{iid}}{\sim} P_\theta$ and define $\widehat{\theta}_n = \mathsf{Median}(X_1, \ldots, X_n)$.

(b) Show that $\sqrt{n}(\widehat{\theta}_n - \theta) = O_{P_\theta}(1)$.

(c) Let δ_n be the one-step estimator for this family. What is its asymptotic distribution?

Question 3.3 (Super-efficiency)**:** In class, we saw the Hodges estimator of the normal mean, which based on a sample $\{X_i\}_{i=1}^n$ is

$$T_n := \begin{cases} \overline{X}_n & \text{if } |\overline{X}_n| \geq n^{-1/4} \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

In this question, you will simulate the Hodges estimator (3.1) to study its performance. Repeat the following experiment $N = 500$ times. For $n \in \{50, 100, 200, 400\}$, generate i.i.d. samples $X_i \stackrel{\text{iid}}{\sim} \mathsf{N}(\theta_n, 1)$, $i = 1, \ldots, n$, where you set θ_n to be the "local" perturbation from $\theta = 0$ given by

$$\theta_n = 0 + \frac{h}{\sqrt{n}}, \quad h \in \{-5, -4.9, -4.8, \ldots, 4.9, 5.0\} = \{-k/10 \mid k \in \{-50, \ldots, 50\}\}.$$

(Thus, you will generate a total of $N \times 4 \times 101$ different samples.) For each sample you generate, compute T_n and $\widehat{\theta}_n = \overline{X}_n$, the sample mean.

(a) Generate three plots, one each for $n = 50, 100, 200$, and plot the (sampled/simulated) mean squared error $\mathbb{E}_h[(\widehat{\theta}_n - \theta_n)^2]$ and $\mathbb{E}_h[(T_n - \theta_n)^2]$ over your simulations as h varies. What do you see? Include your plots in your homework submission.

(b) Using the same errors as before, plot the rescaled mean squared error $n \cdot \mathbb{E}_h[(\widehat{\theta}_n - \theta_n)^2]$ and $n \cdot \mathbb{E}_h[(T_n - \theta_n)^2]$ as h varies *on the same plot*. What do you see? Which estimator do you prefer? Include your plots in your homework submission.

Question 3.4 (Corrupted observations, or the data processing inequality)**:** Let $\{P_\theta\}_{\theta \in \Theta}$, where $\Theta \subset \mathbb{R}^d$ is open and convex (or whatever nice properties you want of it) be a family of models, and assume that we have Fisher information $I_\theta = \mathbb{E}_\theta[\nabla \ell_\theta \nabla \ell_\theta^T] = -\mathbb{E}_\theta[\nabla^2 \ell_\theta]$. Suppose that instead of observing a sample $X_i \stackrel{\text{iid}}{\sim} P_\theta$, there is a *channel* $Q(\cdot \mid x)$, which given $X = x$ draws $Y \mid X = x$ and outputs Y according to the distribution $Q(\cdot \mid x)$. Let

$$I_\theta^{(Q)}$$

be the Fisher information associated with the observation of Y according to this corrupted observation. That is, the process is that $X \sim P_\theta$, and then $Y \sim Q(\cdot \mid X)$, and we observe Y. You may assume for simplicity that Q has a density for all x or has a p.m.f. for all x (that is, Y is discrete with common support for all x) and ignore other measurability issues. Assume that $\{P_\theta\}$ have densities $p_\theta = \frac{dP_\theta}{d\mu}$ w.r.t. a measure μ.

(a) Show that $I_\theta^{(Q)} \preceq I_\theta$ in the positive semidefinite order, meaning that $v^T I_\theta^{(Q)} v \le v^T I_\theta v$ for all vectors v.

(b) Consider randomized response, in which we wish to estimate the parameter $\theta \in [0, 1]$ of a Bernoulli random variable $X_i \stackrel{\text{iid}}{\sim} \mathsf{Bernoulli}(\theta)$, but instead of observing X_i we observe Y_i with corrupted conditional distribution

$$Q(Y_i = x \mid X = x) = \frac{1 + \epsilon}{2}, \quad Q(Y_i = 1 - x \mid X = x) = \frac{1 - \epsilon}{2}$$

where $\epsilon \in (0, 1)$. What are I_θ and $I_\theta^{(Q)}$ in this case?

(c) Based on a sample $Y_1, \ldots, Y_n$ in the setting of part (b), give a consistent estimator of θ based on $Y_1, \ldots, Y_n$. Is your estimator efficient?

(d) Give a situation in which such a procedure might be useful.

Question 3.5 (An average treatment effect estimator)**:** In the Neyman-Rubin (potential outcomes) approach to causal estimation, one treats estimation as a missing data problem. Let $A \in \{0, 1\}$ be an action (often called the treatment or intervention). The *potential outcomes* are the pair $(Y(0), Y(1)) \in \mathbb{R}$, where $Y(0)$ is the response when action $A = 0$ is chosen and $Y(1)$ the response when $A = 1$ is chosen. Thus, for any individual, we observe a *single* response: under action $A = a$, we observe $Y(a)$ but never $Y(1 - a)$. The *average treatment effect* is the difference

$$\tau := \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)],$$

where the expectation is taken over the population of individuals we might intervene on. Here, $A = 1$ is the treatment, while $A = 0$ indicates the control (untreated) action, and we may use the notation $Y\mathbf{1}\{A = a\} = Y(a)\mathbf{1}\{A = a\}$.

The "gold standard" approach is a randomized experiment, where for individuals $i = 1, 2, \ldots, n$, one chooses $A_i \in \{0, 1\}$ uniformly and observes $Y_i(A_i) \in \mathbb{R}$. We assume that individuals are i.i.d.

(a) Show that for $a \in \{0, 1\}$, we have $\mathbb{E}[Y_i(a)\mathbf{1}\{A_i = a\}] = \frac{1}{2}\mathbb{E}[Y(a)]$ in the randomized experiment setting, and hence that $\tau = 2(\mathbb{E}[Y(1)\mathbf{1}\{A = 1\}] - \mathbb{E}[Y(0)\mathbf{1}\{A = 0\}])$.

We consider two mean-based estimators. For $a \in \{0, 1\}$, define the sets $S_a = \{i \in [n] \mid A_i = a\}$ (i.e. the treatment and control groups). The basic estimator is

$$\widehat{\tau}_n := \frac{1}{n} \sum_{i \in S_1} 2Y_i - \frac{1}{n} \sum_{i \in S_0} 2Y_i$$

(b) Give the asymptotic distribution of $\widehat{\tau}_n$. (That is, give the limit distribution of $\sqrt{n}(\widehat{\tau}_n - \tau)$.)

We also consider the slightly more nuanced mean-based estimator, which normalizes by the sample sizes,

$$\widehat{\tau}_n^{\text{norm}} := \frac{1}{|S_1|} \sum_{i \in S_1} Y_i - \frac{1}{|S_0|} \sum_{i \in S_0} Y_i.$$

(c) For $a \in \{0, 1\}$, give the asymptotic distribution of

$$\sqrt{n}\left(\frac{n}{2|S_a|} - 1\right).$$

(d) Give the asymptotic distribution of the mean-based estimator $\widehat{\tau}_n^{\text{norm}}$. *Hint:* it may be useful to split the quantities by considering the means $\tau_a = \mathbb{E}[Y(a)]$ for $a \in \{0, 1\}$ separately.

(e) In the preceding parts, you have shown that

$$\sqrt{n}(\widehat{\tau}_n - \tau) \xrightarrow{d} \mathsf{N}(0, \sigma^2), \quad \sqrt{n}(\widehat{\tau}_n^{\text{norm}} - \tau) \xrightarrow{d} \mathsf{N}(0, \sigma_{\text{norm}}^2).$$

Show that if the means $\tau_a = \mathbb{E}[Y(a)]$ satisfy $\tau_0 \neq -\tau_1$, then $\sigma^2 > \sigma_{\text{norm}}^2$.

Question 3.6 (A weighted average treatment effect estimator)**:** We consider the same setting as in problem 3.5, but take an alternative approach, where we may differentially sample individuals based on their covariates X. To that end, consider a *propensity score* (the propensity for being treated)

$$e(x) := \mathbb{P}(A = 1 \mid X = x). \tag{3.2}$$

Now, we assume that given an individual with covariates $X = x$, we assign treatment A consistionally according to the propensity score (3.2), that is, $\mathbb{P}(A = a \mid X = x) = e(x)$, so that $(Y(0), Y(1)) \perp A \mid X$, that is, the potential responses $(Y(0), Y(1))$ are independent of A given X.

(a) Show that the average treatment $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ also equals

$$\tau = \mathbb{E}\left[\frac{Y(A)\mathbf{1}\left\{A = 1\right\}}{e(X)}\right] - \mathbb{E}\left[\frac{Y(A)\mathbf{1}\left\{A = 0\right\}}{1 - e(X)}\right].$$

(b) Define the conditional second moments $v_2(x, a) := \sqrt{\mathbb{E}[Y(a)^2 \mid X = x]}$, and consider the propensity weighted estimator

$$\widehat{\tau}_n^{\text{ps}} := \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i\mathbf{1}\left\{A_i = 1\right\}}{e(X_i)} - \frac{Y_i\mathbf{1}\left\{A_i = 0\right\}}{1 - e(X_i)}\right].$$

Compute the asymptotic variance σ_{ps}^2 in

$$\sqrt{n}\left(\widehat{\tau}_n^{\text{ps}} - \tau\right) \xrightarrow{d} \mathsf{N}\left(0, \sigma_{\text{ps}}^2\right)$$

as a function (with appropriate expectations) of $v_2(x, a)$ and $e(x)$.

(c) What choice of propensity score $e(x)$ minimizes the asymptotic variance σ_{ps}^2? Give a one-sentence (heuristic) intuition for this choice. When does this improve over the "gold standard" approach of the pure randomized experiment in part (b) in Q. 3.5?

Question 3.7 (A constrained risk inequality (Brown and Low [4]))**:** In this question, we develop some results that help to show the penalties in estimation rates for super-efficient estimators. We begin with the most abstract setting, specializing it presently. Let $\Theta \subset \mathbb{R}^d$ and $\mathcal{P}$ be a collection of probability distributions. We are given a loss function $L : \Theta \times \mathcal{P} \to \mathbb{R}_+$ with the property that

$$\inf_{\theta \in \Theta} L(\theta, P) = 0$$

for all $P \in \mathcal{P}$, that is, the minimal value of L is 0. We will develop various lower bounds on the expected loss of estimators $\widehat{\theta}$, that is, for distributions $\mathcal{P}$ on a space $\mathcal{X}$, on the quantity

$$\mathbb{E}_P[L(\widehat{\theta}, P)] = \mathbb{E}_P[L(\widehat{\theta}(X), P)]$$

for $\widehat{\theta} : \mathcal{X} \to \Theta$, where the expectation is taken over $X \sim P$. Given such a loss, the *separation* between distributions P_0, P_1 the loss induces is

$$d_L(P_0, P_1) := \inf_{\theta \in \Theta}\left\{L(\theta, P_0) + L(\theta, P_1)\right\},$$

that is, the minimal value a parameter θ can achieve simultaneously on P_0 and P_1.

(a) Consider estimating a parameter $\theta : \mathcal{P} \to \mathbb{R}$ of a distribution and using the squared error $L_{\text{sq}}(\theta, P) = \frac{1}{2}(\theta - \theta(P))^2$. Show that for $L = L_{\text{sq}}$ and $\Theta = \mathbb{R}$, we have

$$d_L(P_0, P_1) = \frac{1}{4}(\theta(P_0) - \theta(P_1))^2.$$

Now we argue that if one achieves small loss on a distribution P_0, one must achieve a loss on P_1 that scales as the separation $d_L(P_0, P_1)$. Define the χ^2-affinity between distributions P, Q by

$$\rho\left(P \| Q\right) := \mathbb{E}_P\left[\frac{dP}{dQ}\right] = \int \frac{dP^2}{dQ} = \int \frac{p(x)^2}{q(x)} d\mu(x),$$

where the last equality holds whenever P, Q have densities p, q w.r.t. a base measure μ. We say $\rho\left(P \| Q\right) = +\infty$ whenever $P \not\ll Q$.

(b) Using the Cauchy-Schwarz inequality and that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for all $a, b \ge 0$, show that for any distributions P_0, P_1, any estimator $\widehat{\theta}$, and any loss function $L : \Theta \times \mathcal{P} \to \mathbb{R}_+$,

$$\sqrt{d_L(P_0, P_1)} \le \sqrt{\mathbb{E}_1[L(\widehat{\theta}, P_1)]} + \sqrt{\rho\left(P_1 \| P_0\right)} \sqrt{\mathbb{E}_0[L(\widehat{\theta}, P_0)]},$$

where $\mathbb{E}_i$ denotes expectation under P_i.

(c) Conclude for any pair of distributions P_0, P_1 that we have the *constrained risk inequality*

$$\mathbb{E}_1\left[L(\widehat{\theta}, P_1)\right] \ge \left(\sqrt{d_L(P_0, P_1)} - \sqrt{\rho\left(P_1 \| P_0\right) \mathbb{E}_0[L(\widehat{\theta}, P_0)]}\right)_+^2. \tag{3.3}$$

We now develop an application of the constrained risk inequality to super-efficient estimation of a normal mean. Suppose that $\widehat{\theta}_n : \mathbb{R}^n \to \mathbb{R}$ is an estimator of a Gaussian mean such that

$$\mathbb{E}_0[(\widehat{\theta}_n - 0)^2] = \mathbb{E}_0[(\widehat{\theta}_n - \mathbb{E}_0[X])^2] \le \frac{\delta_n}{n}$$

under i.i.d. sampling from $P_0 = \mathsf{N}(0, 1)$, where $\delta_n \ge 0$ is a sequence with $\delta_n \to 0$.

(d) Show that for two Gaussian distributions $P_0 = \mathsf{N}(\theta_0, 1)$ and $P_1 = \mathsf{N}(\theta_1, 1)$ we have

$$\rho\left(P_1^n \| P_0^n\right) = \exp\left(n(\theta_1 - \theta_0)^2\right).$$

(e) Give a sequence of means θ_n such that under i.i.d. sampling $X_i \overset{\text{iid}}{\sim} \mathsf{N}(\theta_n, 1)$, $i = 1, \ldots, n$, we have

$$n \cdot \mathbb{E}_{\theta_n}[(\widehat{\theta}_n - \theta_n)^2] \to \infty \quad \text{as} \quad n \to \infty.$$

That is, $\widehat{\theta}_n$ does not uniformly enjoy the $1/n$ rate of convergence (for squared error) that we might expect, e.g., from the sample mean. *Hint:* in our solution, we get a lower bound that scales as $\log \frac{1}{\delta_n}$.

Question 3.8 (An application of the constrained risk inequality)**:** Let $\mathcal{P}$ denote the location family of Laplace distributions, that is, probabilities with densities

$$p_\theta(x) = \frac{1}{2} \exp(-|x - \theta|)$$

on $\mathbb{R}$. In this question, you will show that a super-efficient estimator of the location at a single point θ must be inaccurate at a large collection of alternative locations θ' with probability tending to 1. We use the notation of Question 3.7.

(a) Show that for $\theta \ge 0$, the χ^2-affinity between Laplace distributions is

$$\rho\left(P_\theta \| P_0\right) = \frac{1}{3}(2e^\theta + e^{-2\theta}) \overset{(\star)}{=} 1 + \theta^2 + O(\theta^3),$$

where equality $(\star)$ holds as $\theta \downarrow 0$. (By appropriate shifts, one therefore immediately obtains $\rho\left(P_{\theta_1} \| P_{\theta_0}\right) = \frac{1}{3}(2e^{|\theta_1 - \theta_0|} + e^{-2|\theta_1 - \theta_0|})$.)

(b) Show that if $\widehat{\theta}_n$ is rate super-efficient for estimating the location at P_0, meaning that there exists some sequence $\delta_n \geq 0$ with $\delta_n \to 0$ such that

$$P_0^n \left(|\widehat{\theta}_n| \geq \frac{1}{\sqrt{n}} \right) \leq \delta_n,$$

then for any $2 \leq C < \infty$, we have

$$\liminf_{n \to \infty} \inf_{\frac{2}{\sqrt{n}} < \theta \leq \frac{C}{\sqrt{n}}} P_\theta^n \left(|\widehat{\theta}_n - \theta| \geq \frac{1}{\sqrt{n}} \right) = 1.$$

That is, the asymptotic probability of being within $1/\sqrt{n}$ of the true location is zero for a large collection of locations θ. *Hint:* Consider the (sequence) of loss functions $L_n : \mathbb{R} \times \mathcal{P} \to \{0, 1\}$, indexed by n, defined by by

$$L_n(t, P_\theta) := \mathbf{1} \left\{ \sqrt{n} |t - \theta| \geq 1 \right\}.$$

Apply the technique in Question 3.7.

4 U- and V-statistics

Question 4.1 (Signed rank statistics, cf. Van der Vaart Ex. 12.4 and q. 12.9)**:** Let $h(x_1, x_2) = \mathbf{1}\{x_1 + x_2 > 0\}$ and define the U-statistic

$$U_n = \binom{n}{2}^{-1} \sum_{|\beta|=2, \beta \subset [n]} h(X_\beta),$$

which is useful for testing symmetry (and continuity) of the distribution of the random variable X with CDF $F(x) = \mathbb{P}(X \le x)$, that is, that $F(x) = 1 - F(-x)$.

(a) Show that if X has a density that is symmetric about 0, then $\theta = \mathbb{E}[U_n]$ satisfies $\theta = \frac{1}{2}$, and

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathsf{N}(0, 1/3)$$

independently of the distribution of X as long as it is symmetric and X has a density.

The Wilcoxon *signed rank test* is defined as follows. Let $R_1^+, \ldots, R_n^+$ denote the ranks of the absolute values $|X_1|, \ldots, |X_n|$ where $R_i^+ = k$ means that $|X_i|$ is the kth smallest of the absolute values in the sample, $R_i^+ = \sum_{j=1}^n \mathbf{1}\{|X_j| \le |X_i|\}$. Then we define $W^+ := \sum_{i=1}^n R_i^+ \mathbf{1}\{X_i > 0\}$.

(b) Show that if no observations are tied, then

$$W^+ = \binom{n}{2} U_n + \sum_{i=1}^n \mathbf{1}\{X_i > 0\}.$$

Question 4.2 (U-statistics, the information, ranking models, and probit regression)**:** Suppose we have a standard linear regression problem with $Y_i = x_i^T \theta + \varepsilon_i$, $\varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$, and $x_i \in \mathbb{R}^d$ are drawn i.i.d. from a distribution with $\mathbb{E}[x_i] = 0$ and $\mathrm{Cov}(x_i) = \Sigma$. Assume that for all $n \ge d$ we have $\frac{1}{n}\sum_{i=1}^n x_i x_i^T$ is invertible (this will occur if the x_i have a density). Let $Y_1, \ldots, Y_n$ be a sample according to this process, $n \ge d$.

(a) Let $\widehat{\theta}_n = \mathrm{argmin}_\theta \frac{1}{2n}\sum_{i=1}^n (x_i^T \theta - Y_i)^2$ be the least-squares minimizer. What is the asymptotic distribution of $\widehat{\theta}_n$?

One model of ranking relative values of items posits that while humans are very bad at assigning numerical scores, we are quite good at performing relative evaluations (i.e. is something more or less than something else). As a consequence, suppose that you do not actually trust the true values of the Y_i, but you do trust their relative values, so you wish to base your estimate of θ on the ordering $Y_i \lesseqgtr Y_j$. Consider the U-statistic-based "log-likelihood"

$$L_n(\theta) := \binom{n}{2}^{-1} \sum_{i,j \le n} \mathbf{1}\{Y_i > Y_j\} \log P_\theta(Y_i > Y_j \mid x_i, x_j).$$

(b) Show that $L_n(\theta)$ is concave in θ. *Hint:* Write it in terms of the Gaussian CDF. You may use the results of Question 2.7.

(c) Let $\widehat{\theta}_n = \mathrm{argmax}_\theta L_n(\theta)$. You may assume that $\widehat{\theta}_n$ is consistent for θ_0 under the true distribution θ_0. What is the asymptotic distribution of $\widehat{\theta}_n$?

(d) Which estimator of parts (a) and (c) do you prefer?

Question 4.3 (Mean-differences in Hilbert spaces)**:** Recall Question 1.10, which defined reproducing Kernel Hilbert spaces (RKHSs). Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with associated RKHS $\mathcal{H}$. Assume that $\mathcal{X}$ is compact. We call k *universal* if it is dense in $\mathcal{C}(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$. That is, for any $\epsilon > 0$ and any continuous function $f : \mathcal{X} \to \mathbb{R}$, there exists a function $h \in \mathcal{H}$ such that $\sup_{x \in \mathcal{X}} |f(x) - h(x)| < \epsilon$.

Define $\varphi(x) = k(\cdot, x)$. (Thus $k(x, z) = \langle \varphi(x), \varphi(z) \rangle$, and $\varphi(x)$ is the representer of evaluation at x, i.e., $\langle h, \varphi(x) \rangle = h(x)$ for all $h \in \mathcal{H}$.) Let $\mathcal{P}$ be the collection of distributions on $\mathcal{X}$ for which $\mathbb{E}_P[\sqrt{k(X, X)}] < \infty$.

(a) Using the Riesz representation theorem for Hilbert spaces, argue that the mean mapping $\mu(P) := \mathbb{E}_P[\varphi(X)]$ exists and is a vector in $\mathcal{H}$. *Hint:* Letting $\|\cdot\|$ denote the norm on $\mathcal{H}$, the Riesz representation theorem for Hilbert spaces says that if $L : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional, meaning that $L(f) \le C \cdot \|f\|$ for some constant C, then there exists some $h_L \in \mathcal{H}$ such that $L(f) = \langle h_L, f \rangle$ for all $f \in \mathcal{H}$.

(b) Assume that $\mathcal{X}$ is compact and that k is universal. Show that the mean embedding

$$P \mapsto \mathbb{E}_P[\varphi(X)] = \int_{\mathcal{X}} \varphi(x) dP(x)$$

is one-to-one, that is, if $P \neq Q$ then $\mathbb{E}_P[\varphi(X)] \neq \mathbb{E}_Q[\varphi(X)]$.

(c) For distributions P and Q, show that

$$\sup_{f \in \mathcal{H}, \|f\| \le 1} \{\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]\} = \sqrt{\mathbb{E}[k(X, X')] + \mathbb{E}[k(Z, Z')] - 2\mathbb{E}[k(X, Z)]},$$

where $X, X' \stackrel{\text{iid}}{\sim} P$ and $Z, Z' \stackrel{\text{iid}}{\sim} Q$.

Question 4.4 (A kernel two-sample test: basic theory)**:** Consider the classical two-sample testing problem, in which we receive two samples

$$X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P \quad \text{and} \quad Z_1, \ldots, Z_n \stackrel{\text{iid}}{\sim} Q$$

(we assume the samples are the same size for simplicity). We would like to test the null

$$H_0 : P = Q$$

(against the alternative $P \neq Q$). Now, consider the U-like two-sample statistic

$$U_n := \binom{n}{2}^{-1} \sum_{i<j} k(X_i, X_j) + \binom{n}{2}^{-1} \sum_{i<j} k(Z_i, Z_j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(X_i, Z_j),$$

where k is a kernel function with associated reproducing kernel Hilbert space $\mathcal{H}$. (Recall Questions 1.10 and 4.3.) We define the *kernel mean discrepancy* as in Question 4.3 (c) by

$$\Delta(P, Q) := \mathbb{E}[k(X, X')] + \mathbb{E}[k(Z, Z')] - 2\mathbb{E}[k(X, Z)]$$

for $X, X' \stackrel{\text{iid}}{\sim} P$ and $Z, Z' \stackrel{\text{iid}}{\sim} Q$.

(a) Show that U_n is unbiased for $\Delta(P, Q)$.

(b) Argue that under the null H_0 we have

$$U_n = O_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Hint: In the definition of U_n, replace the kernel k with $\widehat{k}(x, x') := \langle \varphi(x) - \mu, \varphi(x') - \mu \rangle$ for an appropriate vector $\mu \in \mathcal{H}$. Does this change U_n? Then bound $\mathbb{E}[U_n^2]$.

(c) Assume that k is a universal kernel (so that $\Delta(P, Q) > 0$ whenever $P \neq Q$). Give a pointwise consistent test T_n of the null $P = Q$ against the alternative $P \neq Q$, that is,

$$\lim_{n\to\infty} \mathbb{P}(T_n \text{ rejects}) = 0$$

if $\mathbb{P}$ is the joint distribution of P and Q when $P = Q$, and otherwise,

$$\lim_{n\to\infty} \mathbb{P}(T_n \text{ rejects}) = 1.$$

Question 4.5 (A kernel two-sample test: performance questions)**:** We consider the performance of a kernel two-sample test with the "favorite" kernel of machine learning, the RBF (Gaussian) kernel, defined on $\mathbb{R}^d \times \mathbb{R}^d$ by

$$k_\tau(x, z) = \exp(-\frac{1}{2\tau^2} \|x - z\|_2^2),$$

which is a universal kernel. Suppose that we have distributions P and Q that are known to be Gaussian on $\mathbb{R}^d$ with identity covariance, where

$$P = \mathsf{N}(0, I) \quad \text{and} \quad Q = \mathsf{N}(\theta, I).$$

We compare the performance of two tests of the null $H_0 : P = Q$, one based on kernel mean discrepancy (Question 4.4) and the other based on a standard normal test. Let

$$X_i \overset{\text{iid}}{\sim} P, \quad Z_i \overset{\text{iid}}{\sim} Q, \ i = 1, \ldots, n$$

and $\mathbb{P}$ denote the joint distribution of (X, Z).

(a) Let T_n be the standard test that $\theta \neq 0$, that is

$$T_n = \begin{cases} \text{reject} & \text{if } \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \right\| \geq t \\ \text{accept} & \text{otherwise.} \end{cases}$$

Give the value t so that T_n is a level α test, that is, under the null $H_0 : P = Q$, so that $\mathbb{P}(T_n \text{ rejects}) = \alpha$.

(b) Another possible test is based on the U-type statistic of problem 4.4,

$$U_n := \binom{n}{2}^{-1} \sum_{i<j} k_\tau(X_i, X_j) + \binom{n}{2}^{-1} \sum_{i<j} k_\tau(Z_i, Z_j) - \frac{2}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_\tau(X_i, Z_j),$$

which is mean zero under the null $H_0 : P = Q$ and $O_{\mathbb{P}}(n^{-1})$ under this null (by Q. 4.4). Define the test

$$\Psi_n = \begin{cases} \text{reject} & \text{if } |U_n| \geq u \\ \text{accept} & \text{otherwise.} \end{cases}$$

For the choice $\tau = 1$ in the kernel k_τ, let the values of the dimension vary over $d = 1, 2, 4, 8, 16, 32, 64, 128$ and sample size vary over $n = 4, 8, 16, 32, 64, 128, 256, 512$ (i.e. 2^k for $k \in \{2, \ldots, 9\}$). Use simulation to estimate the thresholds $u_{n,d}$ such that under the null (in our Gaussian family) $H_0 : P = Q$,

$$\mathbb{P}(\Psi_n \text{ rejects}) = \alpha.$$

Report your thresholds.

(c) Let us now do a power simulation for the tests T_n and Ψ_n. Let $\mathbb{P}_\theta$ be the joint distribution of P and Q when $Q = \mathsf{N}(\theta, I)$. Define the power values

$$\pi_n^T(\theta) := \mathbb{P}_\theta(T_n \text{ rejects}) \quad \text{and} \quad \pi_n^\Psi(\theta) := \mathbb{P}_\theta(\Psi_n \text{ rejects})$$

(leaving the dimension d implicit). For dimensions $d = 2, 16, 128$ and for each $n \in \{4, 8, \ldots, 512\}$, use your thresholds t from part (a) and (b) to define the tests T_n and Ψ_n, and let $\theta_{n,d} \in \mathbb{R}^d$ be an arbitrary vector with $\|\theta_{n,d}\| = 3/\sqrt{n}$. Plot (based on simulation) the powers $\pi_n^T(\theta_{n,d})$ and $\pi_n^\Psi(\theta_{n,d})$ for these n and d.

(d) Explain, in one or two sentences, the behavior in part (c).

Question 4.6 (Relative efficiencies for signed rank tests)**:** Define the kernel function $h(x, y) = \mathbf{1}\{x + y > 0\}$ and U-statistic $U_n = \binom{n}{2}^{-1} \sum_{|\beta|=2} h(X_\beta)$. Consider a null hypothesis H_0 that X has a continuous symmetric density, so that $\theta := \mathbb{E}[h(X_1, X_2)] = \frac{1}{2}$. (See [7, Example 12.4] for asymptotics of this U-statistic.) The signed rank test allows us to test the null that X has symmetric continuous density, and rejects if the null if U_n is large. In this question, we investigate its asymptotic power under local alternatives.

(a) Let P_0 satisfy the null, and suppose that $\{P_t\}_{t \in \mathbb{R}}$ is quadratic mean differentiable at P_0 with score g. Show that for $C_0 = \mathrm{Cov}_0(F(X), g(X))$, where F denotes the CDF of X under P_0,

$$\left(\sqrt{n}(U_n - \theta), \log \frac{dP^n_{t/\sqrt{n}}}{dP_0^n} \right) \xrightarrow[P_0]{d} \mathsf{N}\left(\begin{bmatrix} 0 \\ -(t^2/2)P_0 g^2 \end{bmatrix} \begin{bmatrix} 1/3 & 2t \cdot C_0 \\ 2t \cdot C_0 & t^2 P_0 g^2 \end{bmatrix} \right).$$

(b) Argue (in about one line) that for any $t \in \mathbb{R}$,

$$\sqrt{3n} \left(U_n - \frac{1}{2} - \frac{2tC_0}{\sqrt{n}} \right) \xrightarrow[P_{t/\sqrt{n}}]{d} \mathsf{N}(0, 1).$$

Under the null H_0 that X has a symmetric density, we have limit $\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathsf{N}(0, 1/3)$ (where $\theta = \frac{1}{2}$). The natural signed rank test of asymptotic level α thus rejects if

$$\sqrt{n}(U_n - \theta) \geq \frac{1}{\sqrt{3}} z_{1-\alpha}$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard Gaussian. Let

$$\pi_n(t) := P_t \left(\sqrt{n}(U_n - \theta) \geq \frac{1}{\sqrt{3}} z_{1-\alpha} \right)$$

denote the power of this test under an alternative P_t.

(c) As above, let $\{P_t\}_{t\in\mathbb{R}}$ be QMD at $t = 0$ and P_0 satisfy the null H_0. Show that

$$\lim_{n\to\infty} \pi_n(t/\sqrt{n}) = \Phi\left(z_\alpha + t \cdot 2\sqrt{3} \cdot C_0\right),$$

where Φ is the standard Gaussian CDF.

Now, answer *at least* one of the following parts (d) or (e) (the integrals are a bit tedious):

(d) Let $P_t = \mathsf{N}(t, 1)$ be a mean t Gaussian with unit variance. Show that in this case, the limiting power under local alternatives of the signed rank test is

$$\lim_{n\to\infty} \pi_n(t/\sqrt{n}) = \Phi\left(z_\alpha + t\sqrt{\frac{3}{\pi}}\right).$$

(e) Let P_t denote a Laplace distribution with mean t, that is, P_t has density $\frac{1}{2}\exp(-|x-t|)$. Show that

$$\lim_{n\to\infty} \pi_n(t/\sqrt{n}) = \Phi\left(z_\alpha + t\sqrt{\frac{3}{4}}\right).$$

(f) *Extra credit:* Is the signed rank test asymptotically most powerful against local alternatives $t/\sqrt{n}$, where $t > 0$, for testing symmetry in either the Gaussian or Laplace location families? If not, what is its relative (Pitman) efficiency to an optimal test?

5 Testing

Question 5.1 (Uniform testing vs. pointwise testing)**:** Let $\{P_\theta\}_{\theta\in\Theta}$ be the collection of normal distributions parameterized by $\theta = (\mu, \sigma^2)$ for $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Let $\Theta_0 = \{\theta = (\mu, \sigma^2) \mid \mu = 0\}$ be the collection of mean-zero Gaussian distributions. Let $T_n : \mathbb{R}^n \to \{0, 1\}$ be a test, where 1 indicates rejection of the null, that takes a sample $(X_1, \ldots, X_n)$ and makes a decision. Define

$$\pi_n(\theta) := P_\theta(T_n = 1)$$

to be the power function (where $X_i \stackrel{\text{iid}}{\sim} P_\theta$) of the test.

(a) Let $\alpha \in [0, 1]$ and assume the uniform level guarantee

$$\sup_{\theta\in\Theta_0} \pi_n(\theta) \le \alpha.$$

Show that for all $\epsilon > 0$ and for all $\mu \in \mathbb{R}$, there exists a variance σ^2 such that for $\theta = (\mu, \sigma^2)$,

$$\pi_n(\theta) \le \alpha + \epsilon.$$

That is, uniform guarantees are impossible in this setting of testing a Gaussian mean.

Hint: Note that for any distributions P and Q, $|P(T_n = 1) - Q(T_n = 1)| \le \|P - Q\|_{\text{TV}}$, and use Question 1.9. What is the Hellinger distance between the n-fold product of $\mathsf{N}(\mu, \sigma^2)$ and $\mathsf{N}(0, \sigma^2)$?

(b) Exhibit a test $\psi_n : \mathbb{R}^n \to \{0, 1\}$ for which

$$\sup_{\theta\in\Theta_0} \limsup_{n\to\infty} P_\theta(\psi_n = 1) \le \alpha \quad \text{and} \quad \inf_{\theta\notin\Theta_0} \liminf_{n\to\infty} P_\theta(\psi_n = 1) = 1.$$

Question 5.2 (Asymptotics and tests)**:** Let $\{P_\theta\}_{\theta\in\Theta}$ be a model family as is standard. For a test statistic T_n with rejection region K_n, meaning we reject the null H_0 if $T_n \in K_n$, we define the power function $\pi_n(\theta) := P_\theta(T_n \in K_n)$, so that for a null $H_0 : \theta \in \Theta_0$, the test is level α if $\sup_{\theta\in\Theta_0} \pi_n(\theta) \le \alpha$ and asymptotically of level α if

$$\limsup_n \sup_{\theta\in\Theta_0} \pi_n(\theta) \le \alpha.$$

Given a sample $X_1, \ldots, X_n \in \mathbb{R}$ we consider the sign and mean statistics

$$T_n := \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad S_n := \frac{1}{n}\sum_{i=1}^n \text{sign}(X_i).$$

Consider the normal location family with $P_\theta = \mathsf{N}(\theta, 1)$ and consider testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$, so $\Theta = [0, \infty)$ and we let $\Theta_0 = \{0\}$ and $\Theta_1 = \Theta \setminus \Theta_0 = (0, \infty)$.

(a) Give rejection regions K_n^T for T_n and K_n^S for S_n that yield asymptotically level α tests.

(b) Let $\pi_n^T : \Theta \to [0, 1]$ and $\pi_n^S : \Theta \to [0, 1]$ be the power functions for the two tests. Give formulae for

$$\lim_{n\to\infty} \pi_n^T(\theta) \quad \text{and} \quad \lim_{n\to\infty} \pi_n^S(\theta) \quad \text{for all } \theta \in \Theta.$$

(c) Based on your answer to part (b), which of the test statistics T_n and S_n should you prefer?

(d) Consider a more uniform power calculation using $\inf_{\theta\in\Theta_1} \pi_n(\theta)$. Give formulae for

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta_1} \pi_n^T(\theta) \quad \text{and} \quad \liminf_{n\to\infty} \inf_{\theta\in\Theta_1} \pi_n^S(\theta).$$

(e) Suppose now that the family includes a nuisance parameter of variance, so we have the model $\{\mathsf{N}(\theta, \sigma^2), \theta \geq 0, \sigma^2 > 0\}$. Now the null is the composite null $H_0 : \{\theta = 0, \sigma^2 > 0\}$. (We abuse notation and write $P \in H_0$ to say that $P = \mathsf{N}(0, \sigma^2)$ for some $\sigma^2 > 0$.) Using the *same* rejection regions K_n^T and K_n^S you developed in part (a), evaluate

$$\limsup_n \sup_{P\in H_0} P(T_n \in K_n^T) \quad \text{and} \quad \limsup_n \sup_{P\in H_0} P(S_n \in K_n^S).$$

(f) Give formulae for $\lim_{n\to\infty} P(T_n \in K_n^T)$ and $\lim_{n\to\infty} P(S_n \in K_n^S)$ for each $P \notin H_0$. Which test do you prefer?

We consider a last comparison. Repeat the following $N = 500$ times. For $n \in \{50, 100, 200, 400\}$, generate i.i.d. samples $X_i \stackrel{\text{iid}}{\sim} \mathsf{N}(\theta_n, 1)$, $i = 1, \ldots, n$, setting θ_n^h to be the local perturbation

$$\theta_n^h = 0 + \frac{h}{\sqrt{n}}, \quad h \in \{0, .1, \ldots, 4.9, 5.0\} = \{k/10 \mid k \in \{0, 1, \ldots, 50\}\}.$$

(A total of $N \times 4 \times 51$ different samples.) For each sample you generate, compute T_n and S_n.

(g) Using your sampled data and rejection regions (with $\alpha = .05$) K_n from above, approximate $\pi_n^T(\theta_n^h)$ and $\pi_n^S(\theta_n^h)$ as h and n vary. Plot the function $h \mapsto \pi_n(\theta_n^h)$ for each $n \in \{50, 100, 200, 400\}$. Which of the tests T_n and S_n do you prefer?

Question 5.3: We have a family of distributions $\mathcal{P}$ on a space $\mathcal{X}$ and a parameter of interest $\theta : \mathcal{P} \to \mathbb{R}$, that is, we would like to test the value of $\theta(P)$—a nonparametric testing problem. We consider (local) perturbations of a fixed distribution $P_0 \in \mathcal{P}$, where we assume the parameter is differentiable in $L^2(P_0)$ for a collection of models around P_0. What this means is that for a bounded function $\phi : \mathbb{R} \to \mathbb{R}_+$ with $\phi(0) = \phi'(0) = 1$, continuously differentiable in a neighborhood of 0, there exists a bounded linear functional $D_0 : L^2(P_0) \to \mathbb{R}$ such that for any P_0-mean-zero $g \in L^2(P_0)$, if we define the tilted distributions around P_0 by

$$dP_t(x) = \frac{1}{C(t)} \phi(tg(x)) dP_0(x), \quad C(t) = \int \phi(tg) dP_0,$$

then

$$\lim_{t\downarrow 0} \frac{\theta(P_t) - \theta(P_0)}{t} = D_0(g),$$

where we take g implicitly in the definition of P_t. (We roughly think of $dP_t = (1 + tg)dP_0$.) By the Riesz representation theorem, it is necessarily the case that there exists a mapping $\dot{\theta}_0 : \mathcal{X} \to \mathbb{R}$ with $\dot{\theta}_0 \in L^2(P_0)$ such that

$$D_0(g) = \int g(x)\dot{\theta}_0(x) dP_0(x),$$

and as the tilts are defined only for g satisfying $P_0 g = 0$, it is no loss of generality (by shifting) to assume $P_0\dot{\theta}_0 = 0$.

(a) Abusing notation to define P_n by the density $dP_n = \phi(g/\sqrt{n})dP_0/C(1/\sqrt{n})$ for $n \in \mathbb{N}$, $g \in L^2(P_0)$ with $P_0 g = 0$ as above, use the results of Question 10.5 to show that

$$\lim_{n\to\infty} d_{\mathrm{hel}}^2(P_n^n, P_0^n) = 1 - \exp\left(-\frac{1}{8}P_0 g^2\right).$$

(b) With the same abuse of notation, use the variation/Hellinger bounds in Question 1.9 to show that

$$\liminf_{n\to\infty} \inf_{T_n} \{P_n^n(T_n \neq 1) + P_0^n(T_n \neq 0)\} \geq 1 - \sqrt{1 - \exp\left(-\frac{1}{4}P_0 g^2\right)},$$

where the infimum is taken over all tests $T_n : \mathcal{X}^n \to \{0, 1\}$. That is, the difficulty of testing the null $H_0 : X_i \overset{\text{iid}}{\sim} P_0$ against the (sequence of) alternative(s) $H_1 : X_i \overset{\text{iid}}{\sim} P_n$ is non-trivial, where P^n denotes the product distribution of $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P^n$.

(c) Show that if $\theta : \mathcal{P} \to \mathbb{R}$ is differentiable in L^2 at P_0 as above and $c < \infty$, then for any sequence of tests $T_n : \mathcal{X}^n \to \{0, 1\}$ of

$$H_0 : \theta(P) = \theta(P_0) \quad \text{versus} \quad H_{1,n} : \theta(P) \geq \theta(P_0) + c\frac{\|\dot{\theta}_0\|_{L^2(P_0)}}{\sqrt{n}},$$

there exist sequences of distributions satisfying $H_0, H_{1,n}$ for all large enough n such that

$$\liminf_n \left\{P_{H_0}^n(T_n \neq 0) + P_{H_{1,n}}^n(T_n \neq 1)\right\} \geq 1 - \sqrt{1 - e^{-c^2/4}} > 0.$$

That is, the tests must have asymptotically non-negligible Type I plus Type II error, and as p is independent of all other problem parameters, the scaling $\|\dot{\theta}_0\|_{L^2(P_0)}/\sqrt{n}$ is the "best" possible. *Hint:* let $g(x) = h\dot{\theta}_0(x)/(P_0\dot{\theta}_0^2)^{1/2}$ for some $h > c$, and consider densities defined by $dP_n = \phi(g/\sqrt{n})dP_0/C(1/\sqrt{n})$.

6 Concentration inequalities

Question 6.1 (Sub-Gaussianity of bounded R.V.s)**:** Let X be a random variable taking values in $[a, b]$ with probability distribution P. You may assume w.l.o.g. that $\mathbb{E}[X] = 0$. Define the cumulant generating function $\varphi(\lambda) := \log \mathbb{E}_P[e^{\lambda X}]$, and let Q_λ be the distribution on X defined by

$$dQ_\lambda(x) := \frac{e^{\lambda x}}{\mathbb{E}_P[e^{\lambda X}]} dP(x).$$

You may assume that differentiation and computation of expectations may be exchanged (this is valid for bounded random variables).

(a) Show that $\mathrm{Var}(Y) \le \frac{(b-a)^2}{4}$ for any random variable Y taking values in $[a, b]$.

(b) Show that $\varphi'(\lambda) = \mathbb{E}_{Q_\lambda}[X]$ and $\varphi''(\lambda) = \mathrm{Var}_{Q_\lambda}(X)$.

(c) Show that $\varphi(\lambda) \le \frac{\lambda^2 (b-a)^2}{8}$ for all $\lambda \in \mathbb{R}$.

With these three parts, you have shown that if $X \in [a, b]$, then X is $\frac{(b-a)^2}{4}$ sub-Gaussian.

Question 6.2: Let X_i be independent mean-zero random variables with $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[|X_i|^k] < \infty$ for some $k \ge 1$. Let $S_n = \sum_{i=1}^n X_i$.

(a) Prove that

$$\mathbb{E}[|S_n|^k] \le C_k \mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{\frac{k}{2}}\right]$$

for a constant C_k that depends only on k.

Show the following consequences of this inequality, which apply when $k \ge 2$:

(b) $\mathbb{E}[|S_n|^k] \le C_k \cdot \frac{1}{n} \sum_{i=1}^n \mathbb{E}[|X_i|^k] \cdot n^{k/2}$.

(c) If $\mathbb{E}[|X_i|^k] \le \sigma^k$ for some $\sigma < \infty$ for all i, then $\mathbb{P}(|n^{-1} S_n| \ge t) \le C_k (\frac{\sigma^2}{nt^2})^{\frac{k}{2}}$. How does this compare to Chebyshev's inequality?

The following exercises require basic knowledge of martingales. If you have not seen martingales, we give a workable definition here that should allow solutions of the exercises. Let $X_1, X_2, \ldots$ be a sequence of random variables, and let $Z_1, Z_2, \ldots$ be another sequence of random variables, where Z_k is a function of $X_1, \ldots, X_k$. Then $\{Z_k\}$ is a martingale sequence adapted to $\{X_k\}$ if

$$\mathbb{E}[Z_k \mid X_1, \ldots, X_{k-1}] = Z_{k-1}$$

for all k. Given a martingale $\{Z_k\}$, we say that $\Delta_k = Z_k - Z_{k-1}$ is the associated martingale difference sequence. Any sequence of random vectors or variables $\{\Delta_k\}$ that is adapted to $\{X_k\}$, meaning that Δ_k is a function of $X_1, \ldots, X_k$, is a *martingale difference sequence* if

$$\mathbb{E}[\Delta_k \mid X_1, \ldots, X_{k-1}] = 0 \quad \text{for all } k.$$

Question 6.3 (Azuma's inequality)**:** Let $\mathcal{F}_k = \{X_1, \ldots, X_k\}$. We say a martingale $\{Z_k\}$ adapted to $\{X_k\}$ is σ_k^2-sub-Gaussian if for $\Delta_k = Z_k - Z_{k-1}$, we have for each k that

$$\mathbb{E}\left[\exp\left(\lambda \Delta_k\right) \mid \mathcal{F}_{k-1}\right] \le \exp\left(\frac{\lambda^2 \sigma_k^2}{2}\right)$$

with probability 1 over the randomness in $X_1, X_2, \ldots$. Let Δ_k be a σ_k^2-sub-Gaussian martingale difference sequence with $Z_k = \sum_{i=1}^k \Delta_i$. Show that Z_k is $\sum_{i=1}^k \sigma_i^2$-sub-Gaussian, and hence

$$\mathbb{P}(Z_k \geq t) \vee \mathbb{P}(Z_k \leq -t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^k \sigma_i^2}\right) \quad \text{for } t \geq 0.$$

Question 6.4 (Doob martingales and the bounded-differences inequality)**:** Let $f : \mathcal{X}^n \to \mathbb{R}$ be an arbitrary function and let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables taking values in $\mathcal{X}$. The *Doob martingale* associated to f is

$$Z_k := \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_k].$$

(a) Show that Z_k is a martingale adapted to $\{X_k\}$ and that $Z_n = f(X_1, \ldots, X_n)$.

Now, suppose the function f satisfies *bounded differences with parameters* c_i, meaning that

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}^{n+1}} |f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i \quad \text{for all } i.$$

(b) Show that the associated Doob martingale has bounded differences with $|Z_k - Z_{k-1}| \leq c_k$.

(c) Prove the bounded differences inequality (also known as McDiarmid's inequality): if $X_1, \ldots, X_n$ are independent, then for all $t \geq 0$,

$$\mathbb{P}(|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Question 6.5 (Orlicz norms)**:** Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex increasing function with $\psi(0) = 0$. (Note that $\psi(t) > 0$ for all $t > 0$ as ψ is increasing.) Then for an $\mathbb{R}$-valued random variable X, the *Orlicz norm* of X is

$$\|X\|_\psi := \inf\{t \in \mathbb{R}_+ \mid \mathbb{E}[\psi(|X|/t)] \leq 1\}.$$

In this question, we identify a few properties of these norms, including that they actually are norms.

(a) Show that if $\psi(x) = x^p$, then the Orlicz norm is the standard L^p norm of a random variable, that is, $\|X\|_\psi = \mathbb{E}[|X|^p]^{1/p}$.

(b) Show that quantity $\|X\|_\psi$ is convex in the random variable X. *Hint:* You may use that the perspective transform of a convex function $f : \mathbb{R}^n \to \mathbb{R}$, given by

$$g(x,t) := \begin{cases} tf(x/t) & \text{if } t > 0 \\ +\infty & \text{otherwise,} \end{cases}$$

is jointly convex in its arguments. Use Question 2.8.

(c) Let h be a function on some vector space $\mathcal{X}$. Show that the following conditions are equivalent.

(i) h is convex, symmetric (so that $h(x) = h(-x)$), and positively homogeneous, meaning that $h(\lambda x) = \lambda h(x)$ for $\lambda \geq 0$.

(ii) h is a seminorm on $\mathcal{X}$.

(d) Show that the Orlicz norm $\|\cdot\|_\psi$ is indeed a norm on the space of random variables.

Question 6.6 (Orlicz norms and moment generating functions)**:** Let the function

$$\psi_q(v) := \exp(|v|^q) - 1$$

for $q \in [1, 2]$. We consider the associated Orlicz norms $\|X\|_{\psi_q}$.

(a) Show that for all $t \geq 0$,

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^q / \|X\|_{\psi_q}^q).$$

Thus random variables with Orlicz norms enjoy strong concentration properties.

(b) Show that if $X_1, \ldots, X_n$ are random variables with $\max_j \|X_j\|_{\psi_q} < \infty$, then

$$\mathbb{E}[\max_{j \leq n} |X_j|^q] \leq \max_{j \leq n} \|X_j\|_{\psi_q}^q \log(2n) \quad \text{and} \quad \mathbb{E}[\max_{j \leq n} |X_j|] \leq \max_{j \leq n} \|X_j\|_{\psi_q} \log^{1/q}(2n).$$

Recall that a mean zero X is σ^2-sub-Gaussian if $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\sigma^2/2)$.

(c) Show that if X is σ^2-sub-Gaussian and mean-zero, then

$$\mathbb{E}[\exp(\lambda X^2)] \leq \frac{1}{\sqrt{(1 - 2\lambda\sigma^2)_+}} \quad \text{for } \lambda \geq 0.$$

Hint: If $Z \sim \mathsf{N}(0, \tau^2)$, then $\mathbb{E}[\exp(\lambda Z)] = \exp(\lambda^2\tau^2/2)$, and it is possible to exactly calculate $\mathbb{E}[e^{\lambda Z^2}]$. Use the quantity $\mathbb{E}[e^{\lambda XZ}]$.

(d) Show that if X is σ^2-sub-Gaussian and mean-zero, then $\|X\|_{\psi_2} \leq C\sigma$ for some $C \leq \sqrt{8/3}$.

(e) Show that if $\|X\|_{\psi_2} \leq \sigma$ and X is mean zero, then X is $C\sigma^2$-sub-Gaussian for some constant C. *Hint:* You may cite results from Section 2.3 of Vershynin [8].

Question 6.7 (Orlicz norms: properties)**:** In this question, we enumerate a few properties of Orlicz norms. Let $\psi_q(t) = e^{t^q} - 1$ as in Question 6.5.

(a) Show that if X and Y are sub-Gaussian random variables, which may be dependent, then

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}.$$

(b) Show that for any random variable X and any increasing convex $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\psi(0) = 0$, $\|X - \mathbb{E}[X]\|_\psi \leq 2\|X\|_\psi$.

(c) Show that the inequality in part (b) is tight, that is, show that for all $\epsilon > 0$ there is a random variable X and ψ such that $\|X - \mathbb{E}[X]\|_\psi \geq (2 - \epsilon)\|X\|_\psi$. (Note that for $\psi(t) = t^2$, then $\|X\|_\psi \geq \|X - \mathbb{E}[X]\|_\psi$, so there are indeed ψ such that the inequality holds with constant 1.)

Question 6.8 (Variance of norms under finite moment assumptions, Vershynin [9], Ex. 3.1.6)**:** Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent coordinates satisfying $\mathbb{E}[X_i^2] = 1$ and $\mathbb{E}[X_i^4] \leq \sigma^4$. Show that

$$\mathrm{Var}(\|X\|_2) \leq C \cdot \sigma^4$$

for a numerical constant C.

Hint: First check (by expansion) that $\mathbb{E}[(\|X\|_2^2 - n)^2] \le \sigma^4 n$. Show that this yields $\mathbb{E}[(\|X\|_2 - \sqrt{n})^2] \le \sigma^4$, then replace $\sqrt{n}$ by $\mathbb{E}[\|X\|_2]$.

Question 6.9: Let Z_i, $i = 1, \ldots, n$, be independent standard Gaussians. Show that $\mathbb{E}[\max_i Z_i] \ge (1 - o(1))\sqrt{2 \log n}$ as $n \uparrow \infty$.

Hint: You may use the inequality for the Gaussian CDF that

$$1 - \frac{1}{\sqrt{2\pi}} \frac{1}{t} e^{-t^2/2} \le \Phi(t) \le 1 - \frac{1}{\sqrt{2\pi}} \frac{t}{t^2 + 1} e^{-t^2/2}$$

valid for all $t \ge 0$. (Try to prove these if you like!)

7 Uniform laws of large numbers and related problems

7.1 Uniform laws of large numbers

Question 7.1: Let the pairs $z = (x, y) \in \mathbb{R}^d \times \{-1, 1\}$, and consider the logistic loss $m_\theta(z) = \log(1 + \exp(-y\theta^T x))$, with population expectation $M(\theta) := \mathbb{E}[m_\theta(X, Y)]$ for $(X, Y) \sim P$.

(a) Show that if $\Theta \subset \mathbb{R}^d$ is a compact set and $\mathbb{E}[\|X\|] < \infty$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$, then

$$\sup_{\theta \in \Theta} |P_n m_\theta(X, Y) - M(\theta)| \xrightarrow{p} 0.$$

(b) Assume that Θ is contained in the norm ball $\{\theta \in \mathbb{R}^d : \|\theta\| \le r\}$ and that X is supported on the dual norm ball $\{x \in \mathbb{R}^d : \|x\|_* \le M\}$.[2] Show that there is a numerical constant $C < \infty$ such that for all $\delta \in (0, 1)$,

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |P_n m_\theta(X, Y) - M(\theta)| \ge \epsilon_n(\delta)\right) \le \delta \quad \text{where} \quad \epsilon_n(\delta) = C\sqrt{\frac{r^2 M^2}{n}\left(d \log n + \log \frac{1}{\delta}\right)}.$$

Question 7.2 (Rademacher complexities)**:** In this question, we explore a way to provide finite-sample uniform convergence guarantees. Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$, and let $\varepsilon_i \in \{-1, 1\}$ be an i.i.d. random sign sequence, (known as *Rademacher* variables). For a distribution P on (independent) random variables $X_1, \ldots, X_n$, we define the (normalized) Rademacher complexities

$$R_n(\mathcal{F} \mid X_{1:n}) := \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \mid X_{1:n}\right] \quad \text{and} \quad R_n(\mathcal{F}) = \mathbb{E}[R_n(\mathcal{F} \mid X_{1:n})].$$

Let P_n denote the empirical expectation function given a sample $X_1, \ldots, X_n$.

(a) Show that $\mathbb{E}[\sup_{f \in \mathcal{F}} |P_n f - Pf|] \le 2R_n(\mathcal{F})$.

(b) Assume that $\mathcal{F}$ satisfies the envelope condition $\sup_{x \in \mathcal{X}} \sup_{f \in \mathcal{F}} |f(x) - Pf| \le M$. Show that $h(X_1, \ldots, X_n) := \sup_{f \in \mathcal{F}} |P_n f - Pf|$ has bounded differences and specify its parameters c_i.

(c) Show that for some numerical constant $c > 0$, for all $t \ge 0$ we have

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |P_n f - Pf| \ge 2R_n(\mathcal{F}) + t\right) \le 2\exp\left(-\frac{cnt^2}{M^2}\right).$$

Question 7.3 (Rademacher complexities of some function classes)**:** For this question, use the normalized Rademacher complexity as in Q. 7.2.

(a) Let X_i be independent with support $\{x \in \mathbb{R}^d : \|x\|_2 \le M\}$. Let $\mathcal{F}$ be functions of the form $x \mapsto \langle \theta, x \rangle$ for $\theta \in \Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le r\}$. Give an upper bound on $R_n(\mathcal{F})$.

(b) Let X_i be independent with support $\{x \in \mathbb{R}^d : \|x\|_\infty \le M\}$. Let $\mathcal{F}$ be functions of the form $x \mapsto \langle \theta, x \rangle$ for $\theta \in \Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le r\}$. Give an upper bound on $R_n(\mathcal{F})$.

[2] Recall that for a norm $\|\cdot\|$ on $\mathbb{R}^d$, the dual norm is $\|y\|_* = \sup_x \{x^T y : \|x\| \le 1\}$.

[*Hint:* Do **not** use chaining.]

Question 7.4 (Margin-based model fitting)**:** Consider a binary classification problem with data in pairs $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$, and let $\phi : \mathbb{R} \to \mathbb{R}_+$ be a 1-Lipschitz non-increasing convex function. (For example, we might take $\phi(t) = \log(1 + e^{-t})$ or $\phi(t) = (1 - t)_+$.) Let $m_\theta(x, y) = \phi(y\theta^\top x)$, and given an i.i.d. sample $\{X_i, Y_i\}_{i=1}^n$, consider the empirical risk minimization procedure

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n m_\theta(X_i, Y_i) = \operatorname*{argmin}_{\theta \in \Theta} P_n m_\theta. \tag{7.1}$$

The following result, known as the Ledoux-Talagrand Rademacher contraction inequality, may be useful for this question. Let $\phi \circ \mathcal{F} = \{h : h(x) = \phi(f(x)), f \in \mathcal{F}\}$ denote the composition of ϕ with functions in $\mathcal{F}$. If φ is an L-Lipschitz function with $\varphi(0) = 0$, then $R_n(\varphi \circ \mathcal{F}) \le L R_n(\mathcal{F})$.

(a) In one word, is the procedure (7.1) likely to give a reasonably good classifier? You may assume $\phi(t)$ is strictly decreasing on $t \in [-1, 1]$.

(b) Let $\Theta \subset \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le r\}$ and let X_i be supported on the ℓ_2-ball $\{x \in \mathbb{R}^d : \|x\|_2 \le M\}$. Give the smallest $\epsilon_n(\delta, d, r, M)$ you can—ignoring numerical constants—such that

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |P_n m_\theta - P m_\theta| \ge \epsilon_n(\delta, d, r, M)\right) \le \delta.$$

How does your ϵ_n compare with Question 7.1?

(c) Let $\Theta \subset \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le r\}$ and let X_i be supported on the ℓ_∞-ball $\{x \in \mathbb{R}^d : \|x\|_\infty \le M\}$. Give the smallest $\epsilon_n(\delta, d, r, M)$ you can—ignoring numerical constants—such that

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |P_n m_\theta - P m_\theta| \ge \epsilon_n(\delta, d, r, M)\right) \le \delta.$$

How does your ϵ_n compare with Question 7.1?

Question 7.5 (Dvoretzky-Kiefer-Wolfowitz inequality)**:** Let $\mathcal{F} = \{\mathbf{1}\{x \le t\} \mid t \in \mathbb{R}\}$ be the collection of indicator functions for $x \leqslant t$. Let the $L_2(P)$ metric on $\mathcal{F}$ be defined by $\|f - g\|_{L_2(P)}^2 = \int (f(x) - g(x))^2 dP(x)$.

(a) Show that the covering numbers for $\mathcal{F}$ in $L_2(P)$-norm satisfy

$$\sup_P \log N(\mathcal{F}, L_2(P), \epsilon) \le C \log\left(1 + \frac{1}{\epsilon}\right),$$

where the supremum is over all probability distributions and C is a numerical constant.

For the next two parts of the question, the following notation and quantity may be helpful. For a sample $X_1, \ldots, X_n$ with empirical distribution P_n, let $\|f\|_{L_2(P_n)}^2 = \int f(x)^2 dP_n(x) = \frac{1}{n}\sum_{i=1}^n f(X_i)^2$.

(b) Show that $R_n(\mathcal{F}) \le \frac{C}{\sqrt{n}}$, where C is a universal (numerical) constant.

(c) Prove a (weaker) version of the Dvoretzky-Kiefer-Wolfowitz inequality, that is, that

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} |P_n(X \le t) - P(X \le t)| \ge \frac{C}{\sqrt{n}} + \epsilon\right) \le 2e^{-cn\epsilon^2},$$

where c, C are absolute constants. (In fact, $c = 2$ is possible using tools we have covered.)

Question 7.6 (Smallest eigenvalue of a random, possibly heavy-tailed matrix)**:** Let X_i be i.i.d. $\mathbb{R}^d$-valued random vectors, mean zero, where $\mathrm{Cov}(X_i) = \Sigma$ for a positive definite Σ. Assume also that $\mathbb{E}[|\langle v, X\rangle|] \geq \kappa\sqrt{v^T\Sigma v}$ for any vector $v \in \mathbb{R}^d$, where $\kappa > 0$ is a constant.

(a) Show that for any vector $v \in \mathbb{R}^d$,

$$\mathbb{P}\left(|\langle v, X\rangle| \geq \frac{\kappa}{2}\sqrt{v^T\Sigma v}\right) \geq \frac{\kappa^2}{4}.$$

(b) Let $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ denote the empirical second-moment matrix of the X_i, and for a symmetric matrix A, let

$$\lambda_{\min}(A) := \inf_v \left\{v^T A v \mid v \in \mathbb{S}^{d-1}\right\}$$

denote the minimum eigenvalue of A, where $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$ denotes the sphere in $\mathbb{R}^d$. Show that there exist constants $C_1, C_2, C_3 \in (0, \infty)$, which may depend on κ, such that

$$\mathbb{P}\left(\lambda_{\min}(\widehat{\Sigma}_n) \geq \left(C_1 - C_2\sqrt{\frac{d}{n}} - C_3 t\right)_+ \lambda_{\min}(\Sigma)\right) \geq 1 - e^{-nt^2}$$

for all $t \geq 0$.

Question 7.7: Let (T, d) be a bounded metric space and let $(\mathcal{P}, \ell)$ be a collection of labeled nested partitions as in class. That is, $\mathcal{P} = \{\mathcal{P}_k\}_{k\in\mathbb{Z}}$, and within each level k, we have $A \in \mathcal{P}_k$ implies $\mathrm{diam}(A) \leq 2^{-k}$, where k_0 is the smallest integer such that $\mathrm{diam}(T) \leq 2^{k_0}$. Recall the $\gamma_q(T, d)$ functional, defined as

$$\gamma_q(T, d) := \inf_{\mathcal{P},\ell} \sup_{t\in T} \sum_{k\in\mathbb{Z}} 2^{-k} \log^{\frac{1}{q}} \ell(A_k(t)).$$

Show that for some numerical constant C,

$$\gamma_q(T, d) \leq C \int_0^\infty \log^{\frac{1}{q}} N(T, d, \epsilon) d\epsilon,$$

where N denotes the covering number for the set T in metric d.

Question 7.8 (Covering numbers for low-rank matrices)**:** Let $\mathcal{M}_{r,d}$ be the collection of rank r matrices $A \in \mathbb{R}^{d\times d}$ with $\|A\|_{\mathrm{Fr}} = 1$, where we recall that the Frobenius norm $\|A\|_{\mathrm{Fr}}^2 = \sum_{i,j} A_{ij}^2 = \mathrm{tr}(A^T A)$ is the usual Euclidean norm applied to the entries of A. Show that the covering numbers $N(\mathcal{M}_{r,d}, \|\cdot\|_{\mathrm{Fr}}, \epsilon)$ of $\mathcal{M}_{r,d}$ in the Frobenius norm satisfy

$$\log N(\mathcal{M}_{r,d}, \|\cdot\|_{\mathrm{Fr}}, \epsilon) \leq 2rd \log\left(1 + \frac{4r}{\epsilon}\right).$$

Hint: Our solution uses the singular value decomposition that $A = U\Sigma V^T = \sum_{i=1}^r u_i \sigma_i v_i^T$, where $\Sigma \succeq 0$ is diagonal and $U = [u_1 \ \cdots u_r]$ and $V = [v_1 \ \cdots v_r] \in \mathbb{R}^{d\times r}$ are orthogonal, i.e., $U^T U = I_r$ and $V^T V = I_r$. *Note:* It is possible to get slightly sharper bounds than these, but we won't worry about that.

Question 7.9 (Low-rank matrix sensing)**:** In this question, we consider the problem of recovering a low-rank matrix from linear observations, showing that (with high probability) this is possible under

a Gaussian random measurement model. We assume we observe triples $(X_i, Z_i, Y_i) \in \mathbb{R}^d \times R^d \times \mathbb{R}$ where

$$Y_i = \langle X_i Z_i^T, \Theta^\star \rangle = \operatorname{tr}(Z_i X_i^T \Theta^\star) = X_i^T \Theta^\star Z_i \tag{7.2}$$

for X_i and $Z_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, I_d)$ and independent, where $\Theta^\star \in \mathbb{R}^d$ is an unknown rank r matrix. (Here we use the standard notation on matrices that $\langle A, B \rangle = \operatorname{tr}(A^T B)$.) There is no noise in this observation model. We would like to recover $\Theta^\star$ from n such measurements.

(a) Show that for any $d \times d$ matrix A,

$$\mathbb{E}[|X^T A Z|] \geq \frac{2}{\pi} \|A\|_{\text{Fr}} \quad \text{and} \quad \mathbb{E}[|X^T A Z|^2] = \|A\|_{\text{Fr}}^2 .$$

Hint: To prove the first inequality, first condition on Z. Then note that for any norm $\|\cdot\|$ and random vector W, $\mathbb{E}[\|W\|] \geq \|\mathbb{E}[|W|]\|$, where $|W|$ is the elementwise absolute value of W. Recognize that $\|w\| := \sqrt{\sum_{i=1}^d \sigma_i^2 w_i^2}$ is a norm on $w \in \mathbb{R}^d$.

(b) Argue that there exist numerical constants $c_0, c_1 > 0$ such that for any fixed matrix $A \in \mathbb{R}^{d \times d}$, we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n |\langle X_i Z_i^T, A \rangle| \leq c_0 \|A\|_{\text{Fr}} \right) \leq \exp\left(-c_1 n\right) .$$

Hint: For a constant $c > 0$, define the random variables $B_i = 1$ if $|\langle X_i Z_i^T, A \rangle| \geq c \|A\|_{\text{Fr}}$ and $B_i = 0$ otherwise. Use the Paley-Zygmund inequality (Ex. 1.11) to show that $\mathbb{P}(B_i = 1) \geq p$, where $p > 0$ is a numerical constant, and then bound $\mathbb{P}(\overline{B}_n \leq \mathbb{E}[B]/2)$.

(c) Using the covering number bounds in Ex. 7.8, show there exist numerical constants $0 < c_0, c_1$ and $C < \infty$ such that with probability at least $1 - e^{-c_1 n}$,

$$\frac{1}{n} \sum_{i=1}^n |X_i^T A Z_i| \geq c_0 \|A\|_{\text{Fr}} \tag{7.3}$$

for all rank r matrices $A \in \mathbb{R}^{d \times d}$ as long as $n \geq Cdr \log(dr)$. You may assume dr is large if that is convenient. You may also use that

$$\frac{1}{n} \sum_{i=1}^n \left\| Z_i X_i^T \right\|_{\text{Fr}} = \frac{1}{n} \sum_{i=1}^n \|Z_i\|_2 \|X_i\|_2 \leq \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2} \|Z_i\|_2^2 + \frac{1}{2} \|X_i\|_2^2 \right) \stackrel{(\star)}{\leq} 2d$$

where inequality $(\star)$ holds with probability at least $1 - e^{-c_0 dn}$. *Hint:* note that inequality (7.3) is homogeneous in A.

(d) Assume that $\Theta^\star$ is rank r in the sensing model (7.2). Argue that there exist numerical constants $0 < c_0, c_1$ and $C < \infty$ such that with probability at least $1 - e^{-cn}$,

$$\frac{1}{n} \sum_{i=1}^n |X_i^T \Theta Z_i - Y_i| \geq c_0 \|\Theta - \Theta^\star\|_{\text{Fr}}$$

simultaneously for all rank r matrices Θ as long as $n \geq Cdr \log(dr)$.

(e) For loss $\ell(t) = |t|$, explain what part (d) tells us about the empirical minimizer

$$\widehat{\Theta}_n := \operatorname*{argmin}_{\Theta \in \mathbb{R}^{d \times d}} \left\{ P_n \ell(\langle X Z^T, \Theta \rangle - Y) \mid \operatorname{rank}(\Theta) \leq r \right\} .$$

In one sentence, compare the sample size n versus the number of parameters in $\Theta^\star \in \mathbb{R}^{d \times d}$.

7.2 Rates of Convergence

Question 7.10: Let $R_n : \Theta \to \mathbb{R}$ be a sequence of (random) criterion functions and $R(\theta) = \mathbb{E}[R_n(\theta)]$ be the associated population criterion. Let $d : \Theta \times \Theta$ be some distance on Θ. Let $\theta_0 = \operatorname{argmin}_\theta R(\theta)$, and for $\delta < \infty$, define $\Theta_\delta = \{\theta : d(\theta, \theta_0) \le \delta\}$. Let $\alpha \in (0, 2)$, $\sigma < \infty$, and $D > 0$, and assume we have the continuity bound

$$\mathbb{E}\left[\sup_{\theta \in \Theta_\delta} |(R_n(\theta) - R(\theta)) - (R_n(\theta_0) - R(\theta_0))|\right] \le \frac{\sigma \delta^\alpha}{\sqrt{n}}$$

for all $\delta \le D$. Assume in addition that for some parameters $\beta \in [1, \infty)$ and $\nu > 0$, we have the growth condition

$$R(\theta) \ge R(\theta_0) + \nu d(\theta, \theta_0)^\beta$$

for $d(\theta, \theta_0) \le D$. Let $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} R_n(\theta)$ and assume that $\widehat{\theta}_n \xrightarrow{p} \theta_0$. Give the largest rate r_n (i.e. a function of n, α, β, ignoring other constants) you can for which

$$r_n d(\widehat{\theta}_n, \theta_0) = O_P(1) \text{ as } n \to \infty.$$

Question 7.11: In some applications (such as imaging), we may often observe noiseless measurements of a linear system, though sometimes (due to sensor failures) we observe simply noise. We would like to estimate the parameters of such a system. More precisely, suppose that we have $X \in \mathbb{R}^d$, and we observe

$$Y_i = X_i^T \theta_0 + \varepsilon_i, \quad \text{where } \varepsilon_i = B_i Z_i.$$

Here $B_i \in \{0, 1\}$ is a Bernoulli variable, independent of Z_i and X_i, indicating failed measurements (though we do not observe this), where $\mathbb{P}(B_i = 0) = p > \frac{1}{2}$ and $\mathbb{P}(B_i = 1) = 1 - p$ (so we are more likely to see a good observation than not). The variables Z_i have arbitrary distribution, independent of X_i, and $\mathbb{E}[|Z_i|] < \infty$. Because of its nice median-like estimating properties, we decide to estimate θ_0 using the absolute loss, $\ell(\theta; x, y) = |x^T \theta - y|$, choosing $\widehat{\theta}_n$ by

$$\widehat{\theta}_n \in \operatorname*{argmin}_\theta R_n(\theta) \text{ where } R_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta; X_i, Y_i).$$

Let $R(\theta) := \mathbb{E}[\ell(\theta; X, Y)] = \mathbb{E}[|X^T \theta - Y|]$ be the population risk (so that R_n is the empirical risk).

(a) Show that for any $\theta \in \mathbb{R}^d$, we have

$$R(\theta) - R(\theta_0) \ge (2p - 1)\mathbb{E}[|X^T(\theta - \theta_0)|].$$

(b) Let $V \in \mathbb{R}$ be any random variable, where $|V| \le D$ with probability 1, and let $\mathbb{E}[V^2] = \sigma^2$. Show that

$$\mathbb{P}(|V| \ge c) \ge \frac{\sigma^2 - c^2}{D^2 - c^2} \text{ for all } c \in [0, \sigma].$$

Now (and for the remainder of the question) we assume that there is a constant $D < \infty$ such that $\|X\|_2 \le D$ with probability 1, i.e. X is supported on the ℓ_2-ball of radius D in $\mathbb{R}^d$. We also assume that the second moment matrix $\mathbb{E}[XX^T] = \Sigma$ where $\Sigma \succ 0$, i.e. Σ is positive definite (full rank).

(c) Show that for any vector $v \in \mathbb{R}^d$,

$$\mathbb{E}[|X^T v|] \ge \rho \cdot \|v\|_2,$$

where $\rho > 0$ is a constant that depends on the distribution of X but is independent of v.

(d) Show that there exists a constant $\sigma < \infty$, which may depend on the diameter D of the support of X and dimension d, such that for all $\delta \geq 0$,

$$\mathbb{E}\left[\sup_{\theta:\|\theta-\theta_0\|\leq\delta} |R_n(\theta) - R(\theta) - (R_n(\theta_0) - R(\theta_0))|\right] \leq \frac{\sigma\delta}{\sqrt{n}}.$$

(e) Based on your answers to parts (c) and (d) and question 7.10, at what rate does $\widehat{\theta}_n$ converge to θ_0? Can you explain this behavior? (You may assume that $\widehat{\theta}_n$ is consistent for θ_0; you may also prove that it is consistent if you like.)

Question 7.12: We wish to estimate the median of the distribution of a random variable X on $\mathbb{R}$, where we assume $\mathbb{E}[|X|] < \infty$. Let the loss function ℓ be defined by

$$\ell(\theta, x) = |\theta - x|.$$

We consider minimizers of the population and empirical risks for the preceding loss, defined by

$$R(\theta) := \mathbb{E}[\ell(\theta, X)] \quad \text{and} \quad R_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} |\theta - X_i|,$$

where X_i are i.i.d. We let $\widehat{\theta}_n = \operatorname{argmin}_\theta R_n(\theta)$ denote the empirical minimizer of the absolute loss.

(a) Show that the risk functional R is minimized at $\theta_0 = \text{Med}(X)$.

(b) Suppose that X has a density f in a neighborhood of θ_0, its median. Show that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathsf{N}\left(0, \frac{1}{4f(\theta_0)^2}\right).$$

Hint: see van der Vaart [7, Theorem 5.23 and Example 5.24], and feel free to cite the results.

Now we consider the problem when X does *not* have a density at its median θ_0. Indeed, assume that

$$\min\left\{P(X \geq \theta_0), P(X \leq \theta_0)\right\} \geq \frac{1}{2} + p_0 \tag{7.4}$$

for some $p_0 > 0$.

(c) Show that under the conditions (7.4),

$$R(\theta) \geq R(\theta_0) + p_0|\theta - \theta_0| \quad \text{for all } \theta \in \mathbb{R}.$$

(d) What is the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ under condition (7.4)?

(e) Give the largest rate r_n you can (only as a function of n) such that $r_n|\widehat{\theta}_n - \theta_0| = O_P(1)$.

Question 7.13 (Moduli of continuity and high probability rates of convergence)**:** In this question, we show how convexity can be extremely helpful for many reasons in estimation and proving rates of convergence, including (more or less) free guarantees of consistency, as well as high-probability convergence possibilities. Let $\theta \in \mathbb{R}^d$ and define

$$f(\theta) := \mathbb{E}[F(\theta; X)] = \int_{\mathcal{X}} F(\theta; x) dP(x)$$

be a function, where $F(\cdot; x)$ is convex in its first argument (in θ) for all $x \in \mathcal{X}$, and P is a probability distribution. We assume $F(\theta; \cdot)$ is integrable for all θ. Recall that a function h is convex

$$h(t\theta + (1-t)\theta') \le th(\theta) + (1-t)h(\theta') \text{ for all } \theta, \theta' \in \mathbb{R}^d, \ t \in [0,1].$$

Let $\theta_0 \in \operatorname{argmin}_\theta f(\theta)$, and assume that f satisfies the following *ν-strong convexity* guarantee:

$$f(\theta) \ge f(\theta_0) + \frac{\nu}{2} \|\theta - \theta_0\|^2 \quad \text{for } \theta \text{ s.t. } \|\theta - \theta_0\| \le \beta,$$

where $\beta > 0$ is some constant. We also assume that the instantaneous functions $F(\cdot; x)$ are L-Lipschitz with respect to the norm $\|\cdot\|$.

Let $X_1, \ldots, X_n$ be an i.i.d. sample according to P, and define $f_n(\theta) := \frac{1}{n} \sum_{i=1}^n F(\theta; X_i)$ and let

$$\widehat{\theta}_n \in \operatorname*{argmin}_\theta f_n(\theta).$$

(a) Show that for *any* convex function h, if there is some $r > 0$ and a point θ_0 such that $h(\theta) > h(\theta_0)$ for all θ such that $\|\theta - \theta_0\| = r$, then $h(\theta') > h(\theta_0)$ for all θ' with $\|\theta' - \theta_0\| > r$.

(b) Show that f and f_n are convex.

(c) Show that θ_0 is unique.

(d) Let

$$\Delta(\theta, x) := [F(\theta; x) - f(\theta)] - [F(\theta_0; x) - f(\theta_0)].$$

Show that $\Delta(\theta, X)$ (i.e. the random version where $X \sim P$) is $4L^2 \|\theta - \theta_0\|^2$-sub-Gaussian.

(e) Show that for some constant $\sigma < \infty$, which may depend on the parameters of the problem (you should specify this dependence in your solution)

$$\mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\| \ge \sigma \cdot \frac{1+t}{\sqrt{n}}\right) \le C \exp\left(-t^2\right)$$

for all $t \le \sigma' \sqrt{n} \beta$, where $\sigma' > 0$ is a constant depending on the parameters of the problem and $C < \infty$ is a numerical constant. *Hint:* The quantity $\Delta_n(\theta) := \frac{1}{n} \sum_{i=1}^n \Delta(\theta, X_i)$ may be helpful, as may be the bounded differences inequality in Question 6.4.

Question 7.14 (Uniform convergence in a quantile regression problem)**:** Let pairs $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}$ and for a fixed $q \in (0,1)$ consider the "pinball" loss

$$\ell(\theta, z) = \ell(\theta, x, y) = q\left(\theta^T x - y\right)_+ + (1-q)\left(y - \theta^T x\right)_+ - q\left(-y\right)_+ - (1-q)\left(y\right)_+, \tag{7.5}$$

where $(a)_+ = \max\{a, 0\}$ is the positive part of its argument. (One uses this loss to fit models that predict quantiles.) Define the population expectation $L(\theta) := \mathbb{E}_P[\ell(\theta, X, Y)]$.

(a) Show that if $\Theta \subset \mathbb{R}^d$ is compact and $\mathbb{E}[\|X\|] < \infty$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$, then

$$\sup_{\theta \in \Theta} |P_n \ell(\theta, X, Y) - L(\theta)| \xrightarrow{p} 0.$$

(b) Explain why we must normalize the losses (7.5) by subtracting $q\left(-y\right)_+ + (1-q)\left(y\right)_+$ to achieve the preceding convergence. (This should only take a sentence or two.)

Now, we derive asymptotics of the empirical minimizer $\widehat{\theta}_n$ of $L_n(\theta) := P_n \ell(\theta, X, Y)$, that is, $\widehat{\theta}_n \in \operatorname{argmin}_{\theta \in \Theta} L_n(\theta)$. You may use the following result:

Lemma 7.14.1 (Bertsekas [3]). *If $H : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ is a convex function with $\mathbb{E}_P[|H(\theta, Z)|] < \infty$ and $\nabla_\theta H(\theta, x)$ exists for P-almost all x, then $h(\theta) := \mathbb{E}_P[H(\theta, Z)]$ is differentiable with gradient*

$$\nabla h(\theta) = \mathbb{E}_P[\nabla H(\theta, Z)] = \int \nabla H(\theta, z) dP(z) = \int_{z \in \mathcal{Z} : \nabla H(\theta, z) \text{ exists}} \nabla H(\theta, z) dP(z).$$

Assume that conditional on $X = x$, the random variable Y has cumulative distribution $F_x(\cdot)$ with continuous bounded positive density $f_x(\cdot)$ on $\mathbb{R}$. Assume additionally that $\operatorname{Cov}(X) \succ 0$, that is, X has full rank covariance with $\mathbb{E}[\|X\|_2^2] < \infty$, and that the population minimizer $\theta^\star = \operatorname{argmin}_{\theta \in \Theta} L(\theta) \in \operatorname{int} \Theta$.

(c) Show that $\widehat{\theta}_n$ is consistent for $\theta^\star$. *Hint:* argue that the Hessian $\nabla^2 L(\theta)$ is positive definite in a neighborhood of $\theta^\star$. Then apply van der Vaart [7, Thm. 5.7]. You may assume you can exchange the order of expectation and differentiation in any integrals you desire. (It is possible to use dominated convergence to prove this valid in any case.)

(d) Show that

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \mathsf{N}\left(0, \nabla^2 L(\theta^\star)^{-1} \operatorname{Cov}(\nabla \ell(\theta^\star, X, Y)) \nabla^2 L(\theta^\star)^{-1}\right).$$

In addition, express the covariance $\operatorname{Cov}(\nabla \ell(\theta^\star, X, Y))$ and Hessian $\nabla^2 L(\theta^\star)$ in terms of expectations involving q and the random variables X, $f_X(\langle \theta^\star, X \rangle)$, and $F_X(\langle \theta^\star, X \rangle)$. *Hint:* you may use van der Vaart [7, Thm. 5.23] to show the claimed convergence.

(e) Suppose there exists $\theta_0 \in \operatorname{int} \Theta$ such that

$$F_x(\theta_0^T x) = 1 - q$$

for P-almost all x, and that the density $f_x(\theta_0^T x) = \rho > 0$ for P-almost all x. This would occur, for example, in the model

$$Y = \langle \beta^\star, X \rangle + \varepsilon, \quad \varepsilon \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$$

so long as x includes the intercept term that $x_1 = 1$ (feel free to convince yourself of this!). Show that your result in part (d) simplifies to

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \mathsf{N}\left(0, \frac{q(1-q)}{\rho^2} \mathbb{E}[X X^T]^{-1}\right).$$

7.3 Comparison inequalities and applications

Question 7.15: We consider a few different contraction inequalities and complexities, relating Gaussian to Rademacher complexities. For this problem, define the Rademacher and Gaussian complexities of a set $T \subset \mathbb{R}^n$ by

$$R_n(T) := \mathbb{E}[\sup_{t \in T} |\langle \varepsilon, t \rangle|] \quad \text{and} \quad G_n(T) := \mathbb{E}[\sup_{t \in T} \langle g, t \rangle]$$

where $\varepsilon_i \overset{\text{iid}}{\sim} \mathsf{Uni}\{\pm 1\}$ and $g \sim \mathsf{N}(0, I_n)$. Note the lack of an absolute value in the Gaussian complexity.

(a) Let $X \sim \mathsf{N}(0, \Sigma)$ be a Gaussian vector. Argue that for any index i_0,

$$\mathbb{E}[\max_{i,j} |X_i - X_j|] = 2\mathbb{E}[\max_i X_i] \quad \text{and} \quad \mathbb{E}[\max_i |X_i|] \le 2\mathbb{E}[\max_i X_i] + \mathbb{E}[|X_{i_0}|].$$

(b) Show that for any[3] set $T \subset \mathbb{R}^n$,

$$R_n(T) \le \sqrt{2\pi} G_n(T) + \sqrt{\frac{\pi}{2}} \inf_{t_0 \in T} \mathbb{E}[|\langle g, t \rangle|].$$

If T is symmetric (so $T = -T$) show that $R_n(T) \le \sqrt{\frac{\pi}{2}} G_n(T)$.

(c) Let $\phi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$, be a M-Lipschitz functions, meaning $|\phi_i(x) - \phi_i(y)| \le M|x - y|$ for $x, y \in \mathbb{R}$, and define $\phi(t) = (\phi_i(t_i))_{i=1}^n$ to be the elementwise application of ϕ. Using the result of part (b), show that

$$R_n(\phi(T)) \le M\sqrt{2\pi} G_n(T) + \sqrt{\frac{\pi}{2}} \inf_{t \in T} \mathbb{E}[|\langle g, \phi(t) \rangle|].$$

(d) For a function class $\mathcal{F} \subset \{\mathbb{R}^d \to \mathbb{R}\}$, define the Rademacher and Gaussian complexities

$$R_n(\mathcal{F} \mid x_1^n) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \quad \text{and} \quad G_n(\mathcal{F} \mid x_1^n) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n g_i f(x_i) \right| \right]$$

for any collection $x_1^n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$. Let the function class $\mathcal{F} = \{f(x) = \langle \theta, x \rangle \mid \|\theta\|_1 \le 1\}$, and let ϕ be 1-Lipschitz with $\phi(0) = 0$. Show that for $\sigma_{n,j}^2 = \sum_{i=1}^n x_{i,j}^2$ (the sum of squares of the jth component of the vectors $x_i \in \mathbb{R}^d$),

$$R_n(\phi \circ \mathcal{F} \mid x_1^n) \le C \sqrt{\max_{j \le d} \sigma_{n,j}^2 \log(2d)}$$

for a numerical constant C.

Question 7.16 (Peeling and normalizing losses)**:** We will investigate a situation in which we can show that a has deviation that (roughly) scales with its expected deviation, allowing us to give a "self-normalizing" high-probability concentration result. For data $(x, y) \in \mathbb{R}^d$, consider losses

$$\ell(\theta, x, y) = |\langle \theta, x \rangle - y|.$$

The population loss of interest is $L(\theta) := P\ell(\theta, X, Y)$, where we assume that $\theta^\star = \operatorname{argmin}_\theta L(\theta)$.[4] Then we have $y = \langle x, \theta^\star \rangle + \xi$, though we do *not* necessarily have that the "noise" ξ is mean zero, symmetric, or independent of x. For empirical loss $L_n = P_n \ell(\cdot, X, Y)$, we would like to give (uniform) deviation bounds on the quantity

$$(L_n(\theta) - L_n(\theta^\star)) - (L(\theta) - L(\theta^\star)) \,.$$

[3] ignoring measurability issues, and assuming that for any random vector X and function f we require that $\mathbb{E}[\sup_{t \in T} f(t, X)] = \sup_{k \in \mathbb{N}} \sup_{|T_0| \le k, T_0 \subset T} \mathbb{E}[\max_{t \in T_0} f(t, X)]$

[4] If Y is not integrable, we could normalize by considering losses $|\langle \theta, x \rangle - y| - |y|$; you should convince yourself this would not change the results of this problem, so these concentration guarantees hold no matter Y's distribution.

(a) For $r \geq 0$, consider the function class

$$\mathcal{F}_r := \{(x, y) \mapsto \ell(\theta, x, y) - \ell(\theta^\star, x, y) \mid \|\theta - \theta^\star\|_2 \leq r\},$$

that is, all functions of the form

$$f(x, y) = \ell(\theta, x, y) - \ell(\theta^\star, x, y) = |\langle \theta, x\rangle - y| - |\langle \theta^\star, x\rangle - y| = |\langle \theta - \theta^\star, x\rangle - \xi| - |\xi|,$$

where $\xi = y - \langle x, \theta^\star\rangle$, as θ ranges over $\|\theta - \theta^\star\|_2 \leq r$. Show that the conditional Rademacher complexity

$$R_n(\mathcal{F} \mid x_1^n, y_1^n) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i f(x_i, y_i)\right|\right]$$

satisfies

$$R_n(\mathcal{F}_r \mid x_1^n, y_1^n) \leq 2r\sqrt{\sum_{i=1}^n \|x_i\|_2^2}.$$

(b) Let $\Theta_r := \{\theta \in \mathbb{R}^d \mid \|\theta - \theta^\star\|_2 \leq r\}$. Assume that under the distribution P, we have $\|X_i\|_2 \leq b$ with probability 1. Show that for some numerical constant $c < \infty$ (i.e., c should be independent of all other parameters in your problem),

$$\mathbb{P}\left(\sup_{\theta \in \Theta_r} |(L_n(\theta) - L_n(\theta^\star)) - (L(\theta) - L(\theta^\star))| \geq c\frac{rb(1+t)}{\sqrt{n}}\right) \leq e^{-t^2}$$

for all $r, t \geq 0$.

(c) Let $0 < \epsilon \leq r < \infty$, and define the ratio

$$\rho_n(r, \epsilon) := \sup_\theta \left\{\frac{|(L_n(\theta) - L_n(\theta^\star)) - (L(\theta) - L(\theta^\star))|}{\|\theta - \theta^\star\|_2} \text{ s.t. } \epsilon \leq \|\theta - \theta^\star\|_2 \leq r\right\}. \tag{7.6}$$

Show that there exists a numerical constant $c < \infty$ (again, independent of all parameters) such that for any $r < \infty$ and $\epsilon > 0$ with $r/\epsilon > \exp(1)$,

$$\mathbb{P}\left(\rho_n(r, \epsilon) \geq cb\sqrt{\frac{1 + \log\log\frac{r}{\epsilon} + u}{n}}\right) \leq e^{-u}$$

for all $u \geq 0$. *Hint.* In the notation of part (b), consider parameter sets of the form $\Theta_{r(k)}$, where $r(k) = 2^k\epsilon$ and k ranges over $\{0, \ldots, \log_2 \frac{r}{\epsilon}\}$. You can assume $\log_2 \frac{r}{\epsilon} \in \mathbb{N}$.

Question 7.17 (Using a normalized concentration inequality)**:** In this question, you will use the results of Question 7.16 to show a few convergence guarantees for empirical minimizers of the robust loss $\ell(\theta, x, y) = |\langle \theta, x\rangle - y|$. Throughout this question, we will make the assumption that

$$Y = \langle X, \theta^\star\rangle + \sigma\xi,$$

where ξ is a symmetric random variable, independent of X, with strictly positive density in a neighborhood of zero (recall Question 2.11), that is, ξ has density π with $\pi(z) \geq p_{\min}$ for $z \in [-\tau, \tau]$, where $p_{\min}, \tau > 0$ are positive. (By a change of variables, this means that $\sigma\xi$ has a density π_σ with $\pi_\sigma(z) \geq \frac{p_{\min}}{\sigma}$ for $z \in [-\tau\sigma, \tau\sigma]$.) We will also assume for simplicity that $X \sim \mathsf{Uni}(\sqrt{d}\mathbb{S}^{d-1})$, that is,

X is uniform on the sphere $\{x \in \mathbb{R}^d \mid \|x\|_2 = \sqrt{d}\}$. In this question, you will develop finite sample convergence rates of the empirical minimizer

$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta} \{L_n(\theta) := P_n \ell(\theta, X, Y)\}$$

to the population minimizer $\theta^\star = \operatorname{argmin}_\theta L(\theta)$, where $L(\theta) = P\ell(\theta, X, Y)$ as usual.

You may assume that $\mathbb{E}[|Y|] < \infty$, though as in footnote 4, we could simply normalize by using losses $|\langle \theta, x\rangle - y| - |y|$. You may also use (without proof or citation) that if $X \sim \mathsf{Uni}(\sqrt{d}\mathbb{S}^{d-1})$, there exists a numerical constant $c_0 > 0$ such that for any fixed vector v, we have $|\langle X, v\rangle| \geq c_0 \|v\|_2$ with probability at least $\frac{1}{2}$.

(a) Let $\sigma > 0$. Show that for a numerical constant $c > 0$, we have

$$L(\theta) - L(\theta^\star) \geq c \frac{p_{\min}}{\sigma} \min\left\{\|\theta - \theta^\star\|_2^2, \sigma\tau \|\theta - \theta^\star\|_2\right\}.$$

Hint: Use Question 2.11.

(b) Let $\rho_n(r, \epsilon)$ be defined as in Eq. (7.6). Show that

$$L_n(\theta) - L_n(\theta^\star) \geq L(\theta) - L(\theta^\star) - \rho_n(r, \epsilon) \|\theta - \theta^\star\|_2$$

simultaneously for all θ satisfying $\|\theta - \theta^\star\|_2 \in [\epsilon, r]$.

(c) Let r be small enough that $r \leq \sigma\tau$ (recall that τ is a parameter of the density π of ξ). Argue that there exists a numerical constant $C < \infty$ such that on the event

$$C \frac{\rho_n(r, \epsilon)\sigma}{p_{\min}} < r,$$

we have

$$L_n(\theta) - L_n(\theta^\star) > 0$$

for all θ satisfying $\|\theta - \theta^\star\|_2 > C \cdot \rho_n(r, \epsilon) \frac{\sigma}{p_{\min}}$. *Hint:* Use Question 2.10 and parts (a)–(b).

(d) Show that there exists a numerical constant $C < \infty$ such that for any $\nu > 0$, for large enough n (which you may feel free to specify), we have

$$\|\widehat{\theta}_n - \theta^\star\|_2^2 \leq C \frac{d}{n} \frac{\sigma^2}{p_{\min}^2} \cdot \nu \log n \quad \text{with probability} \quad \geq 1 - \frac{1}{n^\nu}.$$

Question 7.18 (A weakened Sudakov-Fernique inequality; Ledoux and Talagrand [6])**:** In this question, we will develop a weakened version of the Sudakov-Fernique inequality. The standard version is as follows. Let $X = \{X_i\}_{i=1}^n$ and $Y = \{Y_i\}_{i=1}^n$ be mean-zero Gaussian vectors, where

$$\mathbb{E}[(X_i - X_j)^2] \leq \mathbb{E}[(Y_i - Y_j)^2] \tag{7.7}$$

for all i, j. Then

$$\mathbb{E}[\max_{i \leq n} X_i] \leq \mathbb{E}[\max_{i \leq n} Y_i].$$

We will demonstrate a weaker version of this, which follows from Slepian's inequality; we will show that under the hypothesis (7.7)

$$\mathbb{E}[\max_{i \leq n} X_i] \leq 2\mathbb{E}[\max_{i \leq n} Y_i]. \tag{7.8}$$

(a) Argue that $\mathbb{E}[\max_{i\le n}(X_i - X_1)] = \mathbb{E}[\max_{i\le n} X_i]$ and similarly $\mathbb{E}[\max_{i\le n}(Y_i - Y_1)] = \mathbb{E}[\max_{i\le n} Y_i]$, so that without loss of generality we may assume $X_1 = Y_1 = 0$.

For the rest of the problem, assume that $X_1 = Y_1 = 0$; note that this means $\max_{i\le n} X_i \ge 0$, $\max_{i\le n} Y_i \ge 0$, and $\max_i |Y_i| \le \max_{i,j} |Y_i - Y_j|$.

(b) Define $\sigma^2 := \max_{i\le n} \mathbb{E}[Y_i^2]$ and consider the perturbed random variables

$$X_i' = X_i + \sigma Z \quad \text{and} \quad Y_i' = Y_i + (\sigma^2 + \mathbb{E}[X_i^2] - \mathbb{E}[Y_i^2])^{1/2} Z$$

where $Z \sim \mathsf{N}(0,1)$ is standard normal. Show that $\mathbb{E}[(X_i')^2] = \mathbb{E}[(Y_i')^2]$ and that

$$\mathbb{E}[\max_{i\le n} X_i'] \le \mathbb{E}[\max_{i\le n} Y_i'].$$

(c) Show that $\sigma^2 \le 2\pi \mathbb{E}[\max_{i\le n} Y_i]^2$.

(d) Show inequality (7.8).

8 High-dimensional problems

Question 8.1: Consider the sub-Gaussian sequence model

$$Y = \theta + \sigma\varepsilon, \tag{8.1}$$

where $\varepsilon \in \mathbb{R}^n$ consists of independent mean-zero 1-sub-Gaussian components (for $\theta \in \mathbb{R}^n$). The soft-thresholding operator, defined for $v \in \mathbb{R}$ by

$$\mathsf{S}_\lambda(v) := \operatorname{sign}(v)(|v| - \lambda)_+ = \begin{cases} v - \lambda & \text{if } v \geq \lambda \\ 0 & \text{if } v \in [-\lambda, \lambda] \\ v + \lambda & \text{if } v \leq -\lambda, \end{cases}$$

gives the soft-thresholding estimator (when applied elementwise)

$$\widehat{\theta} := \mathsf{S}_\lambda(Y) = \operatorname*{argmin}_\theta \left\{ \frac{1}{2}\|\theta - Y\|_2^2 + \lambda\|\theta\|_1 \right\}.$$

In this question, we give high-probability bounds on the error of $\widehat{\theta}$ for the sub-Gaussian sequence model (8.1) when θ is k-sparse, meaning that $\|\theta\|_0 = \operatorname{card}\{j \in [n] \mid \theta_j \neq 0\} \leq k$.

(a) Show that if $\lambda \geq \sigma\|\varepsilon\|_\infty$, then

$$\|\widehat{\theta} - \theta\|_2^2 \leq 4k\lambda^2.$$

(b) Show that if

$$\lambda = \lambda(t) := \sqrt{2\sigma^2 \log(2n) + 2\sigma^2 t},$$

where $t \geq 0$, then

$$\sup_{\|\theta\|_0 \leq k} \mathbb{P}\left(\|\widehat{\theta} - \theta\|_2 \geq 2\sqrt{k}\lambda(t) \right) \leq e^{-t}.$$

Conclude that probability at least $1 - \frac{1}{2n}$, in the sub-Gaussian sequence model (8.1), the soft-thresholding estimator with $\lambda = 2\sqrt{\sigma^2 \log(2n)}$ satisfies $\|\widehat{\theta} - \theta\|_2^2 \leq 16k\sigma^2 \log(2n)$.

Question 8.2: A matrix $A \in \mathbb{R}^{n \times d}$, $A = [a_1 \ \cdots \ a_d]$, where $a_i \in \mathbb{R}^d$ are the columns of A, is μ-pairwise incoherent if

$$\delta_{\text{pw}}(A) := \left\| \frac{1}{n} A^T A - I_{d \times d} \right\|_\infty$$

satisfies $\delta_{\text{pw}}(A) \leq \mu$, where $\|\cdot\|_\infty$ denotes the entrywise ℓ_∞ norm (maximum absolute value). Recall that for a set $S \subset [d]$, we define $A_S = [a_i]_{i \in S} \in \mathbb{R}^{n \times |S|}$ to be the matrix whose columns are indexed by S.

(a) Let $S \subset [d]$ have cardinality $|S| = k$. Show that if $\delta_{\text{pw}}(A) \leq \mu$, then the minimal eigenvalue $\lambda_{\min}$ of $\frac{1}{n} A_S^T A_S$ satisfies

$$\lambda_{\min}(n^{-1} A_S^T A_S) \geq 1 - k\mu.$$

(b) Show that if $\delta_{\text{pw}}(A) = \mu < \frac{1}{2k}$, then A satisfies the restricted nullspace property with respect to any set $S \subset [d]$ with $|S| \leq k$. That is, if

$$\mathbb{C}(S) := \{x \in \mathbb{R}^d \mid \|x_{S^c}\|_1 \leq \|x_S\|_1\},$$

then $\operatorname{null}(A) \cap \mathbb{C}(S) = \{0\}$.

Question 8.3 (The square root Lasso)**:** The square-root Lasso chooses the estimator $\widehat{\theta}$ via

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{\sqrt{n}} \|y - X\theta\|_2 + \gamma \|\theta\|_1 \right\}.$$

Assume that $y = X\theta^\star + \varepsilon$ for some vector $\theta^\star$ with support $S = \{j : \theta^\star_j \neq 0\}$ and $\varepsilon \in \mathbb{R}^n$.

(a) Show that the square-root Lasso is equivalent to choosing

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \inf_{\lambda \geq 0} \left\{ \frac{1}{2n} \frac{\|y - X\theta\|_2^2}{\lambda} + \frac{\lambda}{2} + \gamma \|\theta\|_1 \right\}.$$

What value does λ take on at $\theta = \theta^\star$? Why might this be a valuable quantity?

(b) Let $\theta^\star$ have support $S = \operatorname{supp} \theta^\star = \{j \in [d] : \theta^\star_j \neq 0\}$. Show that

$$\frac{1}{\sqrt{n}} \|X\widehat{\theta} - y\|_2 - \frac{1}{\sqrt{n}} \|X\theta^\star - y\|_2 \leq \gamma \left(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1 \right),$$

where $\widehat{\theta} = \theta^\star + \Delta$.

(c) Show that if $y = X\theta^\star + \varepsilon$, then

$$\|X\widehat{\theta} - y\|_2 \geq \|X\theta^\star - y\|_2 - \frac{\left\|X^T\varepsilon\right\|_\infty}{\|\varepsilon\|_2} \|\Delta\|_1.$$

(d) Letting

$$\mathbb{C}_3(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq 3 \|\Delta_S\|_1\}$$

denote (a scaled version of) the critical cone, show that $\Delta \in \mathbb{C}_3(S)$ whenever $\gamma \geq 2\frac{\left\|X^T\varepsilon\right\|_\infty}{\sqrt{n}\|\varepsilon\|_2}$.

(e) Show that any solution $\widehat{\theta}$ to the square-root Lasso satisfies

$$\frac{\frac{1}{n}X^T(X\widehat{\theta} - y)}{\frac{1}{\sqrt{n}}\|X\widehat{\theta} - y\|_2} + \gamma z = 0$$

for some $z \in \partial\|\widehat{\theta}\|_1$, the subdifferential of the ℓ_1-norm at $\widehat{\theta}$.

(f) Using the previous part, derive the following extension of the basic inequality for the square root lasso:

$$\begin{aligned} \frac{1}{n} \|X\Delta\|_2^2 &\leq \frac{1}{n}\langle \Delta, X^T\varepsilon \rangle + \frac{\gamma}{\sqrt{n}} \|X\Delta - \varepsilon\|_2 \left(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1 \right) \\ &\leq \frac{1}{n}\langle \Delta, X^T\varepsilon \rangle + \frac{\gamma}{\sqrt{n}} \|\varepsilon\|_2 \|\Delta_S\|_1 + \gamma^2 \|\Delta_S\|_1^2. \end{aligned}$$

(You should prove both inequalities.)

(g) Suppose that X satisfies the restricted strong convexity condition that $\frac{1}{n}\|X\Delta\|_2^2 \geq \mu \|\Delta\|_2^2$ for all $\Delta \in \mathbb{C}_3(S)$ and $\gamma \geq 2\frac{\|X^T\varepsilon\|_\infty}{\sqrt{n}\|\varepsilon\|_2}$. Show that if $k = |S|$, then

$$(\mu - k\gamma^2) \|\Delta\|_2 \leq 3\frac{\gamma \|\varepsilon\|_2 \sqrt{k}}{\sqrt{n}}.$$

(h) Assume $X \in \mathbb{R}^{n \times d}$ has columns with norm $\|X_i\|_2^2 = n$ and satisfies restricted strong convexity as in part (g); assume also that ε_i are independent, mean zero, $C\sigma^2$-sub-Gaussian, and $\mathbb{E}[\varepsilon_i^2] = \sigma^2$. Argue that if $\gamma = C\frac{\sqrt{\log d}}{\sqrt{n}}$ for a constant C, then $\|\widehat{\theta} - \theta^\star\|_2^2 \lesssim \sigma^2 \frac{k \log d}{n}$ with high probability.

(i) What advantages does this have (in one sentence) over the standard Lasso program?

Question 8.4 (ℓ_∞ bounds on the Lasso)**:** Consider the sparse linear regression model $Y = X\theta^\star + \varepsilon$, $X \in \mathbb{R}^{n \times d}$, where $\varepsilon \in \mathbb{R}^n$ consists of independent σ^2-sub-Gaussian noise and $\mathrm{supp}(\theta^\star) = S \subset [d]$. Let $\widehat{\Sigma} := \frac{1}{n} X^T X$, and assume that the diagonal $\mathrm{diag}(\widehat{\Sigma}) \le 1$ (i.e. the entries are uniformly bounded by 1) and that we have the ℓ_∞ growth condition

$$\|\widehat{\Sigma}\Delta\|_\infty \ge \mu \|\Delta\|_\infty \quad \text{for all } \Delta \in \mathbb{C}_3(S), \tag{8.2}$$

where we recall that $\mathbb{C}_3(S) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1\}$.

(a) Show that with regularization parameter $\lambda_n = 4\sigma\sqrt{\frac{\log d}{n}}$, any Lasso solution satisfies the ℓ_∞ bound

$$\|\widehat{\theta} - \theta^\star\|_\infty \le C\frac{\sigma}{\mu}\sqrt{\frac{\log d}{n}}$$

with high probability, where C is a numerical constant.

(b) Under the same conditions, show that if $|S| \le k$, we have the ℓ_1 bound

$$\|\widehat{\theta} - \theta^\star\|_1 \le C'\frac{\sigma}{\mu} k\sqrt{\frac{\log d}{n}}$$

for a numerical constant C'.

Hint: Use the subdifferential optimality condition for convex optimization that x minimizes $f(x)$ if and only if $0 \in \partial f(x)$.

Question 8.5 (Verifying the ℓ_∞ growth condition)**:** Let $X \in \mathbb{R}^{n \times d}$ have i.i.d. $\mathsf{N}(0,1)$ entries. Show that with high probability, $\widehat{\Sigma} := \frac{1}{n} X^T X$ satisfies

$$\|\widehat{\Sigma}\Delta\|_\infty \ge \mu \|\Delta\|_\infty \quad \text{for all } \Delta \in \mathbb{C}_3(S)$$

with high probability as long as $n \ge Ck^2 \log d$, where $0 < \mu, C < \infty$ are numerical constants and $k = |S|$.

Hint: Do not use chaining or anything like that.

Question 8.6: Perform the following simulation, doing each experiment a total of $T = 20$ times. Draw a data matrix $X \in \mathbb{R}^{n \times d}$, where $n = 50$ and $d = 200$, and $k = 5$, and choose $\theta^\star \in \mathbb{R}^d$ uniformly at random from k-sparse vectors with $\|\theta^\star\|_2 = 1$. For values of $\sigma^2 \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^2, 10^3\}$, set

$$Y = X\theta^\star + \sigma\varepsilon, \quad \varepsilon \sim \mathsf{N}(0, I_n).$$

Then solve the following two lasso problems:

$$\text{[Square-root]} \quad \underset{\theta}{\text{minimize}} \ \frac{1}{\sqrt{n}} \|X\theta - Y\|_2 + \lambda_n \|\theta\|_1$$

$$\text{[Standard]} \quad \underset{\theta}{\text{minimize}} \ \frac{1}{2n} \|X\theta - y\|_2^2 + \lambda_n \|\theta\|_1,$$

where $\lambda_n = 2\sqrt{\frac{\log d}{n}}$ for both.

For each value of σ^2, plot the mean error $\|\widehat{\theta} - \theta^\star\|_2$ for each of the square-root and standard lassos as well as the variance across your experiments. Explain your results in a few sentences.

Hint: You may wish to use the `CVX` package, available for R (https://rviews.rstudio.com/2017/11/27/introduction-to-cvxr/), Julia (https://github.com/JuliaOpt/Convex.jl), Python (https://cvxopt.org/), or Matlab (http://cvxr.com/cvx/).

9 Convergence in Distribution in Metric Spaces and Uniform CLTs

Question 9.1: Let $\mathcal{F}$ be the collection of cumulative distribution functions on the real line, and let $\|F - G\|_\infty = \sup_t |F(t) - G(t)|$ be the usual sup-norm on $\mathcal{F}$. Recall that a functional $\gamma : \mathcal{F} \to \mathbb{R}$ is continuous in the sup-norm at F if for all $\epsilon > 0$ there exists $\delta > 0$ such that $\|G - F\|_\infty \le \delta$ implies $|\gamma(F) - \gamma(G)| \le \epsilon$.

(a) Let F_n be the empirical distribution of an i.i.d. sample $X_1, \ldots, X_n$ drawn from distribution with CDF F. Show that if γ is continuous in the sup-norm, then

$$\gamma(F_n) \xrightarrow{p} \gamma(F).$$

(b) Which of the following functionals are sup-norm continuous? Prove or give a counterexample.

(i) The mean functional $F \mapsto \int x dF(x)$.

(ii) The Cramér-von Mises functional $F \mapsto \int (F(x) - F_0(x_0))^2 dF_0(x)$.

(iii) The quantile functional $Q_p(F) := \inf\{t \in \mathbb{R} \mid F(t) \ge p\}$.

Question 9.2: We consider estimation of median-like quantities in dimension $d \ge 1$. Let $\|\cdot\|_2$ denote the typical ℓ_2-norm, defined by $\|x\|_2^2 = \sum_{j=1}^d x_j^2$, and consider the loss function

$$\ell_\theta(x) := \|x - \theta\|_2$$

and risk $R(\theta) := \mathbb{E}[\ell_\theta(X)]$ for $X \sim P$. We will consider the asymptotics of the minimizer

$$\widehat{\theta}_n := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell_\theta(X_i).$$

We assume that $\mathbb{E}[\|X\|_2^2] < \infty$ for simplicity, though this is not strictly necessary. In this exercise, you may use the following facts (see, for example, the paper of Bertsekas [3]): if $\ell_\theta(x)$ is convex in θ for all x and for P-almost every x is differentiable in a neighborhood of a point θ_0 with derivative $\dot{\ell}_\theta(x) = \nabla_\theta \ell_\theta(x)$, then

$$\nabla R(\theta) = \mathbb{E}[\dot{\ell}_\theta(X)].$$

Similarly, if the Hessian $\ddot{\ell}_\theta = \nabla_\theta^2 \ell_\theta$ exists with P-probability 1 near θ_0, then $\nabla^2 R(\theta) = \mathbb{E}[\ddot{\ell}_\theta(X)]$.

(a) Show that the set $\operatorname{argmin}_\theta R(\theta) = \{\theta_0 \in \mathbb{R}^d \mid R(\theta_0) \le \inf_\theta R(\theta)\}$ is non-empty.

For the remainder of the question, we will assume that P has a density $f(x)$ for x in a neighborhood (i.e. some ball in $\mathbb{R}^d$) of a point $\theta_0 \in \operatorname{argmin}_\theta R(\theta)$.

(b) Show that θ_0 is unique. *Hint:* Question 7.13.(c) may be useful, as ℓ_θ is convex in θ.

(c) Give an asymptotic expansion of $\widehat{\theta}_n$, that is, show that

$$\widehat{\theta}_n - \theta_0 = \frac{1}{n} \sum_{i=1}^n \psi(X_i) + o_P(n^{-\frac{1}{2}}),$$

and specify the functions $\psi : \mathbb{R}^d \to \mathbb{R}^d$.

(d) What is the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$?

(e) Suppose that the vectors X_i are i.i.d. $\mathsf{N}(0, I)$, Gaussian with identity covariance I and the dimension $d \geq 3$. Show that $\theta_0 = 0$ and that $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathsf{N}(0, c_d I)$, where c_d is a constant that you should specify.

(f) Compare the asymptotic distribution of $\|\widehat{\theta}_n\|_2^2$ to that of $\left\|\overline{X}_n\right\|_2^2$, the sample mean, when $X_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I)$. What sample size $m(n)$ is required (as a function of n) for $\widehat{\theta}_{m(n)}$ to have the same asymptotic performance as $\overline{X}_n$?

Question 9.3 (Elliptical classes are Donsker)**:** Recall that a collection $\mathcal{F}$ of functions is P-Donsker if the process $\mathbb{G}_n := \sqrt{n}(P_n - P)$, viewed as a mapping $\mathbb{G}_n : \mathcal{F} \to \mathbb{R}$ via $\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - Pf)$, converges to a tight Gaussian process $\mathbb{G}$ in $L^\infty(\mathcal{F})$. For this, it is sufficient (by our arguments in class and asymptotic stochastic equi-continuity) that $\mathcal{F}$ be totally bounded for the $L^2(P)$ metric, and for the localized class

$$\mathcal{F}_\delta := \left\{ (f - g) \mid f, g \in \mathcal{F}, \|f - g\|_{L^2(P)} \leq \delta \right\},$$

where we recall $\|f - g\|_{L^2(P)}^2 = P(f - g)^2$, we have

$$\lim_{\delta \downarrow 0} \limsup_n \mathbb{E}\left[\sup_{(f-g) \in \mathcal{F}_\delta} \mathbb{G}_n(f - g) \right] = 0$$

and that each $f \in \mathcal{F}$ has a second moment under P so that finite-dimensional convergence holds.

Let $\{\varphi_i\}_{i \in \mathbb{N}}$ be a collection of functions $\varphi_i : \mathcal{X} \to \mathbb{R}$ with $P\varphi_i \varphi_j = 0$ for all $i \neq j$ and $\sum_{i=1}^\infty P\varphi_i^2 < \infty$. Define the elliptical class of functions

$$\mathcal{F} := \left\{ \sum_{i=1}^\infty c_i \varphi_i \mid \sum_{i=1}^\infty c_i^2 \leq 1 \text{ and the series converges pointwise} \right\}.$$

We will show that $\mathcal{F}$ is P-Donsker.

(a) Show that $\sum_{i=1}^\infty c_i \varphi_i$ converges in $L^2(P)$.

(b) Show that for any $f \in \mathcal{F}$ and $\epsilon > 0$, there exists some $m \in \mathbb{N}$, $\alpha \in \mathbb{R}^m$, and $g = \sum_{i=1}^m \alpha_i \varphi_i$ such that

$$P(f - g)^2 \leq \epsilon^2.$$

Argue that $\mathcal{F}$ is totally bounded for $L^2(P)$.

(c) Show that for any pair $f, g \in \mathcal{F}$, there exists a numerical constant $C < \infty$ such that for all $k \in \mathbb{N}$,

$$|\mathbb{G}_n(f) - \mathbb{G}_n(g)|^2 \leq C \left[P(f - g)^2 \sum_{i=1}^k \frac{\mathbb{G}_n^2(\varphi_i)}{P\varphi_i^2} + \sum_{i=k+1}^\infty \mathbb{G}_n^2(\varphi_i) \right].$$

(d) Argue that for any $\epsilon > 0$, we can choose $\delta > 0$ such that

$$\mathbb{E}\left[\sup_{(f-g) \in \mathcal{F}_\delta} \mathbb{G}_n^2(f - g) \right] \leq \epsilon,$$

whence the ellitical class $\mathcal{F}$ is Donsker.

Question 9.4: Consider a multiple hypothesis testing problem, where we observe a number n of statistics of interest, and wish to test whether and how many of them are significant. One approach to this is to use a Kolmogorov-Smirnov-type test (or an Anderson-Darling test [1]), which proceeds as follows: under the null H_0, the p-values of each of our statistics are independent and uniform; call them $U_1, U_2, \ldots, U_n$. Then for empirical CDF $F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{U_i \le t\}$, we can consider

$$K_n := \sup_{t\in[0,1]} \sqrt{n}\,|F_n(t) - t|\,.$$

Then we might hope that if there is a departure from the (known) limit distribution, we would detect it. In this problem, we consider more nuanced departures from the null

$$H_0 : U_i \overset{\text{iid}}{\sim} \mathsf{Uni}([0,1]).$$

In particular, we consider a situation where a few of the p-values may be drawn from an alternative distribution Q_n instead of the null $\mathsf{Uni}([0,1])$, but this may be a very small number. As motivation, one might consider testing for contamination in a part of a city's water supply: one does not know where to look for the contamination, so that one needs to perform tests of many individuals, but if enough have elevated levels of some contaminant, one knows to do a more careful investigation.

To that end, consider the following sequence of alternative distributions indexed by n:

$$H_{1,n} : U_i \overset{\text{iid}}{\sim} (1-\varepsilon_n)\mathsf{Uni}([0,1]) + \varepsilon_n Q_n,$$

that is, with probability $(1-\varepsilon_n)$ the U_i are drawn from $\mathsf{Uni}([0,1])$ as in the null H_0 and otherwise $U_i \sim Q_n$, where Q_n is some other distribution. We shall assume throughout that

$$\varepsilon_n = n^{-\beta}$$

for some $\beta \in (\frac{1}{2}, 1)$, so that the number of observations off the null is much smaller than the typical $1/\sqrt{n}$ scaling one needs for central limit theorems.

(a) Argue that for $F(t) = t$, the empirical process

$$\sqrt{n}(F_n(\cdot) - F(\cdot))$$

(indexed by $t \in [0,1]$) has the same limit under both the i.i.d. null H_0 and the alternative $H_{1,n}$.

(b) Show that the KS statistic K_n has the same limit distribution under both H_0 and $H_{1,n}$.

Hint. You should not need to check any asymptotic stochastic equicontinuity or tightness of the empirical process to do (a)–(b).

We now develop an example to show it may be possible to detect small contaminants, though we cannot necessarily reject individual null hypotheses. To that end, consider the hypotheses

$$H_0 : X_i \overset{\text{iid}}{\sim} \mathsf{N}(0,1) \quad \text{and} \quad H_{1,n} : X_i \overset{\text{iid}}{\sim} (1-\varepsilon_n)\mathsf{N}(0,1) + \varepsilon_n \mathsf{N}(\mu_n, 1),$$

where the mean $\mu_n = \sqrt{2r\log n}$ for some $r < 1$.[5] Define the Bernoulli random variables

$$B_i^n := \mathbf{1}\{X_i \ge \mu_n\}$$

for all $i \le n$ and $n \in \mathbb{N}$. For simplicity, assume that $\frac{1}{2} < \beta < r < 1$.

[5] To obtain p-values, simply invert the Gaussian CDF. Our choice of $r < 1$ prevents a naive test that simply looks at the maximum of the X_i to distinguish H_0 and $H_{1,n}$, as even under H_0 we would expect $\max_i X_i \approx \sqrt{2\log n}$.

(c) Develop a test statistic T_n based on $\sum_{i=1}^n B_i^n$ and threshold t_n such that

$$P_0(T_n \geq t_n) \to 0 \quad \text{and} \quad P_{1,n}(T_n \geq t_n) \to 1 \tag{9.1}$$

as $n \to \infty$, where $P_{1,n}$ denotes sampling under the alternative $H_{1,n}$. Prove that your statistic satisfies the desiderata (9.1). *Hint.* There are many ways to do this, including via concentration inequalities, Chebyshev inequalities, or multiplicative Hoeffding bounds. Your value t_n may not need to depend on n, but it's fine if it does.

As a piece of fun culture, there are tests that can distinguish H_0 and $H_{1,n}$ with high probability, while also being adaptive to the contamination rate ε_n and distribution Q_n. Deriving this is well beyond the scope of this course, but one prominent example is Donoho and Jin's "Higher Criticism," which for a given level α (typically $\alpha = .05$) uses

$$\mathrm{HC}_\alpha := \sup_{t \in [0,\alpha]} \frac{\sqrt{n}(F_n(t) - t)}{\sqrt{t(1-t)}}.$$

The statistic actually diverges even under the null H_0 of i.i.d. uniform sampling, but not by much: the correction $\mathrm{HC}_\alpha / \sqrt{2 \log \log n} \xrightarrow{p} 1$ as $n \to \infty$. Thus, while we may not be able to reject the null in a classical $p \leq .05$ Fisherian sense, we can recognize "something is funny."

10 Contiguity and Quadratic Mean Differentiability

Question 10.1: Let P_n and Q_n be sequences of probability measures with $\|P_n - Q_n\|_{\text{TV}} \to 0$. Show that P_n and Q_n are mutually contiguous.

Question 10.2: Recall that a family $\{P_\theta\}_{\theta \in \Theta}$ of distributions on $\mathcal{X}$ is *quadratic mean differentiable (QMD)* at $\theta \in \mathbb{R}^d$ if there exists a score function $\dot{\ell}_\theta : \mathcal{X} \to \mathbb{R}^d$ such that

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^\top \dot{\ell}_\theta \sqrt{p_\theta} \right)^2 d\mu = o(\|h\|^2).$$

Let P^n denote the n-fold product of P (i.e. n i.i.d. observations from P).

(a) Show that

$$\lim_{n \to \infty} d_{\text{hel}}^2(P_{\theta_0}^n, P_{\theta_0 + h/\sqrt{n}}^n) = 1 - \exp\left(-\frac{1}{8} h^\top I_{\theta_0} h \right).$$

(b) Give conditions on h (and prove them) such that for any sequence of tests $\psi_n : \mathcal{X} \to \{0, 1\}$, we have the asymptotically non-negligible error guarantee that

$$\liminf_n \left\{ P_{\theta_0}^n(\psi_n \neq 0) + P_{\theta_0 + h/\sqrt{n}}^n(\psi_n \neq 1) \right\} > 0.$$

Question 10.3: Let P_θ denote the uniform distribution on $[0, \theta]$, defined whenever $\theta > 0$. Let $\theta > 0$ and consider the "local" alternatives $P_{\theta + h/\sqrt{n}}$, where $h \in \mathbb{R}$. Letting $\psi_n : \mathbb{R} \to \{0, 1\}$ be a sequence of tests, give upper and lower bounds on the limit infimum

$$\liminf_{n \to \infty} \inf_{\psi_n} \left\{ P_{\theta + h/\sqrt{n}}^n(\psi_n \neq 1) + P_\theta^n(\psi_n \neq 0) \right\}.$$

Explain your result in the light of Question 10.2.

Question 10.4 (Extending Lemma 7.6 of van der Vaart [7])**:** Let $\Theta \subset \mathbb{R}^k$ be open and p_θ be a μ-probability density on $\mathcal{X}$. Assume that $\theta \mapsto s_\theta(x) := \sqrt{p_\theta(x)}$ is absolutely continuous for all x, and that for each θ, we have

$$\mu(\{x \in \mathcal{X} : \dot{p}_\theta(x) \text{ fails to exist}\}) = 0.$$

Assume additionally that the elements of $I_\theta := \int (\dot{p}_\theta / p_\theta)(\dot{p}_\theta / p_\theta)^T p_\theta d\mu$ are continuous in θ. Prove that the map $\theta \mapsto \sqrt{p_\theta}$ is differentiable in quadratic mean with $\dot{\ell}_\theta = \dot{p}_\theta / p_\theta$, that is,

$$\int \left(\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2} h^T \dot{\ell}_\theta(x) \sqrt{p_\theta(x)} \right)^2 d\mu(x) = o(\|h\|^2) \text{ as } h \to 0.$$

Question 10.5 (Non-parametrics and Hellinger Divergences)**:** Recall that the Hellinger distance between distributions P and Q is

$$d_{\text{hel}}^2(P, Q) = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 d\mu,$$

where $p = dP/\mu$, $q = dQ/d\mu$, and μ is any measure dominating P and Q. Let P_0 be a probability distribution on a measurable space $\mathcal{X}$.

(a) For a bounded function g with mean zero under P_0, that is, such that $P_0 g = 0$ and $\|g\|_\infty = \sup_{x \in \mathcal{X}} |g(x)| < \infty$, define the tilted distributions

$$dP_t(x) = (1 + tg(x))dP_0(x),$$

which are evidently valid distributions whenever $t \leq \frac{1}{\|g\|_\infty}$. Show that as $t \downarrow 0$,

$$d_{\rm hel}^2(P_t, P_0) = \frac{1}{8} t^2 P_0 g^2 + O(t^3).$$

Hint: the expansion $\sqrt{1+a} = 1 + \frac{a}{2} - \frac{a^2}{8} \pm |a|^3$, valid for $|a| \leq \frac{1}{2}$, may be useful.

We now give a more advanced variant of this expansion, including (roughly) a derivative of $\sqrt{dP_t}$ in $L^2(P_0)$, showing that this mapping is quadratic mean differentiable (though the particular concept is not necessary for this exercise). Let $\phi : \mathbb{R} \to \mathbb{R}_+$ be bounded, continuous, continuously differentiable in a neighborhood of 0 and satisfy $\phi(0) = \phi'(0) = 1$. (For example, the functions $\phi(t) = \frac{2}{1+e^{-2t}}$ and $\phi(t) = \min\{(1+t)_+, 2\}$ satisfy these conditions.) Let $g \in L^2(P_0)$, that is, $P_0 g^2 < \infty$ and assume $P_0 g = 0$. Define

$$dP_t(x) = \frac{1}{C(t)} \phi(tg(x)) dP_0(x), \quad \text{where } C(t) := \int \phi(tg(x)) dP_0(x).$$

(b) Show that $C(t) = 1 + o(t)$ as $t \downarrow 0$.

(c) Show that as $t \downarrow 0$, for all x we have

$$h_t(x) := \frac{1}{t^2} \left(\sqrt{\phi(tg(x))/C(t)} - 1 - \frac{1}{2} tg(x) \right)^2 \to 0.$$

(d) Give a dominating function for $h_t(x)$ and argue that $\lim_{t \downarrow 0} \int h_t(x) dP_0(x) = 0$, and so

$$\int \left(\sqrt{dP_t} - \sqrt{dP_0} - \frac{1}{2} tg \sqrt{dP_0} \right)^2 = o(t^2).$$

(e) Use the preceding result to show that

$$d_{\rm hel}^2(P_t, P_0) = \frac{1}{8} t^2 P_0 g^2 + o(t^2).$$

11 Local Asymptotic Normality, Efficiency, and Minimaxity

Question 11.1: Let P_0 and P_1 be arbitrary distributions. Show Le Cam's first lemma, that

$$\inf_T \{P_0(T \neq 0) + P_1(T \neq 1)\} = 1 - \|P_0 - P_1\|_{\rm TV},$$

where the infimum is taken over all tests $T : \mathcal{X} \to \{0, 1\}$.

Question 11.2 (Estimating a non-differentiable function)**:** Let $\mathcal{P}$ be a collection of distributions on a space $\mathcal{X}$ and $\psi : \mathcal{P} \to \mathbb{R}$ be a functional of interest that we wish to estimate. Let $\{P_t\}_{t \geq 0} \subset \mathcal{P}$ be a sub-model of $\mathcal{P}$, and assume that it is a QMD sub-model in the sense that there is a score $g : \mathcal{X} \to \mathbb{R}$, $P_0 g = 0$, $P_0 g^2 < \infty$ with

$$\int \left(\sqrt{dP_t} - \sqrt{dP_0} - \frac{1}{2} t g \sqrt{dP_0}\right)^2 = o(t)^2.$$

We illustrate some of the difficulties in estimation of $\psi(P)$ along a path $t \mapsto P_t$ for which $\psi(P_t)$ is *not* differentiable. For simplicity, we assume that $\psi(P_0) = 0$ (this is no loss of generality) and that the path is such that

$$\lim_{h \downarrow 0} \sup_{t \in [0,h]} \frac{|\psi(P_t) - \psi(P_0)|}{h} = +\infty.$$

Though the results will hold under this assumption, you may instead assume that the limit is actually infinite:

$$\lim_{t \downarrow 0} \frac{|\psi(P_t) - \psi(P_0)|}{t} = +\infty.$$

(a) Letting P^n denote the n-fold product of P, give the limit $\lim_{n \to \infty} d_{\rm hel}^2(P^n_{t/\sqrt{n}}, P_0^n)$, where $t \in \mathbb{R}_+$.

(b) Show that for any t, we have

$$\inf_T \{P_t^n(T \neq 0) + P_0^n(T \neq 1)\} \geq 1 - d_{\rm hel}(P_t^n, P_0^n)\sqrt{2 - d_{\rm hel}^2(P_t^n, P_0^n)}.$$

(c) Show that for any sequence $\epsilon_n \downarrow 0$, we have

$$\inf_{0 \leq t \leq \epsilon_n/\sqrt{n}} \inf_T \{P_t^n(T \neq 0) + P_0^n(T \neq 1)\} \to 1.$$

(d) Show that we have the local minimax lower bound

$$\lim_{\epsilon \downarrow 0} \liminf_{K \to \infty} \liminf_{n \to \infty} \sup_{0 \leq t \leq \frac{\epsilon}{\sqrt{n}}} \inf_{\widehat{\psi}_n} \max_{s \in \{0,t\}} P_s^n \left(|\widehat{\psi}_n - \psi(P_s)| \geq \frac{K}{\sqrt{n}}\right) \geq \frac{1}{2}.$$

(e) Give a one-sentence description of this result.

Question 11.3 (Score and influence functions for regression)**:** Consider the prediction problem of finding θ to best predict a scalar y from $x \in \mathbb{R}^d$ via the model $\widehat{y} = \theta^\top x$. We study the local asymptotic minimax risk for estimation of the parameter

$$\theta(P) := \operatorname*{argmin}_\theta \mathbb{E}_P[(Y - X^T\theta)^2],$$

where $(X, Y) \sim P$, but the standard linear regression model need not hold. You may assume that $\mathbb{E}[\|X\|^4] < \infty$ and $\mathbb{E}[Y^2 \|X\|^2] < \infty$ for simplicity.

(a) What is $\theta(P)$?

(b) Let the function $\phi(t) = \min\{2, \max\{1+t, 0\}\}$, and let $g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ satisfy $Pg = 0$ and $Pg^2 < \infty$. For $t \geq 0$, define

$$dP_t(x, y) = c(t)\phi(tg(x, y))dP(x, y),$$

where $c(t)$ is a normalizing function. Show that as $t \downarrow 0$,

$$\int \left(\sqrt{dP_t} - \sqrt{dP} - \frac{1}{2}tg\sqrt{dP}\right)^2 = o(t^2).$$

(c) Give the limit

$$\lim_{t \to 0} \frac{\theta(P_t) - \theta(P)}{t}.$$

(d) Let $\mathcal{G}$ be the collection of functions $g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, $Pg^2 < \infty$, and $Pg = 0$. Let $L : \mathbb{R}^d \to \mathbb{R}_+$ be a symmetric quasi-convex, bowl-shaped and Lipschitz loss. For functions $g_1, \ldots, g_k \in \mathcal{G}$ and $h \in \mathbb{R}^k$, define the distributions

$$dP_h(x, y) \propto \phi(h^\top g(x, y))dP(x, y),$$

normalized appropriately, where $g(x, y) = [g_1(x, y) \; \cdots \; g_k(x, y)]^\top$. Let $\theta_h = \theta(P_h)$ for shorthand. Construct an influence function for the parameter θ_h, that is, a function $\psi^{\text{Inf}} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ such that

$$\theta(P_h) - \theta(P) = \mathbb{E}_P\left[\psi^{\text{Inf}}(X, Y)g(X, Y)^\top\right] h + o(\|h\|).$$

(You may do this for g mapping into $\mathbb{R}$ rather than $\mathbb{R}^k$.)

(e) Let $\pi_{n,c,k}$ be a uniform distribution on $\{h \in \mathbb{R}^k \mid \|h\| \leq c/\sqrt{n}\}$. Let $\theta_h = \theta(P_h)$ for shorthand. Give a (tight) lower bound on

$$\sup_{k \in \mathbb{N}, g_1, \ldots, g_k \in \mathcal{G}} \liminf_{c \to \infty} \liminf_n \inf_{\widehat{\theta}_n} \int \mathbb{E}_{P_h^n}\left[L(\sqrt{n}(\widehat{\theta}_n - \theta_h))\right] d\pi_{n,c}(h).$$

What does your lower bound become when the model $y = x^\top\theta + \varepsilon$ holds, where ε is a mean-zero independent noise with $P\varepsilon^2 = \sigma^2$?

Question 11.4 (Anderson's Lemma)**:** In this question, we derive a more general version of Anderson's lemma from the Prékopa-Leindler (PL) inequality, a deep result in functional analysis and convex geometry (see, e.g. Ball [2] or Gardner [5] for discussions of the inequality and attendant results). The PL inequality is as follows. Suppose that for some $\lambda \in [0, 1]$, functions $f, g, h : \mathbb{R}^n \to \mathbb{R}_+$ satisfy

$$h((1-\lambda)x + \lambda y) \geq f(x)^{1-\lambda}g(y)^\lambda \;\; \text{for all } x, y \in \mathbb{R}^n.$$

Then

$$\int h(x)dx \geq \left(\int f(x)dx\right)^{1-\lambda}\left(\int g(x)dx\right)^\lambda. \tag{11.1}$$

Consequences of this inequality include the Brunn-Minkowski inequality, which Anderson first used to prove his eponymous inequality. We will give an alternative approach using inequality (11.1).

A function $f : \mathbb{R}^n \to \mathbb{R}_+$ is *log concave* if $\log f$ is concave, meaning that

$$f(\lambda x + (1-\lambda)y) \geq f(x)^\lambda f(y)^{1-\lambda}$$

whenever $\lambda \in [0,1]$ and $x, y \in \operatorname{dom} f$. (We treat $f(x) = 0$ when $x \notin \operatorname{dom} f$, so that $\log f(x) = -\infty$, and assume $\operatorname{dom} f$ is a convex set.)

(a) Let f, g be log-concave. Show that the convolution

$$c(x) = (f * g)(x) := \int f(x-y)g(y)dy$$

is log-concave in x.

(b) Let C be a convex set and $g(x) = \mathbf{1}\{x \in C\}$ (i.e. $g(x) = 1$ if $x \in C$ and 0 otherwise). Show that g is log-concave.

(c) Let f be a log-concave density, meaning $\int f(x)dx = 1$, and let X be a random variable with density f. Show that the function $h(v) = \mathbb{P}(X + v \in C)$ is log concave for any convex set C.

With these preliminaries, we now come to Anderson's lemma. Recall that a function $L : \mathbb{R}^k \to \mathbb{R}_+$ is quasiconvex if the sublevel sets $\{x \in \mathbb{R}^k : L(x) \leq t\}$ are convex for each $t \in \mathbb{R}$, and that L is symmetric if $L(x) = L(-x)$ for each x. The basic Anderson's lemma presented in the text is the following:

Lemma 11.4.1. *Let $X \sim \mathsf{N}(0, \Sigma)$ and $L : \mathbb{R}^k \to \mathbb{R}_+$ be quasiconvex and symmetric. Then for any matrix A,*

$$\inf_v \mathbb{E}[L(AX - v)] = \mathbb{E}[L(AX)].$$

The extension that we will prove is the following:

Lemma 11.4.2 (Anderson's lemma)**.** *Let $X \in \mathbb{R}^d$ be a random vector with a symmetric log-concave density f, and let $L : \mathbb{R}^k \to \mathbb{R}_+$ be quasiconvex and symmetric. Then*

$$\inf_v \mathbb{E}[L(AX - v)] = \mathbb{E}[L(AX)]$$

for any matrix $A \in \mathbb{R}^{k \times d}$.

(d) Prove Lemma 11.4.2.

Question 11.5 (The influence function in stochastic optimization)**:** In this problem, we develop the (nonparametric) influence function for stochastic optimization, or M-estimation, problems, assuming sufficient smoothness and convexity.

We will develop the idea via the implicit function theorem, a variant of which follows. To state the theorem, let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be a $\mathcal{C}^1$ function in a neighborhood of the point $(u_0, v_0) \in \mathbb{R}^n \times \mathbb{R}^m$, where $f(u_0, v_0) = 0$. Let $D_{\mathsf{u}} f(u, v) \in \mathbb{R}^{m \times n}$ be the Jacobian (derivative matrix) of f with respect to u (i.e. its first n coordinates) and $D_{\mathsf{v}} f(u, v) \in \mathbb{R}^{m \times m}$ that with respect to v (i.e. its last m coordinates), so that

$$f(u + \Delta_u, v + \Delta_v) = f(u, v) + D_{\mathsf{u}} f(u, v)\Delta_u + D_{\mathsf{v}} f(u, v)\Delta_v + o(\|\Delta_u\| + \|\Delta_v\|).$$

We then have the implicit function theorem.

Theorem 11.5.1 (Implicit function theorem)**.** *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ satisfy the above conditions. Then there exists an open neighborhood $U \subset \mathbb{R}^n$ of u_0 and a $\mathcal{C}^1$ function $h : U \to \mathbb{R}^m$ such that $h(u_0) = v_0$, and for all $u \in U$ we have both $f(u, h(u)) = 0$ and*

$$\dot{h}(u) = -(D_{\mathsf{v}} f(u, h(u)))^{-1} D_{\mathsf{u}} f(u, h(u)) \in \mathbb{R}^{m \times n}.$$

We use Theorem 11.5.1 to develop the influence function of M-estimators. Let $\mathcal{P}$ be a family of distributions on $\mathcal{X}$ and $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$ a loss function convex in its first argument, where $\Theta \subset \mathbb{R}^d$ is an open convex set. Define the population losses $L_P(\theta) := P\ell(\theta, X)$ and let

$$\theta(P) := \operatorname*{argmin}_{\theta \in \Theta} \{L_P(\theta) = P\ell(\theta, X)\}$$

be the parameter of interest. Fix $P_0 \in \mathcal{P}$, and for a bounded function $g : \mathcal{X} \to \mathbb{R}$ with $P_0 g = 0$, define the tilted distributions P_t by

$$dP_t(x) := (1 + tg(x))dP_0(x),$$

as in, e.g., Question 10.5 or [7, Example 25.16]. Assume that $\nabla \ell(\theta, x)$ and $\nabla^2 \ell(\theta, x)$ are $M_1(x)$- and $M_2(x)$-Lipschitz in θ, respectively, meaning that (for a norm $\|\cdot\|$ whose choice is immaterial)

$$\begin{aligned} \left\|\nabla \ell(\theta', x) - \nabla \ell(\theta, x)\right\| &\le M_1(x) \left\|\theta - \theta'\right\| \\ \left\|\nabla^2 \ell(\theta', x) - \nabla^2 \ell(\theta, x)\right\| &\le M_2(x) \left\|\theta - \theta'\right\|, \quad \text{all } \theta, \theta', \end{aligned}$$

where M_1 and M_2 are P_0-integrable, and that the Hessian $\nabla^2 L_{P_0}(\theta(P_0)) \succ 0$, is positive definite. Assume also that the objective ℓ is locally Lipschitz around $\theta_0 = \theta(P_0)$, or, what is simpler and sufficient, that $M_0(x)^2 := \|\nabla \ell(\theta_0, x)\|^2$ is P_0-integrable. Construct the influence function $\dot{\theta}_0 : \mathcal{X} \to \mathbb{R}^d$ of $\theta(\cdot)$, that is, give a square-integrable function $\dot{\theta}_0$ such that

$$\lim_{t \downarrow 0} \frac{\theta(P_t) - \theta(P_0)}{t} = P_0 \dot{\theta}_0 g = \int \dot{\theta}_0(x) g(x) dP_0(x).$$

An interpretation of this is that we may view $\theta(P_t)$ (asymptotically) as a linear function of the parameter t in the model family $\{P_t\}_{t \in \mathbb{R}}$.

You may assume that integrals and derivatives may be exchanged without comment (they can! We just don't require a proof). *Hint.* The minimizers $\theta(P)$ satisfy $\nabla L_P(\theta(P)) = \dot{L}_P(\theta(P)) = 0$.

Question 11.6 (Asymptotic efficiency with different losses in regression)**:** Consider data generated according to

$$Y_i = \langle X_i, \theta \rangle + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} Q, \quad X_i \stackrel{\text{iid}}{\sim} \mu, \tag{11.2}$$

where Q has a continuous density q with respect to Lebesgue measure, $t \mapsto \sqrt{q(t)}$ is absolutely continuous, and $q(t), \dot{q}(t) \to 0$ as $t \to \pm\infty$. Assume also that for $a(t) = \log q(t)$, the Fisher information $J_q = Q\dot{a}^2 = Q(\dot{q}/q)^2$ for location under q exists. You should *not* need to assume that q has any monotonicity properties. Assume also that $\mathbb{E}_\mu X X^T = \Sigma \succ 0$.

(a) Give the Fisher information for θ in this model and argue that the family is QMD.

(b) Give the local asymptotic minimax lower bound for estimation in the family (11.2).

Let P_θ denote the joint distribution over (X, Y) in model (11.2). Now we consider the question of estimating θ without knowing the distribution Q of the noise using different loss functions. In particular, for different symmetric convex functions $f : \mathbb{R} \to \mathbb{R}_+$, we consider M-estimators of the form

$$\widehat{\theta}_n := \operatorname*{argmin}_\theta F_n(\theta) = P_n f(Y - X^T \theta).$$

(c) Let $f(t) = |t|$ and assume that the density $q(0) > 0$. Give an asymptotically linear expansion of $\widehat{\theta}_n$, that is, write $\widehat{\theta}_n - \theta = P_n Z + o_P(1/\sqrt{n})$ and specify the random vectors Z. Give the asymptotic distribution of $\widehat{\theta}_n$ under local alternatives in model (11.2), that is, under $P_{\theta + h_n/\sqrt{n}}$ for $h_n \to h \in \mathbb{R}^d$.

(d) Give a density q for which the absolute loss in part (c) is arbitrarily inefficient.

(e) Let $f_u(t)$ be the Huber loss with threshold $u > 0$, that is,

$$f_u(t) = \begin{cases} \frac{1}{2u} t^2 & \text{if } |t| \le u \\ |t| - u/2 & \text{if } |t| > u \end{cases}.$$

Repeat the same analysis, *mutatis mutandis*, as in part (c).

(f) Give a loss function $f : \mathbb{R} \to \mathbb{R}_+$ such that whenever the preceding conditions on q hold, we have the locally uniform convergence

$$\sqrt{n}\left(\widehat{\theta}_n - (\theta + h_n/\sqrt{n})\right) \underset{P_{\theta + h_n/\sqrt{n}}}{\xrightarrow{d}} \mathsf{N}(0, \sigma^2(q)\Sigma^{-1}),$$

where $\sigma^2(q) < \infty$ may depend on q and the loss.

Question 11.7: Consider the setting of Question 11.6. The results of the question guarantee that there exists a $\sqrt{n}$-consistent estimator. Give sufficient conditions on the density q of the noise that you may construct one-step estimator δ_n, of the form in Question 3.2, which is asymptotically optimal and regular. That is, give an explicit one-step estimator δ_n such that

$$\sqrt{n}(\delta_n - \theta_0) = \frac{1}{n} \sum_{i=1}^n I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X_i, Y_i) + o_{P_0}(1),$$

where $\dot{\ell}$ is the score function in the model (11.2). Conclude that we have the locally uniform convergence

$$\sqrt{n}(\delta_n - (\theta_0 + h_n/\sqrt{n})) \underset{\theta_0 + h_n/\sqrt{n}}{\xrightarrow{d}} \mathsf{N}(0, I_{\theta_0}^{-1})$$

under local alternatives $h_n \to h$.

References

[1] T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23(2):193–212, 1952.

[2] K. Ball. An elementary introduction to modern convex geometry. In S. Levy, editor, *Flavors of Geometry*, pages 1–58. MSRI Publications, 1997.

[3] D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

[4] L. D. Brown and M. G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *Annals of Statistics*, 24(6):2524–2535, 1996.

[5] R. J. Gardner. The Brunn-Minkowski inequality. *Bulletin of the American Mathematical Society*, 39(3):355–405, 2002.

[6] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[7] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

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

[9] R. Vershynin. High dimensional probability: An introduction with applications in data science. 2019.