Exercises for Theory of Statistics (Stats300b)

John Duchi
Winter Quarter 2019

Contents

1 Background 2
2 Convex Analysis and Statistics 6
3 Asymptotic Efficiency 10
4 U- and V-statistics 12
5 Testing 16
6 Concentration inequalities 18
7 Uniform laws of large numbers and related problems 21
8 High-dimensional problems 26
9 Convergence in Distribution in Metric Spaces and Uniform CLTs 29
10 Contiguity and Quadratic Mean Differentiability 31
11 Local Asymptotic Normality, Efficiency, and Minimaxity 32
1 Background

**Question 1.1:** Let \( p_n \) and \( q_n \) be (a sequence of) densities with respect to some base measure \( \mu \). 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 \( \epsilon \in (0, 1/2) \), we have

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

Here we define \( G^{-1}(\alpha) = \inf\{x \in \mathbb{R} : G(x) \geq \alpha\} \) 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 \( F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq t\} \). 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. \( \text{Beta}(\theta, 1) \) for some \( \theta > 0 \). Letting \( \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \) denote the sample mean and \( \hat{\theta}_n = \frac{\overline{X}_n}{1-\overline{X}_n} \), give the limiting distribution of the sequence

\[
\sqrt{n} \left( \hat{\theta}_n - \theta \right)
\]

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 \( X_i^n \overset{iid}{\sim} \text{Bernoulli}(\theta_n) \) and \( \theta_n = 1/\sqrt{n} \). Define \( \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i^n \). Is \( n^{3/4}(\hat{\theta}_n - \theta_n) \) 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 \( \sigma^2 \)-sub-Gaussian if \( \mathbb{E}[\exp(\lambda X)] \leq \exp(\frac{\lambda^2 \sigma^2}{2}) \) for all \( \lambda \in \mathbb{R} \).

(a) Show that if \( Z \) is mean-zero Gaussian with variance \( \sigma^2 \), 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 \( \sigma^2 \)-sub-Gaussian random variables, then \( \mathbb{E}[\max_{i \leq n} X_i] \leq \sqrt{2\sigma^2 \log n} \).
Question 1.8: Let $\|\cdot\|_{TV}$ be the total variation distance, that is,

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

for probability distributions $P$ and $Q$. Let $\mu$ 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 $\mu$. Show the following equalities, where $\wedge$ denotes min and $\vee$ denotes max.

(a) $2\|P - Q\|_{TV} = \int |p - q| d\mu$.
(b) $\sup_{\|f\|_\infty \leq 1} \int f(x)(dP(x) - dQ(x)) = 2\|P - Q\|_{TV}$.
(c) $2\|P - Q\|_{TV} = \int [p - q]_+ d\mu + \int [q - p]_+ d\mu$.
(d) $\|P - Q\|_{TV} = \int (p \vee q) d\mu - 1$.
(e) $\|P - Q\|_{TV} = 1 - \int (p \wedge q) d\mu$.

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

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

Show that

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

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 $\alpha^T K \alpha \geq 0$. 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 $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$ 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

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

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

\[
f(x) := \lim_{n \to \infty} f_n(x)
\]

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) exists.

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

\[
k(x, z) := \langle r_x, r_x \rangle
\]

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

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

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

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

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

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

It may be easier to show the stronger inequality

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

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

Let $\text{dom} f = \{ x : f(x) < +\infty \}$, 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 $y \geq x > x_0$, write $x = \lambda y + (1 - \lambda)x_0$]

(b) Define the left and right derivatives

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

Show that

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

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

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

We say that $f$ is strictly convex at the point $x$ if for all $x_0, x_1 \neq x$ 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 $\mathbb{E}[X] \in \text{dom} f$, we have $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$. 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 $\mu$ 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 $D_{\text{kl}}(P|Q) \geq 0$, 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 $p_\theta$ be an exponential family density (with respect to some base measure $\mu$ 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 $\Theta = \text{dom} A := \{ \theta \in \mathbb{R}^d : A(\theta) < \infty \}$, 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. [\textit{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 $\theta$, then the parameter $\theta$ is identifiable. [\textit{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 $V$ 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 $V$ is complete, meaning that Cauchy sequences in $V$ converge.

Let $C \subset V$ 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

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

(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 $\pi_C(x) := \arg\min_{y \in C} \{\|x - y\|^2\}$, the projection of $x$ onto $C$, which is characterized by

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

Draw a picture of your result.

**Question 2.5**: Let $\mathcal{X}$ be a measurable space and $X_i \overset{\text{iid}}{\sim} P$, 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 $\theta$. Let $\theta^* = \arg\min_{\theta \in \Theta} R(\theta)$ and assume that the Hessian $\nabla^2 R(\theta^*) \succ 0$, that is, the Hessian of the risk is positive definite at the point $\theta^*$, and assume that $\theta^* \in \text{int} \Theta$. 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 $\theta$, that is,

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

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

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

where $X_i \overset{\text{iid}}{\sim} P$, and $\hat{\theta}_n = \arg\min_{\theta \in \Theta} \hat{R}_n(\theta)$, then we have asymptotic normality of $\hat{\theta}_n$. 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 $\theta$.

(b) Using the above assumptions, show that $\widehat{\theta}_n \xrightarrow{P} \theta^*$. You may use the following result: if a function $f$ is convex and satisfies $\nabla^2 f(\theta) \geq \lambda I$ for all $\theta$ satisfying $\|\theta - \theta_0\| \leq c$, then
\[
f(\theta) \geq f(\theta_0) + \nabla f(\theta_0)^T (\theta - \theta_0) + \frac{\lambda}{2} \min \left\{ \|\theta_0 - \theta\|^2, c \|\theta_0 - \theta\| \right\}.
\]

(c) Assuming that $\widehat{\theta}_n \xrightarrow{P} \theta^*$, use a Taylor expansion to show that
\[
\sqrt{n}(\widehat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N} \left( 0, (\nabla^2 R(\theta^*))^{-1} \right)
\]
where $\Sigma = \Cov(\nabla \ell(\theta^*, X))$ is the covariance matrix of the gradient of the loss.

**Question 2.6** (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 $\frac{d^2}{dt^2} \log F(t) \leq 0$ for all $t \in (a, b)$.

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

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

(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 2.7** (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 $Y \subset \mathbb{R}^m$ the function $g(x) := \inf_{y \in Y} f(x, y)$ 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 2.8 (Subgradients):** Let \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) 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 \( x^* \) if and only if \( 0 \in \partial f(x^*) \).

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

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

(a) What is the asymptotic distribution of \( T_n \)?

(b) Suppose we know that \( X_i \overset{iid}{\sim} \mathcal{N}(\theta, 1) \) for some unknown \( \theta \). Letting \( \Phi(x) = P(Z \leq x) \) be the standard normal CDF, another possible estimator is \( G_n := \Phi(x - X_n) \) where \( X_n = \frac{1}{n} \sum_{i=1}^{n} X_i \).

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 \( \mathcal{N}(\theta, 1) \). Give a consistent estimator \( \hat{\theta}_n \) of \( \theta \) based on \( T_n \). What is the asymptotic distribution of your estimator? What is its efficiency relative to the mean \( \bar{X}_n \)?

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 \( X_i \overset{iid}{\sim} P_{\theta_0} \), where \( P_\theta \) has density \( p_\theta \) w.r.t. the measure \( \mu \) as usual. Assume that \( \ell_\theta(x) = \log p_\theta(x) \) is twice continuously differentiable in \( \theta \) 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 \( \hat{\theta}_n \) is a \( \sqrt{n} \)-consistent estimator, that is,

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = O_{P_{\theta_0}}(1).
\]

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

\[
\nabla L_n(\hat{\theta}_n) + \nabla^2 L_n(\hat{\theta}_n)(\delta_n - \hat{\theta}_n) = 0.
\]

(a) What is the asymptotic distribution of \( \delta_n \)?

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

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

(c) Let \( \delta_n \) 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

\[
T_n := \begin{cases} 
X_n & \text{if } |X_n| \geq n^{-1/4} \\
0 & \text{otherwise.}
\end{cases} \tag{2}
\]
In this question, you will simulate the Hodges estimator (2) to study its performance. Repeat the following experiment \(N = 500\) times. For \(n \in \{50, 100, 200, 400\}\), generate i.i.d. samples \(X_i \sim N(\theta_n, 1)\), \(i = 1, \ldots, n\), where you set \(\theta_n\) 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 \(\hat{\theta}_n = \bar{X}_n\), the sample mean.

(a) Generate three plots, one each for \(n = 50, 100, 200\), and plot the (sampled/simulated) mean squared error \(n \cdot E_h[(\hat{\theta}_n - \theta_n)^2]\) and \(n \cdot \mathbb{E}_h[(T_n - \theta_n)^2]\) 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 \(n \cdot \mathbb{E}_h[(\hat{\theta}_n - \theta_n)^2]\) and \(n \cdot E_h[(T_n - \theta_n)^2]\) 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 \(X_i \sim P_\theta\), 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 \(\mu\).

(a) Show that \(I_\theta^{(Q)} \preceq I_\theta\) in the positive semidefinite order, meaning that \(v^T I_\theta^{(Q)} v \leq v^T I_\theta v\) 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 \(X_i \sim \text{Bernoulli}(\theta)\), 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 \(I_\theta\) 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 \(\theta\) based on \(Y_1, \ldots, Y_n\). Is your estimator efficient?

(d) Give a situation in which such a procedure might be useful.
4 U- and V-statistics

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

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

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

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

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 $k$th smallest of the absolute values in the sample, $R_i^+ = \sum_{j=1}^n 1 \{|X_j| \leq |X_i|\}$. Then we define $W^+ := \sum_{i=1}^n R_i^+ 1 \{X_i > 0\}$.

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

$$W^+ = \left(\frac{n}{2}\right) U_n + \sum_{i=1}^n 1 \{X_i > 0\}.$$ 

**Question 4.2** (U-statistics, the information, ranking models, and probit regression): Suppose we have a standard linear regression problem with $Y_i = x_i^T \theta + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$, and $x_i \in \mathbb{R}^d$ are drawn i.i.d. from a distribution with $\mathbb{E}[x_i] = 0$ and $\text{Cov}(x_i) = \Sigma$. Assume that for all $n \geq d$ 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, $n \geq d$.

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

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 $\theta$ on the ordering $Y_i \preceq Y_j$. Consider the U-statistic-based “log-likelihood”

$$L_n(\theta) := \left(\frac{n}{2}\right)^{-1} \sum_{i,j \leq n} 1 \{Y_i > Y_j\} \log P_\theta(Y_i > Y_j | x_i, x_j).$$

(b) Show that $L_n(\theta)$ is concave in $\theta$. [Hint: Write it in terms of the Gaussian CDF.]

(c) Let $\hat{\theta}_n = \arg \max_\theta L_n(\theta)$. You may assume that $\hat{\theta}_n$ is consistent for $\theta_0$ under the true distribution $\theta_0$. What is the asymptotic distribution of $\hat{\theta}_n$?

(d) Which estimator of parts (a) and (c) do you prefer?
Assume that $X$ is compact. We call $k$ universal if it is dense in $C(\mathcal{X})$, 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 $L(f) \leq C \cdot \|f\|$ 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

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

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

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

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

(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

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

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 $X, X' \overset{iid}{\sim} P$ and $Z, Z' \overset{iid}{\sim} Q$.

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

13
(b) Argue that under the null \( H_0 \) we have

\[ U_n = O_p \left( \frac{1}{n} \right). \]

Hint: In the definition of \( U_n \), replace the kernel \( k \) with \( \hat{k}(x, x') := \langle \varphi(x) - \mu, \varphi(x') - \mu \rangle \) 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

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

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

\[ P = \mathcal{N}(0, I) \quad \text{and} \quad Q = \mathcal{N}(\theta, I). \]

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

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

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

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

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

Give the value \( t \) so that \( T_n \) is a level \( \alpha \) 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 \( O_p(n^{-1}) \) 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 \( k_\tau \), 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 \( \Psi_n \). Let \( \mathbb{P}_\theta \) be the joint distribution of \( P \) and \( Q \) when \( Q = N(\theta, I) \). Define the power values

\[
\pi^T_n(\theta) := \mathbb{P}_\theta(T_n \text{ rejects}) \quad \text{and} \quad \pi^\Psi_n(\theta) := \mathbb{P}_\theta(\Psi_n \text{ rejects})
\]

(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 \( \Psi_n \), 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 \( \pi^T_n(\theta_{n,d}) \) and \( \pi^\Psi_n(\theta_{n,d}) \) for these \( n \) and \( d \).

(d) Explain, in one or two sentences, the behavior in part (c).
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 \( X_i \overset{iid}{\sim} P_\theta \)) of the test.

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

\[
\sup_{\theta \in \Theta_0} \pi_n(\theta) \leq \alpha.
\]

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

\[
\pi_n(\theta) \leq \alpha + \epsilon.
\]

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

Hint: Note that for any distributions \( P \) and \( Q \),

\[
|P(T_n = 1) - Q(T_n = 1)| \leq \|P - Q\|_{TV},
\]

and use Question 1.9. What is the Hellinger distance between the \( n \)-fold product of \( N(\mu, \sigma^2) \) and \( N(0, \sigma^2) \)?

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

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

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 \( \alpha \) if \( \sup_{\theta \in \Theta_0} \pi_n(\theta) \leq \alpha \) and asymptotically of level \( \alpha \) if

\[
\limsup_n \sup_{\theta \in \Theta_0} \pi_n(\theta) \leq \alpha.
\]

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 \( P_\theta = N(\theta, 1) \) 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 \( \alpha \) 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

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

(e) Suppose now that the family includes a nuisance parameter of variance, so we have the model \( \{N(\theta, \sigma^2), \Theta \} \). 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 $P = N(0, \sigma^2)$ for some $\sigma^2 > 0$.) Using the same rejection regions $K_n^T$ and $K_n^S$ you developed in part (a), evaluate

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

(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 $X_i \stackrel{iid}{\sim} N(\theta_n, 1), i = 1, \ldots, n$, setting $\theta_n^h$ 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?
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 $E[X] = 0$. Define the cumulant generating function $\varphi(\lambda) := \log E_P[e^{\lambda X}]$, and let $Q_\lambda$ be the distribution on $X$ defined by

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

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

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

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

(c) Show that $\varphi(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}$ 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 $E[X_i] = 0$ and $E[|X_i|^k] < \infty$ for some $k \geq 1$. Let $S_n = \sum_{i=1}^n X_i$.

(a) Prove that

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

for a constant $C_k$ that depends only on $k$.

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

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

(c) If $E[|X_i|^k] \leq \sigma^k$ for some $\sigma < \infty$ for all $i$, then $P(|n^{-1}S_n| \geq t) \leq C_k (\frac{\sigma^2}{nt^2})^{\frac{k}{2}}$. 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

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

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 $\Delta_k$ is a function of $X_1, \ldots, X_k$, is a martingale difference sequence if

$$E[\Delta_k \mid X_1, \ldots, X_{k-1}] = 0$$

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 $\sigma_k^2$-sub-Gaussian if for $\Delta_k = Z_k - Z_{k-1}$, we have for each $k$ that

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

18
with probability 1 over the randomness in $X_1, X_2, \ldots$. Let $\Delta_k$ be a $\sigma^2_k$-sub-Gaussian martingale difference sequence with $Z_k = \sum_{i=1}^{k} \Delta_i$. Show that $Z_k$ is $\sum_{i=1}^{k} \sigma^2_i$-sub-Gaussian, and hence

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

**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

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

(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 $\psi$ 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 2.7.

(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 $\sigma^2$-sub-Gaussian if $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 \sigma^2/2)$.

(c) Show that if $X$ is $\sigma^2$-sub-Gaussian and mean-zero, then
\[ \mathbb{E}[\exp(\lambda X^2)] \leq \frac{1}{\sqrt{[1 - 2\lambda \sigma^2]_+}}. \]

*Hint:* If $Z \sim \mathcal{N}(0, \tau^2)$, 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 $\sigma^2$-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 [4].

**Question 6.7** (Orlicz norms: properties): In this question, we enumerate a few properties of Orlicz norms. Let $\psi_q(t) = e^{qt} - 1$ 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 $\psi$ 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 $\psi$ such that the inequality holds with constant 1.)
7 Uniform laws of large numbers and related problems

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

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

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 \( f : X \rightarrow \mathbb{R} \), 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 Rademacher complexities

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

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

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

(b) Assume that \( \mathcal{F} \) satisfies the envelope condition \( \sup_{x \in X} \sup_{f \in \mathcal{F}} |f(x) - Pf| \leq M \). 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 \( t \geq 0 \) we have

\[
\mathbb{P}\left( \sup_{f \in \mathcal{F}} |P_n f - Pf| \geq 2R_n(\mathcal{F}) + t \right) \leq 2 \exp\left( -\frac{c t^2}{M^2} \right).
\]

Question 7.3 (Rademacher complexities of some function classes):

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

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

\(^1\)Recall that for a norm \( \|\cdot\| \) on \( \mathbb{R}^d \), the dual norm is \( \|y\|_* = \sup_{x} \{x^T y : \|x\| \leq 1\} \).
[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 \(m_\theta(x, y) = \phi(y \theta^T x)\), and given an i.i.d. sample \(\{X_i, Y_i\}_{i=1}^n\), consider the empirical risk minimization procedure

\[
\hat{\theta}_n = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n m_\theta(X_i, Y_i) = \arg\min_{\theta \in \Theta} P_n m_\theta. \tag{3}
\]

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 \(\phi\) with functions in \(\mathcal{F}\). If \(\phi\) is an \(L\)-Lipschitz function, then \(R_n(\phi \circ \mathcal{F}) \leq LR_n(\mathcal{F})\).

(a) In one word, is the procedure (3) likely to give a reasonably good classifier? You may assume \(\phi(t)\) is strictly decreasing on \(t \in [-1, 1]\).

(b) Let \(\Theta \subset \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq r\}\) and let \(X_i\) be supported on the \(\ell_2\)-ball \(\{x \in \mathbb{R}^d : \|x\|_2 \leq M\}\). Give the smallest \(\epsilon_n(\delta, d, r, M)\) you can—ignoring numerical constants—such that

\[
\mathbb{P} \left( \sup_{\theta \in \Theta} \left| P_n m_\theta - P m_\theta \right| \geq \epsilon_n(\delta, d, r, M) \right) \leq \delta.
\]

How does your \(\epsilon_n\) compare with Question 7.1?

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

\[
\mathbb{P} \left( \sup_{\theta \in \Theta} \left| P_n m_\theta - P m_\theta \right| \geq \epsilon_n(\delta, d, r, M) \right) \leq \delta.
\]

How does your \(\epsilon_n\) compare with Question 7.1?

**Question 7.5** (Dvoretzky-Kiefer-Wolfowitz inequality): Let \(\mathcal{F} = \{1 \{x \leq t\} \mid t \in \mathbb{R}\}\) be the collection of indicator functions for \(x \leq t\). 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

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

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 \(R_n(\mathcal{F}) \leq \frac{C}{\sqrt{n}}\), where \(C\) is a universal (numerical) constant.

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

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

where \(c, C\) are absolute constants. (In fact, \(c = 2\) is possible using tools we have covered.)
**Question 7.6:** Let $R_n : \Theta \rightarrow \mathbb{R}$ 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 $\Theta$. Let $\theta_0 = \arg\min_\theta R(\theta)$, and for $\delta < \infty$, define $\Theta_\delta = \{ \theta : d(\theta, \theta_0) \leq \delta \}$. Let $\alpha \in (0, 2)$, $\sigma < \infty$, and $D > 0$, and assume we have the continuity bound

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

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

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

for $d(\theta, \theta_0) \leq D$. Let $\hat{\theta}_n = \arg\min_{\theta \in \Theta} R_n(\theta)$ and assume that $\hat{\theta}_n \overset{p}{\to} \theta_0$. Give the largest rate $r_n$ (i.e. a function of $n, \alpha, \beta$, ignoring other constants) you can for which

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

**Question 7.7:** 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 $\theta_0$ using the absolute loss, $\ell(\theta; x, y) = |x^T \theta - y|$, choosing $\hat{\theta}_n$ by

$$
\hat{\theta}_n \in \arg\min_{\theta} R_n(\theta) \quad \text{where } R_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta; X_i, Y_i).
$$

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

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

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

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

Now (and for the remainder of the question) we assume that there is a constant $D < \infty$ such that $\|X\|_2 \leq D$ with probability 1, i.e. $X$ is supported on the $\ell_2$-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. $\Sigma$ is positive definite (full rank).

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

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

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

23
(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 7.6, at what rate does $\hat{\theta}_n$ converge to $\theta_0$? Can you explain this behavior? (You may assume that $\hat{\theta}_n$ is consistent for $\theta_0$; you may also prove that it is consistent if you like.)

**Question 7.8** (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 $\text{Cov}(X_i) = \Sigma$ for a positive definite $\Sigma$. Assume also that $\mathbb{E}[|\langle v, X \rangle|^2] \geq \kappa \sqrt{v^T \Sigma v}$ for any vector $v \in \mathbb{R}^d$, where $\kappa > 0$ is a constant.

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

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

(b) Let $\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^T$ 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 $\kappa$, such that

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

for all $t \geq 0$.

**Question 7.9:** 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 $\ell$ 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 $\hat{\theta}_n = \text{argmin}_\theta R_n(\theta)$ 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 $\theta_0$, its median. Show that

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

*Hint:* see van der Vaart [2, 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 $\theta_0$. Indeed, assume that
\[
\min \{ P(X \geq \theta_0), P(X \leq \theta_0) \} \geq \frac{1}{2} + p_0
\] (4)
for some $p_0 > 0$.

(c) Show that under the conditions (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 $\sqrt{n}(\hat{\theta}_n - \theta_0)$ under condition (4)?

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

**Question 7.10** (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 $\theta$) for all $x \in \mathcal{X}$, and $P$ is a probability distribution. We assume $F(\theta; \cdot)$ is integrable for all $\theta$. Recall that a function $h$ is convex
\[
h(t\theta + (1-t)\theta') \leq th(\theta) + (1-t)h(\theta') \quad \text{for all } \theta, \theta' \in \mathbb{R}^d, \ t \in [0, 1].
\]

Let $\theta_0 \in \arg\min_{\theta} f(\theta)$, and assume that $f$ satisfies the following $\nu$-strong convexity guarantee:
\[
f(\theta) \geq f(\theta_0) + \frac{\nu}{2} \|\theta - \theta_0\|^2 \quad \text{for } \theta \text{ s.t. } \|\theta - \theta_0\| \leq \beta,
\]
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
\[
\hat{\theta}_n \in \arg\min_{\theta} f_n(\theta).
\]

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

(b) Show that $f$ and $f_n$ are convex.

(c) Show that $\theta_0$ 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)
\[
P\left( \|\hat{\theta}_n - \theta_0\| \geq \sigma \cdot \frac{1 + t}{\sqrt{n}} \right) \leq C \exp\left(-t^2\right)
\]
for all $t \leq \sigma'\sqrt{n}\beta$, 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.
8 High-dimensional problems

Question 8.1: Consider the sub-Gaussian sequence model

\[ Y = \theta + \sigma \varepsilon, \quad (5) \]

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

\[ S_\lambda(v) := \text{sign}(v) \begin{cases} v - \lambda & \text{if } v \geq \lambda \\ 0 & \text{if } v \in [-\lambda, \lambda] \\ v + \lambda & \text{if } v \leq -\lambda, \end{cases} \]

gives the soft-thresholding estimator (when applied elementwise)

\[ \hat{\theta} := S_\lambda(Y) = \arg\min_{\theta} \left\{ \frac{1}{2} \| \theta - Y \|_2^2 + \lambda \| \theta \|_1 \right\}. \]

In this question, we give high-probability bounds on the error of \( \hat{\theta} \) for the sub-Gaussian sequence model (5) when \( \theta \) is \( k \)-sparse, meaning that \( \| \theta \|_0 = \text{card} \{ j \in [n] \mid \theta_j \neq 0 \} \leq k \).

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

\[ \| \hat{\theta} - \theta \|_2^2 \leq 4k\lambda^2. \]

(b) Show that if

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

where \( t \geq 0 \), then

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

Conclude that probability at least \( 1 - \frac{1}{2n} \), in the sub-Gaussian sequence model (5), the soft-thresholding estimator with \( \lambda = 2\sqrt{\sigma^2 \log(2n)} \) satisfies

\[ \| \hat{\theta} - \theta \|_2^2 \leq 16k\sigma^2 \log(2n). \]

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 \( \mu \)-pairwise incoherent if

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

satisfies \( \delta_{pw}(A) \leq \mu \), where \( \| \cdot \|_\infty \) denotes the entrywise \( \ell_\infty \) 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 \( \delta_{pw}(A) \leq \mu \), 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 \( \delta_{pw}(A) = \mu < \frac{1}{2\sqrt{d}} \), 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 null(\( A \)) \cap \mathbb{C}(S) = \{ 0 \}. 
**Question 8.3 (The square root Lasso):** The square-root Lasso chooses the estimator \( \hat{\theta} \) via

\[
\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{\sqrt{n}} \| y - X\theta \|_2 + \gamma \| \theta \|_1 \right\}.
\]

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

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

\[
\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^d} \inf_{\lambda \geq 0} \left\{ \frac{1}{2n} \frac{\| y - X\theta \|_2^2}{\lambda} + \frac{\lambda}{2} + \gamma \| \theta \|_1 \right\}.
\]

What value does \( \lambda \) take on at \( \theta = \theta^* \)? Why might this be a valuable quantity?

(b) Let \( \theta^* \) have support \( S = \text{supp} \theta^* = \{ j \in [d] : \theta^*_j \neq 0 \} \). Show that

\[
\frac{1}{\sqrt{n}} \| X\hat{\theta} - y \|_2 - \frac{1}{\sqrt{n}} \| X\theta^* - y \|_2 \leq \gamma (\| \Delta S \|_1 - \| \Delta S^c \|_1),
\]

where \( \hat{\theta} = \theta^* + \Delta \).

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

\[
\| X\hat{\theta} - y \|_2 \geq \| X\theta^* - y \|_2 - \frac{\| X^T\varepsilon \|_\infty}{\| \varepsilon \|_2} \| \Delta \|_1.
\]

(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 \( \gamma \geq 2 \frac{\| X^T\varepsilon \|_\infty}{\sqrt{n} \| \varepsilon \|_2} \).

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

\[
\frac{1}{\sqrt{n}} X^T(X\hat{\theta} - y) + \gamma z = 0
\]

for some \( z \in \partial \| \hat{\theta} \|_1 \), the subdifferential of the \( \ell_1 \)-norm at \( \hat{\theta} \).

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

\[
\frac{1}{n} \| X\Delta \|_2^2 \leq \frac{1}{n} \langle \Delta, X^T\varepsilon \rangle + \gamma \frac{\| X\Delta - \varepsilon \|_2}{\sqrt{n}} (\| \Delta S \|_1 - \| \Delta S^c \|_1)
\]

\[
\leq \frac{1}{n} \langle \Delta, X^T\varepsilon \rangle + \gamma \frac{\| \varepsilon \|_2}{\sqrt{n}} \| \Delta S \|_1 + \gamma^2 \| \Delta S^c \|_1^2.
\]

(You should prove both inequalities.)

(g) Suppose that \( X \) satisfies the restricted strong convexity condition that \( \| X\Delta \|_2 \geq \mu \| \Delta \|_2 \) 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

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

27
(h) If $X \in \mathbb{R}^{n \times d}$ has columns with norm $\|X_i\|_2^2 = n$, and $\varepsilon_i$ are independent with mean zero, $C\sigma^2$-sub-Gaussian, and $E[\varepsilon_i^2] = \sigma^2$, argue that if $\gamma = C\sqrt{\log d} \sqrt{\frac{\log d}{n}}$ for a constant $C$, then $\|\hat{\theta} - \theta^*\|_2^2 \lesssim \sigma \frac{\sqrt{\log d}}{\sqrt{n}}$ with high probability.

(i) What advantages does this have (in one sentence) over the standard Lasso program?
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 $\|G - F\|_\infty \leq \delta$ implies $|\gamma(F) - \gamma(G)| \leq \epsilon$.

(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 $\gamma$ 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.

1. The mean functional $F \mapsto \int x dF(x)$.
2. The Cramér-von Mises functional $F \mapsto \int (F(x) - F_0(x_0))^2 dF_0(x)$.
3. The quantile functional $Q_p(F) := \inf\{t \in \mathbb{R} \mid F(t) \geq p\}$.

**Question 9.2:** We consider estimation of median-like quantities in dimension $d \geq 1$. Let $\|\cdot\|_2$ denote the typical $\ell_2$-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

$$\hat{\theta}_n := \arg\min_{\theta \in \mathbb{R}^d} R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell_\theta(X_i).$$

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 [1]): if $\ell_\theta(x)$ is convex in $\theta$ for all $x$ and for $P$-almost every $x$ is differentiable in a neighborhood of a point $\theta_0$ with derivative $\hat{\ell}_\theta(x) = \nabla_\theta \ell_\theta(x)$, then

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

Similarly, if the Hessian $\dddot{\ell}_\theta = \nabla_\theta^2 \ell_\theta$ exists with $P$-probability 1 near $\theta_0$, then $\nabla^2 R(\theta) = \mathbb{E}[\dddot{\ell}_\theta(X)]$.

(a) Show that the set $\arg\min_{\theta} R(\theta) = \{\theta_0 \in \mathbb{R}^d \mid R(\theta_0) \leq \inf_{\theta} R(\theta)\}$ 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 $\theta_0 \in \arg\min_{\theta} R(\theta)$.

(b) Show that $\theta_0$ is unique. Hint: Question 7.10.(c) may be useful, as $\ell_\theta$ is convex in $\theta$.

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

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

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

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

(f) Compare the asymptotic distribution of \( \| \hat{\theta}_n \|^2 \) to that of \( \| X_n \|_2^2 \), the sample mean, when \( X_i \sim \mathcal{N}(0, I) \). What sample size \( m(n) \) is required (as a function of \( n \)) for \( \hat{\theta}_{m(n)} \) to have the same asymptotic performance as \( X_n \)?

**Question 9.3 (Elliptical classes are Donsker):** Recall that a collection \( \mathcal{F} \) of functions is \( P \)-Donsker if the process \( \mathcal{G}_n := \sqrt{n} (P_n - P) \), viewed as a mapping \( \mathcal{G}_n : \mathcal{F} \to \mathbb{R} \) via \( \mathcal{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Pf) \), converges to a tight Gaussian process \( \mathcal{G} \) 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 \( \| f - g \|_{L^2(P)} = P(f - g)^2 \), we have

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

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 \( k \in \mathbb{N} \),

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

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

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

whence the elliptical class \( \mathcal{F} \) is Donsker.
10 Contiguity and Quadratic Mean Differentiability

**Question 10.1:** Let $P_n$ and $Q_n$ be sequences of probability measures with $\|P_n - Q_n\|_{TV} \to 0$. 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 $\ell_\theta : \mathcal{X} \to \mathbb{R}^d$ such that
\[
\int \left( \sqrt{p_\theta + h} - \sqrt{p_\theta} - \frac{1}{2} h^\top \ell_\theta \sqrt{p_\theta} \right)^2 d\mu = o(\|h\|^2).
\]
Let $P^n$ denote the $n$-fold product of $P$ (i.e. $n$ i.i.d. observations from $P$).

(a) Show that
\[
\lim_{n \to \infty} d^2_{hel}(P^n_{\theta_0}, P^n_{\theta_0 + h/\sqrt{n}}) = 1 - \exp \left( -\frac{1}{8} h^\top I_{\theta_0} h \right).
\]

(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
\[
\liminf_n \left\{ P^n_{\theta_0} (\psi_n \neq 0) + P^n_{\theta_0 + h/\sqrt{n}} (\psi_n \neq 1) \right\} > 0.
\]

**Question 10.3:** Let $P_\theta$ 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
\[
\liminf_{n \to \infty} \inf_{\psi_n} \left\{ P^n_{\theta + h/\sqrt{n}} (\psi_n \neq 1) + P^n_{\theta} (\psi_n \neq 0) \right\}.
\]
Explain your result in the light of Question 10.2.
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

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

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

Question 11.2 (Estimating a non-differentiable function): Let $\mathcal{P}$ be a collection of distributions on a space $\mathcal{X}$ and $\psi : \mathcal{P} \rightarrow \mathbb{R}$ 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 $g : \mathcal{X} \rightarrow \mathbb{R}$, $Pg = 0$, $Pg^2 < \infty$

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

We illustrate some of the difficulties in estimation of $\psi(P)$ along a path $t \mapsto P_t$ for which $\psi(P)$ is not differentiable. For simplicity, we assume that $\psi(P) = 0$ (this is no loss of generality) and that the path is such that

$$\limsup_{t \downarrow 0} \frac{\left|\psi(P_t) - \psi(P)\right|}{t} = +\infty.$$

(a) Letting $P^n_t$ denote the $n$-fold product of $P$, give the limit $\lim_{n \to \infty} d^2_{hel}(P^n_t \sqrt{n}, P^n)$, where $t \in \mathbb{R}_+$.

(b) Show that for any $t$, we have

$$\inf_T \{P^n_t(T \neq 0) + P^n(T \neq 1)\} \geq 1 - d^2_{hel}(P^n_t, P^n_0).$$

(c) Show that for any sequence $\epsilon_n \downarrow 0$, we have

$$\sup_{0 \leq t \leq \epsilon_n / \sqrt{n}} \inf_T \{P^n_t(T \neq 0) + P^n(T \neq 1)\} \to 1.$$

(d) Show that we have the local minimax lower bound

$$\lim_{\epsilon \downarrow 0} \lim_{K \to \infty} \lim_{n \to \infty} \sup_{0 \leq t \leq \epsilon_n / \sqrt{n}} \inf_{\hat{\psi}_n} \max_{t \in \{0, t\}} P^n_t \left(\left|\hat{\psi}_n - \psi(P_t)\right| \geq \frac{K}{\sqrt{n}}\right) \geq \frac{1}{2}.$$

(e) Give a one-sentence description of this result.

Question 11.3 (Score and influence functions for regression): Consider the prediction problem of finding $\theta$ to best predict a scalar $y$ from $x \in \mathbb{R}^d$ via the model $\hat{y} = \theta^\top x$. We study the local asymptotic minimax risk for estimation of the parameter

$$\theta(P) := \arg\min_\theta \mathbb{E}_P[(Y - X^T \theta)^2],$$

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 \( G \) be the collection of functions \( g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \), \( P g^2 < \infty \), and \( Pg = 0 \). Let \( L : \mathbb{R}^d \to \mathbb{R}_+ \) be a symmetric quasi-convex, bowl-shaped and Lipschitz loss. For functions \( g_1, \ldots, g_k \in G \) 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 \( g(x, y) = [g_1(x, y) \cdot \cdots \cdot g_k(x, y)]^\top \). Let \( \theta_h = \theta(P_h) \) for shorthand. Construct an influence function for the parameter \( \theta_h \), that is, a function \( \psi_{\text{inf}} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \) 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

\[
\sup_{k \in \mathbb{N}, g_1, \ldots, g_k \in G} \lim_{c \to \infty} \lim_{n \to \infty} \lim_{\theta \to \theta_h} \inf \mathbb{E}_{P_n^k} \left[ L(\sqrt{n} (\hat\theta_n - \theta_h)) \right] d\pi_{n,c}(h).
\]

What does your lower bound become when the model \( y = x^\top \theta + \varepsilon \) holds, where \( \varepsilon \) is a mean-zero independent noise with \( P\varepsilon^2 = \sigma^2 \)?


References


