There are five questions. You only need to do three. Circle the three questions you want to be graded:

1  2  3  4  5

Name: ________________________________
Problem 1: Let $X_1, \ldots, X_n$ be a random sample where $-B \leq X_i \leq B$ for some finite $B > 0$. For every real number $a$ define

$$R(a) = \mathbb{E}|X - a|, \quad \hat{R}_n(a) = \frac{1}{n} \sum_{i=1}^{n} |X_i - a|.$$ 

Let $a^*$ minimize $R(a)$ and let $\hat{a}$ minimize $\hat{R}_n(a)$. That is,

$$a^* = \arg\min_{-B \leq a \leq B} R(a), \quad \hat{a} = \arg\min_{-B \leq a \leq B} \hat{R}_n(a).$$

In this question you will show that, with high probability, $R(\hat{a}) \leq R(a^*) + O(\sqrt{\log n/n})$ with high probability.

(a) Let $P_n$ be the empirical distribution. Thus $P_n(A) = (\text{number of } X_i \in A)/n$. Show that

$$\sup_{-B \leq a \leq B} |R(a) - \hat{R}_n(a)| \leq 2B \sup_{A \in \mathcal{A}} |P_n(A) - P(A)|$$

where

$$\mathcal{A} = \left\{ \{x : g_a(x) > t\} : a \in [-B, B], \ t > 0 \right\}$$

and $g_a(x) = |x - a|$.

Hint: Note that

$$R(a) = \mathbb{E}(g_a(X)) = \int_0^{2B} \mathbb{P}(g_a(X) > t)dt$$

and

$$\hat{R}_n(a) = \int_0^{2B} P_n(g_a(X) > t)dt = \int_0^{2B} \frac{1}{n} \sum_{i=1}^{n} I_{g_a(X_i) > t} dt.$$

(There is workspace on the next page.)
Workspace for part (a).

Ans.
Using the hint, we know that
\[
|R(a) - \tilde{R}_n(a)| = |\int_0^{2B} P(g_a(X) > t) - P_n(g_a(X) > t)dt|
\leq \int_0^{2B} |P(g_a(X) > t) - P_n(g_a(X) > t)|dt
\leq \int_0^{2B} \sup_{t \geq 0} |P(g_a(X) > t) - P_n(g_a(X) > t)|dt
= 2B \sup_{t \geq 0} |P(g_a(X) > t) - P_n(g_a(X) > t)|
\]
Since this inequality is true for all \(a\), we get that
\[
\sup_{-B \leq a \leq B} |R(a) - \tilde{R}_n(a)| \leq 2B \sup_{-B \leq a \leq B} \sup_{t \geq 0} |P_n(g_a(X) > t) - P(g_a(X) > t)|
\leq 2B \sup_{A \in \mathcal{A}} |P_n(A) - P(A)|
\]
(b) Compute the VC dimension of $\mathcal{A}$. **Ans.** $\mathcal{A}$ is defined as \{\{x : |x - a| > t\} : a \in [-B, B], t > 0\}. This is the set of all two-sided intervals with gap in the center.

![Figure 1: Example of an element of the set $\mathcal{A}$](image)

It is clear that such family of intervals can shatter any set of 2 numbers in $[-B, B]$. Let $x_1 < x_2 < x_3 \in [-B, B]$; it is also easy to see that $\{x_2\}$ cannot be picked out by any elements of $\mathcal{A}$.

Hence, VC-dimension of $\mathcal{A}$ is 2.
(c) Recall that if $\mathcal{A}$ has VC dimension $d$ then
\[
\Pr \left( \sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon \right) \leq c_1 n^d e^{-c_2 n \epsilon^2}
\]
for some $c_1$ and $c_2$. Use this fact, together with the results from (a) and (b) to show that
\[
\sup_a |\hat{R}_n(a) - R(a)| < \epsilon
\]
with high probability.

**NOTE:** there was a typo in the bound, it should be $n^d$ instead of $d^n$ as stated in the exam. We will accept both as correct but only work with $n^d$ in the solutions.

**Ans.**
\[
P(\sup_{-B \leq a \leq B} |\hat{R}_n(a) - R(a)| > \epsilon) \leq P(2B \sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon) = P(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \frac{\epsilon}{2B}) \leq c_1 n^3 \exp(-c_2 n \frac{\epsilon^2}{4B^2})
\]
Where we used that the VC dimension $d = 3$. 

---

5
(d) Find $z(n)$ such that
\[ R(\hat{a}) \leq R(a_*) + z(n) \]
with probability at least $1 - \delta$.

**Ans.** We set $\delta = c_1 n^3 \exp(-c_2 n \frac{a^2}{4B^2})$ and work through a little algebra to find that $\epsilon = \sqrt{\frac{4B^2}{nc_2} (3 \log n + \log \frac{c_1}{\delta})}$.

Hence, with probability at least $1 - \delta$, we know that for all $a \in [-B, B]$, $|\hat{R}_n(a) - R(a)| \leq z'(n)$ where $z'(n) = \sqrt{\frac{4B^2}{nc_2} (3 \log n + \log \frac{c_1}{\delta})}$

By definition of $\hat{a}$ and $a_*$, we can conclude that with probability at least $1 - \delta$:
\[
R(\hat{a}) - R(a_*) = R(\hat{a}) - \hat{R}_n(\hat{a}) + \hat{R}_n(\hat{a}) - \hat{R}_n(a_*) + \hat{R}_n(a_*) - R(a_*) \\
\leq |R(\hat{a}) - \hat{R}_n(\hat{a})| + |\hat{R}_n(a_*) - R(a_*)| \\
\leq 2z'(n)
\]
where we used the fact that $\hat{a}$ is the empirical risk minimizer and hence $\hat{R}_n(\hat{a}) \leq \hat{R}_n(a_*)$.

Set $z(n) = 2z'(n)$ and we get the desired bound.
Problem 2: Let $P_1$ and $P_2$ be two distributions with densities $p_1$ and $p_2$. Recall that $\text{TV}(P_1, P_2) = \sup_A |P_1(A) - P_2(A)|$.

(a) Show that
$$\int p_1 \wedge p_2 = 1 - \text{TV}(P_1, P_2)$$

where $p_1(x) \wedge p_2(x) = \min \{p_1(x), p_2(x)\}$.

Ans.

Note that for any $A \subset \mathbb{R}$, $P_1(A) - P_2(A) = (1 - P_1(A^c)) - (1 - P_2(A^c)) = P_2(A^c) - P_1(A^c)$.

Hence, $\sup_A P_1(A) - P_2(A) = \sup_A P_2(A) - P_1(A) = \sup_A |P_1(A) - P_2(A)|$.

Now, $\sup_A P_1(A) - P_2(A) = \sup_A \int_{x \in A} p_1(x) - p_2(x) dx$ and it is clear that $A = \{x : p_1(x) > p_2(x)\}$.

$$1 - \text{TV}(P_1, P_2) = \int_A p_1(x) dx + \int_{A^c} p_1(x) dx - \left( \int_A p_1(x) - p_2(x) dx \right)$$
$$= \int_{A^c} p_1(x) dx + \int_A p_2(x) dx$$
$$= \int p_1 \wedge p_2 dx$$

Where the last equality follow from the observation that $A = \{x : p_1(x) > p_2(x)\}$ and that $A \cup A^c = \mathbb{R}$. We performed our analysis assuming support is $\mathbb{R}$ but it can generalize to any measure space.
Let \( \mathcal{P} \) be a set of distributions. Let \( P_1 \) and \( P_2 \) be two arbitrary distributions in \( \mathcal{P} \). Let \( X \sim P \) for some \( P \in \mathcal{P} \). Let \( \theta : \mathcal{P} \rightarrow \mathbb{R} \) and let \( \hat{\theta} = \hat{\theta}(X) \) denote an estimator of \( \theta(P) \).

Show that
\[
\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P |\hat{\theta} - \theta(P)| \geq \frac{|\theta(P_1) - \theta(P_2)|}{4} \left(1 - \text{TV}(P_1, P_2)\right).
\]

**Ans.** We first finitize and discretize:

\[
\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P |\hat{\theta}(X) - \theta(P)| \geq \inf_{\hat{\theta}} \max_{P \in \{P_1, P_2\}} \mathbb{E}_P |\hat{\theta}(X) - \theta(P)|
\]

\[
\geq \inf_{\hat{Z}} \max_{P \in \{P_1, P_2\}} P_i(Z(X) \neq i) |\frac{\theta(P_1) - \theta(P_2)}{2}|
\]

\[
\geq \inf_{\hat{Z}} [P_1(Z(X) \neq 1) + P_2(Z(X) \neq 2)] |\frac{\theta(P_1) - \theta(P_2)}{4}|
\]

where \( Z \) is a binary function of the data.

By Neyman-Pearson lemma, the estimator \( Z^* \) that minimizes \( P_1(Z(X) \neq 1) + P_2(Z(X) \neq 2) \) is \( Z^*(X) = 1 \) if \( p_1(X) > p_2(X) \) and \( Z^*(X) = 2 \) if \( p_2(X) > p_1(X) \).

Hence, \( P_1(Z^*(X) \neq 1) = \int_{x : p_1(x) < p_2(x)} p_1(x) dx \) and \( P_2(Z^*(X) \neq 2) \geq \int_{x : p_2(x) < p_1(x)} p_2(x) dx \).

Combining these two results, we have that \( P_1(Z^*(X) \neq 1) + P_2(Z^*(X) \neq 2) = \int p_1 \wedge p_2 dx \).

Thus, \( \inf_{\hat{Z}} [P_1(Z(X) \neq 1) + P_2(Z(X) \neq 2)] \geq \int p_1 \wedge p_2 dx \) and we get the desired bound.
Problem 3. In class, we saw that a kernel density estimate can achieve a mean square error (MSE) rate of $n^{-2/(2+d)}$ for Lipschitz densities. The same rate is true for a histogram density estimate as well. Moreover if the density has compact support, the same is true for mean integrated square error (MISE) $\mathbb{E}[\int |\hat{p}(x) - p(x)|^2 dx]$ which is a global measure of accuracy.

In this problem, you will derive the rate of MISE convergence for densities that are piecewise-smooth, i.e. they are Lipschitz everywhere, except for a few points where the densities can have a discontinuity.

Consider univariate ($d = 1$) densities supported on the unit interval $[0, 1]$ that satisfy $|p(x) - p(x')| \leq L|x - x'|$ for all $x \in [0, 1]$, except for $N$ (a finite number of) points where it may jump. You may assume that the density is bounded from above, i.e. $p(x) \leq B < \infty$.

Consider a histogram density estimator based on $n$ samples $\{X_i\}_{i=1}^n$ drawn i.i.d. from the density as follows:

$$\hat{p}(x) = \frac{m}{n} \sum_{j=1}^m \hat{p}_j I(x \in B_j)$$

where $\hat{p}_j = \frac{m}{n} \sum_{i=1}^n I(X_i \in B_j)$ and $B_1 = [0, 1/m), B_2 = [1/m, 2/m), \ldots, B_m = [(m-1)/m, 1)$. Denote its mean by $\bar{p}(x) = \mathbb{E}[\hat{p}(x)]$.

(a) Compute the integrated square bias $\int |\hat{p}(x) - p(x)|^2 dx$ of the histogram density estimator.

**Ans.** We first look at $\bar{p}(x)$:

$$\bar{p}(x) = \mathbb{E}[\hat{p}(x)] = \sum_{j=1}^m \mathbb{E}[\hat{p}_j] I(x \in B_j) = \sum_{j=1}^m mP(B_j) I(x \in B_j),$$

where $P(B_j) := \int_{y \in B_j} p(y) dy$. Then, we have the integrated squared bias

$$\int |\bar{p}(x) - p(x)|^2 dx = \int \left| \sum_{j=1}^m mP(B_j) I(x \in B_j) - p(x) \right|^2 dx = \sum_{j=1}^m \int_{x \in B_j} |mP(B_j) - p(x)|^2 dx.$$

For each $B_j$, let us consider two cases.

1. $B_j$ contains none of the $N$ discontinuities. Using the Lipschitz property, we get

$$\int_{x \in B_j} |mP(B_j) - p(x)|^2 dx = \int_{x \in B_j} \left| m \int_{y \in B_j} (p(y) - p(x)) dy \right|^2 dx \leq \int_{x \in B_j} \left( m \int_{y \in B_j} |p(y) - p(x)| dy \right)^2 dx \leq \int_{x \in B_j} \left( m \int_{y \in B_j} \frac{L}{m} |y - x| dy \right)^2 dx \leq \int_{x \in B_j} \frac{L^2}{m^2} m^2 dx = \frac{L^2}{m^3}.$$
(2) $B_j$ contains at least one of the $N$ discontinuities. Using the assumption that $p(x) \leq B < \infty$, we get

$$
\int_{x \in B_j} |mP(B_j) - p(x)|^2 \, dx = \int_{x \in B_j} \left| m \int_{y \in B_j} (p(y) - p(x)) \, dy \right|^2 \, dx
$$

$$
\leq \int_{x \in B_j} \left( m \int_{y \in B_j} |p(y) - p(x)| \, dy \right)^2 \, dx
$$

$$
\leq \int_{x \in B_j} \left( m \int_{y \in B_j} B \, dy \right)^2 \, dx
$$

$$
\leq \int_{x \in B_j} B^2 \, dx = \frac{B^2}{m}.
$$

Since $N$ is finite, we have that

$$
\int |\bar{p}(x) - p(x)|^2 \, dx \leq \frac{cNB^2}{m}
$$

for some constant $c$ and large $m$.

(b) Compute the integrated variance $\int \mathbb{E}[(\hat{p}(x) - \bar{p}(x))|^2] \, dx$.

**Ans.**

$$
\int \mathbb{E}[(\hat{p}(x) - \bar{p}(x))|^2] \, dx = \int \mathbb{E} \left[ \sum_{j=1}^m (\hat{p}_j - mP(B_j))I(x \in B_j) \right]^2 \, dx
$$

$$
= \sum_{j=1}^m \frac{\mathbb{E}[|\hat{p}_j - mP(B_j)|^2]}{m}
$$

$$
= \sum_{j=1}^m m \mathbb{E} \left[ \frac{\hat{p}_j}{m} - P(B_j) \right]^2
$$

$$
= \sum_{j=1}^m m \mathbb{V} \left[ \sum_{i=1}^n I(X_i \in B_j) \right]
$$

$$
= \sum_{j=1}^m \frac{m}{n} P(X \in B_j)(1 - P(X \in B_j))
$$

$$
\leq \sum_{j=1}^m \frac{m}{n} P(X \in B_j) = \frac{m}{n}.
$$

(c) Derive the rate of mean integrated square error (MISE) convergence.

**Ans.** The MISE is the integrated squared bias plus the integrated variance. To get the
optimal \( m \), we let

\[
\frac{m}{n} = \frac{cN B^2}{m} \iff m = B \sqrt{cN} \sqrt{n},
\]

leading to \( \text{MISE} \in O(n^{-1/2}) \).

(d) How does this rate compare to the MISE rate for estimating a Lipschitz smooth density?

Comment.

\textbf{Ans.} The MISE rate for estimating a Lipschitz smooth density, when \( d = 1 \), is \( n^{-2/3} \), which is faster than our rate \( n^{-1/2} \) here. The reason is that discontinuous points increase the bias in the estimate from \( O(1/m^2) \), which is the case for smooth densities, to \( O(1/m) \). The variances in both cases are the same.
Problem 4. Let \( x^\top = [x_A^\top \ x_B^\top] \) be a random vector following a zero-mean Gaussian distribution with precision (inverse covariance)
\[
\Omega = \begin{bmatrix}
\Omega_{AA} & \Omega_{AB} \\
\Omega_{BA} & \Omega_{BB}
\end{bmatrix},
\]
where \( A \) and \( B \) form a partition of the variables.

(a) Write the conditional density \( p(x_A|x_B) \) in terms of \( \Omega_{AA}, \Omega_{AB}, \Omega_{BA}, \Omega_{BB} \).
\[
\log p(x_A, x_B) \\
\propto -\frac{1}{2} [x_A \ x_B^\top] \begin{bmatrix}
\Omega_{AA} & \Omega_{AB} \\
\Omega_{BA} & \Omega_{BB}
\end{bmatrix} [x_A \ x_B] \\
= -\frac{1}{2} (x_A^\top \Omega_{AA} x_A + 2x_B^\top \Omega_{BA} x_A + x_B^\top \Omega_{BB} x_B) \\
= -\frac{1}{2} \left( (x_A + \Omega_{AA}^{-1} \Omega_{AB} x_B)^\top \Omega_{AA} (x_A + \Omega_{AA}^{-1} \Omega_{AB} x_B) + x_B^\top (\Omega_{BB} - \Omega_{BA}(\Omega_{AA})^{-1} \Omega_{AB}) x_B \right).
\]
This suggests that the marginal distribution \( p(x_B) \), obtained by integrating \( p(x_A, x_B) \) over \( x_A \), is a zero mean Gaussian with inverse covariance
\[
\Omega_{BB} - \Omega_{BA}(\Omega_{AA})^{-1} \Omega_{AB},
\]
which then gives that
\[
p(x_A|x_B) = \frac{p(x_A, x_B)}{p(x_B)} = \mathcal{N}(-\Omega_{AA}^{-1} \Omega_{AB} x_B, \Omega_{AA}^{-1}).
\]

(b) Show that the precision matrix of \( x_A \) given \( x_B \) does NOT depend on the value of \( x_B \).
\[
\text{Ans.} \quad \text{From (a) we know the precision matrix of } x_A \text{ given } x_B \text{ is } \Omega_{AA}, \text{ which does not depend on the value of } x_B.
\]

(c) Write the marginal density \( p(x_A) \) in terms of \( \Omega_{AA}, \Omega_{AB}, \Omega_{BA}, \Omega_{BB} \).
\[
\text{Ans.} \quad \text{Switching } x_A \text{ and } x_B \text{ in the derivation in (a), we get that}
\]
\[
p(x_A) = \mathcal{N}(0, (\Omega_{AA} - \Omega_{AB}(\Omega_{BB})^{-1} \Omega_{BA})^{-1}).
\]

(d) Assume the variables in \( x_A \) are mutually independent of one another conditioning on \( x_B \). Would the variables in \( x_A \) be mutually independent? Why or why not?
\[
\text{Ans.} \quad \text{The variables in } x_A \text{ are mutually independent of one another conditioning on } x_B \text{ if and only if the precision matrix of the condition distribution, which has been shown in (a) to be } \Omega_{AA}, \text{ is diagonal. The variables in } x_A \text{ are mutually independent if and only if the precision matrix of the marginal, } \Omega_{AA} - \Omega_{AB}(\Omega_{BB})^{-1} \Omega_{BA}, \text{ is diagonal. Obviously, } \Omega_{AA} \text{ being diagonal does not guarantee } \Omega_{AA} - \Omega_{AB}(\Omega_{BB})^{-1} \Omega_{BA} \text{ to be diagonal, so the answer is no.}
\]
Problem 5. Let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{p \times n}$. The Lasso problem is to solve, for a given regularization parameter $\lambda$,

$$
\Phi(\lambda) = \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} ||Y - X\beta||_2^2 + \lambda ||\beta||_1.
$$

In this problem, we show that one can equivalently solve

$$
\Psi(t) = \min_{\beta \in \mathbb{R}^p: ||\beta||_1 \leq t} \frac{1}{2n} ||Y - X\beta||_2^2.
$$

(a) Show that both optimizations are convex. Ans.

We know that $h(x) = ||x||_2^2$ is convex since gradient of $f$ at $x_0$ is $2x_0$ and the Hessian of $f$ at $x_0$ is $2I$. Since composition of a convex function with an affine function is convex, we know that $f(\beta) = ||Y - X\beta||_2^2$ is convex for all $Y, X$.

Finally, since $|| \cdot ||_1$ is a norm, it is convex and thus, $\Phi(\lambda)$ contains a convex optimization. Likewise, the constraint in $\Psi(t)$ is convex and thus the second optimization is convex as well.
(b) Prove that for a fixed $t_0$, there exist a unique $\lambda_0$ such that if $\hat{\beta}$ minimizes $\frac{1}{2n}||Y - X\beta||_2^2$ for $||\beta||_1 \leq t_0$ then $\hat{\beta}$ also minimizes $\frac{1}{2n}||Y - X\beta||_2^2 + \lambda_0||\beta||_1$. Show that

$$\lambda_0 = \text{argsup}_{\lambda \geq 0} \Phi(\lambda) - \lambda t_0.$$

(Hint: Use strong duality.)

**Ans.** We first take the constrained form and write down the Lagrangian:

$$L(\beta, \lambda) = \frac{1}{2n}||Y - X\beta||_2^2 + \lambda(||\beta||_1 - t_0)$$

$$= \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1 - \lambda t_0$$

Since both optimizations are convex, by strong duality we have

$$\Psi(t_0) = \min_{\beta} \sup_{\lambda} L(\beta, \lambda) = \sup_{\lambda} \min_{\beta} L(\beta, \lambda) = \sup_{\lambda} \Phi(\lambda) - \lambda t_0$$

Let $(\beta^*, \lambda_0)$ be a pair of primal-dual optimal solution. Then by KKT conditions, it must be that subgradient of $L(\beta, \lambda_0)$ at $\beta^*$ contains 0 and hence $\beta^*$ is the global optimum of the optimization in $\Phi(\lambda_0)$.

Since $\lambda_0$ is the dual optimum, it must be that $\lambda_0$ optimizes $\sup_{\lambda} \Phi(\lambda) - \lambda t_0$.

By strong duality, we know that $\lambda_0$ is global dual optimum, and by the fact that $\Phi(\lambda) - \lambda t_0$ is strongly convex in $\lambda$, we know that $\lambda_0$ is unique.
(c) Is it true that $\Psi(t_0) = \Phi(\lambda_0)$? Explain.

**Ans.** $\Phi(\lambda_0) = \Psi(t_0) + \lambda_0 t_0$ and hence the two are not equal.
(Extra Blank Paper.)