10-702/36-702
Midterm Exam Solutions

March 2 2011

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

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

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

$$a_* = \underset{-B \leq a \leq B}{\operatorname{argmin}} R(a), \qquad \widehat{a} = \underset{-B \leq a \leq B}{\operatorname{argmin}} \widehat{R}_n(a).$$

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

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

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

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

(There is workspace on the next page.)

Workspace for part (a).

Ans.

Using the hint, we know that

$$\begin{aligned}
|R(a) - \widehat{R}_n(a)| &= |\int_0^{2B} P(g_a(X) > t) - P_n(g_a(X) > t)dt| \\
&\leq \int_0^{2B} |P(g_a(X) > t) - P_n(g_a(X) > t)|dt \\
&\leq \int_0^{2B} \sup_{t\geq 0} |P(g_a(X) > t) - P_n(g_a(X) > t)|dt \\
&= 2B \sup_{t\geq 0} |P(g_a(X) > t) - P_n(g_a(X) > t)|
\end{aligned}$$

Since this inequality is true for all a, we get that

$$\begin{aligned}
\sup_{-B\leq a\leq B} |R(a) - \widehat{R}_n(a)| &\leq 2B \sup_{-B\leq a\leq B} \sup_{t\geq 0} |P_n(g_a(X) > t) - P(g_a(X) > t)| \\
&\leq 2B \sup_{A\in\mathcal{A}} |P_n(A) - P(A)|
\end{aligned}$$

(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.

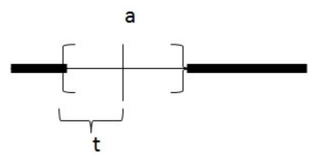

Figure 1: Example of an element of the set $\mathcal{A}$

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

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq c_1 n^d e^{-c_2 n\epsilon^2}$$

for some c_1 and c_2. Use this fact, together with the results from (a) and (b) to show that $\sup_a |\widehat{R}_n(a) - R(a)| < \epsilon$ 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.

$$\begin{aligned}
P(sup_{-B\leq a\leq B}|\widehat{R}_n(a) - R(a)| > \epsilon) &\leq P(2B\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon) \\
&= P(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \frac{\epsilon}{2B}) \\
&\leq c_1 n^3 \exp(-c_2 n\frac{\epsilon^2}{4B^2})
\end{aligned}$$

Where we used that the VC dimension $d = 3$.

(d) Find $z(n)$ such that

$$R(\widehat{a}) \leq R(a_*) + z(n)$$

with probability at least $1 - \delta$.

Ans. We set $\delta = c_1 n^3 \exp(-c_2 n \frac{\epsilon^2}{4B^2})$ 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 $a \in [-B, B]$, $|\widehat{R}_n(a) - R(a)| \leq z'(n)$ where $z'(n) = \sqrt{\frac{4B^2}{nc_2}(3\log n + \log\frac{c_1}{\delta})}$

By definition of $\widehat{a}$ and a_*, we can conclude that with probability at least $1 - \delta$:

$$\begin{aligned}
R(\widehat{a}) - R(a_*) &= R(\widehat{a}) - \widehat{R}_n(\widehat{a}) + \widehat{R}_n(\widehat{a}) - \widehat{R}_n(a_*) + \widehat{R}_n(a_*) - R(a_*) \\
&\leq |R(\widehat{a}) - \widehat{R}_n(\widehat{a})| + |\widehat{R}_n(a_*) - R(a_*)| \\
&\leq 2z'(n)
\end{aligned}$$

where we used the fact that $\widehat{a}$ is the empirical risk minimizer and hence $\widehat{R}_n(\widehat{a}) \leq \widehat{R}_n(a_*)$.

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 $\mathsf{TV}(P_1, P_2) = \sup_A |P_1(A) - P_2(A)|$.

(a) Show that

$$\int p_1 \wedge p_2 = 1 - \mathsf{TV}(P_1, P_2)$$

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 $A = \{x : p_1(x) > p_2(x)\}$.

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

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.

(b) 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 $\theta : \mathcal{P} \to \mathbb{R}$ and let $\widehat{\theta} = \widehat{\theta}(X)$ denote an estimator of $\theta(P)$. Show that

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P |\widehat{\theta} - \theta(P)| \geq \frac{|\theta(P_1) - \theta(P_2)|}{4} \ (1 - \mathsf{TV}(P_1, P_2)) .$$

Ans. We first finitize and discretize:

$$\begin{aligned}
\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P |\widehat{\theta}(X) - \theta(P)| &\geq \inf_{\widehat{\theta}} \max_{P \in \{P_1, P_2\}} \mathbb{E}_P |\widehat{\theta}(X) - \theta(P)| \\
&\geq \inf_{Z} \max_{P_i \in \{P_1, P_2\}} P_i(Z(X) \neq i) \frac{|\theta(P_1) - \theta(P_2)|}{2} \\
&\geq \inf_{Z} [P_1(Z(X) \neq 1) + P_2(Z(X) \neq 2)] \frac{|\theta(P_1) - \theta(P_2)|}{4}
\end{aligned}$$

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, $P_1(Z^*(X) \neq 1) = \displaystyle\int_{x: p_1(x) < p_2(x)} p_1(x) dx$ and $P_2(Z^*(X) \neq 2) \geq \displaystyle\int_{x: p_2(x) < p_1(x)} p_2(x) dx$. Combining these two results, we have that $P_1(Z^*(X) \neq 1) + P_2(Z^*(X) \neq 2) = \displaystyle\int p_1 \wedge p_2 dx$

Thus, $\inf_Z [P_1(Z(X) \neq 1) + P_2(Z(X) \neq 2)] \geq \displaystyle\int p_1 \wedge p_2 dx$ 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) $\mathbb{E}[\int |\widehat{p}(x) - p(x)|^2 dx]$ 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:

$$\widehat{p}(x) = \sum_{j=1}^{m} \widehat{p}_j I(x \in B_j) \text{ where } \widehat{p}_j = \frac{m}{n} \sum_{i=1}^{n} I(X_i \in B_j)$$

and $B_1 = [0, 1/m), B_2 = [1/m, 2/m), \ldots, B_m = [(m-1)/m, 1)$. Denote its mean by $\bar{p}(x) = \mathbb{E}[\widehat{p}(x)]$.

(a) Compute the integrated square bias $\int |\bar{p}(x) - p(x)|^2 dx$ of the histogram density estimator.
Ans. We first look at $\bar{p}(x)$:

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

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

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

For each B_j, let us consider two cases.

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

$$\begin{aligned}
\int_{x \in B_j} |m P(B_j) - p(x)|^2 \, dx \;&=\; \int_{x \in B_j} \left| m \int_{y \in B_j} (p(y) - p(x)) dy \right|^2 dx \\
&\leq\; \int_{x \in B_j} \left(m \int_{y \in B_j} |p(y) - p(x)| dy \right)^2 dx \\
&\leq\; \int_{x \in B_j} \left(m \int_{y \in B_j} \frac{L}{m} dy \right)^2 dx \\
&\leq\; \int_{x \in B_j} \frac{L^2}{m^2} dx \;=\; \frac{L^2}{m^3}.
\end{aligned}$$

(2) B_j contains at least one of the N discontinuities. Using the assumption that $p(x) \leq B < \infty$, we get

$$\begin{aligned}
\int_{x \in B_j} |mP(B_j) - p(x)|^2 \, dx &= \int_{x \in B_j} \left| m \int_{y \in B_j} (p(y) - p(x)) dy \right|^2 dx \\
&\leq \int_{x \in B_j} \left(m \int_{y \in B_j} |p(y) - p(x)| dy \right)^2 dx \\
&\leq \int_{x \in B_j} \left(m \int_{y \in B_j} B dy \right)^2 dx \\
&\leq \int_{x \in B_j} B^2 dx = \frac{B^2}{m}.
\end{aligned}$$

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 $\int \mathbb{E}[|\widehat{p}(x) - \bar{p}(x)|^2] dx$.
Ans.

$$\begin{aligned}
\int \mathbb{E}[|\widehat{p}(x) - \bar{p}(x)|^2] dx &= \int \mathbb{E}\left[\left| \sum_{j=1}^{m} (\widehat{p}_j - mP(B_j)) I(x \in B_j) \right|^2 \right] dx \\
&= \sum_{j=1}^{m} \frac{\mathbb{E}[|\widehat{p}_j - mP(B_j)|^2]}{m} \\
&= \sum_{j=1}^{m} m\mathbb{E}\left[\left| \frac{\widehat{p}_j}{m} - P(B_j) \right|^2 \right] \\
&= \sum_{j=1}^{m} m\mathbb{V}\left[\frac{\sum_{i=1}^{n} I(X_i \in B_j)}{n} \right] \\
&= \sum_{j=1}^{m} \frac{m}{n} P(X \in B_j)(1 - P(X \in B_j)) \\
&\leq \sum_{j=1}^{m} \frac{m}{n} P(X \in B_j) = \frac{m}{n}.
\end{aligned}$$

(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

$$\frac{m}{n} = \frac{cNB^2}{m} \qquad \Longleftrightarrow \qquad m = B\sqrt{cN}\sqrt{n},$$

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

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

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 $\mathbf{x}^\top = [\mathbf{x}_A^\top \ \mathbf{x}_B^\top]$ 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 $p(\mathbf{x}_A|\mathbf{x}_B)$ in terms of Ω_{AA}, Ω_{AB}, Ω_{BA}, Ω_{BB}.
Ans.

$$\begin{aligned}
& \log p(\mathbf{x}_A, \mathbf{x}_B) \\
\propto\ & -\frac{1}{2}[\mathbf{x}_A^\top \ \mathbf{x}_B^\top] \begin{bmatrix} \Omega_{AA} & \Omega_{AB} \\ \Omega_{BA} & \Omega_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{bmatrix} \\
=\ & -\frac{1}{2}(\mathbf{x}_A^\top \Omega_{AA}\mathbf{x}_A + 2\mathbf{x}_B^\top \Omega_{BA}\mathbf{x}_A + \mathbf{x}_B^\top \Omega_{BB}\mathbf{x}_B) \\
=\ & -\frac{1}{2}\Big((\mathbf{x}_A + \Omega_{AA}^{-1}\Omega_{AB}\mathbf{x}_B)^\top \Omega_{AA}(\mathbf{x}_A + \Omega_{AA}^{-1}\Omega_{AB}\mathbf{x}_B) + \mathbf{x}_B^\top\big(\Omega_{BB} - \Omega_{BA}\big(\Omega_{AA}\big)^{-1}\Omega_{AB}\big)\mathbf{x}_B\Big).
\end{aligned}$$

This suggests that the marginal distribution $p(\mathbf{x}_B)$, obtained by integrating $p(\mathbf{x}_A, \mathbf{x}_B)$ over $\mathbf{x}_A$, is a zero mean Gaussian with inverse covariance

$$\Omega_{BB} - \Omega_{BA}\big(\Omega_{AA}\big)^{-1}\Omega_{AB},$$

which then gives that

$$p(\mathbf{x}_A|\mathbf{x}_B) \;=\; \frac{p(\mathbf{x}_A, \mathbf{x}_B)}{p(\mathbf{x}_B)} \;=\; \mathcal{N}(-\Omega_{AA}^{-1}\Omega_{AB}\mathbf{x}_B, \Omega_{AA}^{-1}).$$

(b) Show that the precision matrix of $\mathbf{x}_A$ given $\mathbf{x}_B$ does NOT depend on the value of $\mathbf{x}_B$.
Ans. From (a) we know the precision matrix of $\mathbf{x}_A$ given $\mathbf{x}_B$ is Ω_{AA}, which does not depend on the value of $\mathbf{x}_B$.

(c) Write the marginal density $p(\mathbf{x}_A)$ in terms of Ω_{AA}, Ω_{AB}, Ω_{BA}, Ω_{BB}.
Ans. Switching $\mathbf{x}_A$ and $\mathbf{x}_B$ in the derivation in (a), we get that

$$p(\mathbf{x}_A) \;=\; \mathcal{N}(\mathbf{0}, (\Omega_{AA} - \Omega_{AB}\big(\Omega_{BB}\big)^{-1}\Omega_{BA})^{-1}).$$

(d) Assume the variables in $\mathbf{x}_A$ are mutually independent of one another conditioning on $\mathbf{x}_B$. Would the variables in $\mathbf{x}_A$ be mutually independent? Why or why not?
Ans. The variables in $\mathbf{x}_A$ are mutually independent of one another conditioning on $\mathbf{x}_B$ if and only if the precision matrix of the condition distribution, which has been shown in (a) to be Ω_{AA}, is diagonal. The variables in $\mathbf{x}_A$ are mutually independent if and only if the precision matrix of the marginal, $\Omega_{AA} - \Omega_{AB}\big(\Omega_{BB}\big)^{-1}\Omega_{BA}$, is diagonal. Obviously, Ω_{AA} being diagonal does not guarantee $\Omega_{AA} - \Omega_{AB}\big(\Omega_{BB}\big)^{-1}\Omega_{BA}$ 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 λ,

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

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 $2Id$.

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 λ_0 such that if $\widehat{\beta}$ minimizes $\frac{1}{2n}||Y - X\beta||_2^2$ for $||\beta||_1 \leq t_0$ then $\widehat{\beta}$ 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:

$$\begin{aligned} Ł(\beta, \lambda) &= \frac{1}{2n}||Y - X\beta||_2^2 + \lambda(||\beta||_1 - t_0) \\ &= \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1 - \lambda t_0 \end{aligned}$$

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 (β^*, λ_0) be a pair of primal-dual optimal solution. Then by KKT conditions, it must be that subgradient of $L(\beta, \lambda_0)$ at β^* contains 0 and hence β^* is the global optimum of the optimization in $\Phi(\lambda_0)$.

Since λ_0 is the dual optimum, it must be that λ_0 optimizes $\sup_\lambda \Phi(\lambda) - \lambda t_0$.

By strong duality, we know that λ_0 is global dual optimum, and by the fact that $\Phi(\lambda) - \lambda t_0$ is strongly convex in λ, we know that λ_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.)