Stats300b Problem Set 7
Due: Thursday, March 28 at beginning of class

Exercises 7.9, 7.10, 6.3, 6.6, 8.1

**Additional question:** Approximately how long did it take you to complete this homework?

**Answer to 7.9:**
For the answer to this question, we use some extraneous knowledge, as it makes our arguments quicker and simpler. There are ways to do this without this knowledge, but it certainly is cleaner. The first result is that convex functions defined on \( \mathbb{R} \) have non-empty subgradient sets everywhere (recall Question 2.1), where for convex \( f: \mathbb{R} \to \mathbb{R} \),

\[
\partial f(x) = \{ g \in \mathbb{R} \mid f(y) \geq f(x) + g(y-x) \text{ for all } y \in \mathbb{R} \}.
\]

We also note that if \( R(\theta) = \mathbb{E}[\ell(\theta, X)] \), then \( \partial R(\theta) = \mathbb{E}[\partial_\theta \ell(\theta, X)] \), where the integral of a set-valued mapping \( T: \mathcal{X} \to 2^{\mathbb{R}} \) is defined by

\[
\int T(x) d\mu(x) := \left\{ \int t(x) d\mu(x) \mid t(x) \in T(x), t \text{ measurable} \right\}.
\]

Moreover, \( \partial_\theta \ell(\theta, x) = \text{sgn}(\theta - x) \), where \( \text{sgn}(t) = 1 \) if \( t > 0 \), \( \text{sgn}(t) = -1 \) if \( t < 0 \), and \( \text{sgn}(t) = [-1, 1] \) otherwise, and \( \theta_0 \) minimizes \( R(\theta) \) if and only if \( 0 \in \partial R(\theta_0) \).

(a) By the above calculations, we have

\[
\partial R(\theta) = \mathbb{E}[\text{sgn}(\theta - X)] = P(\theta > X) - P(\theta < X) + [-P(\theta = X), P(\theta = X)].
\]

At \( \theta_0 = \text{Med}(X) \), we have

\[
\partial R(\theta_0) = [P(X > \theta_0) - P(X \leq \theta_0), P(X \geq \theta_0) - P(X < \theta_0)] \supset [-p_0, p_0]. \tag{1}
\]

(b) This is literally \( \square \), Example 5.24.

(c) Using the inequality \( \square \), we note that

\[
R(\theta) \geq R(\theta_0) + g(\theta - \theta_0) \quad \text{for } g \in \partial R(\theta_0).
\]

If \( \theta > \theta_0 \), then taking \( g = P(X \geq \theta_0) - P(X < \theta_0) \geq p_0 \) yields

\[
R(\theta) \geq R(\theta_0) + p_0(\theta - \theta_0) = R(\theta_0) + p_0|\theta - \theta_0|.
\]

The reverse inequality is similar.

(d) We use the next answer to see that \( \sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} 0 \), the point mass at \( 0 \).

(e) There are many ways to do this. One is to use the modulus of continuity, as in class, which will yield that any \( r_n \) which is polynomial in \( n \) satisfies \( r_n(\hat{\theta}_n - \theta_0) = O_P(1) \). The other is to use a concentration inequality, which we do in this answer. Indeed, we have that the median of \( P_n \), the empirical distribution, is exactly \( \theta_0 \) if

\[
\text{card}\{i \in [n] \mid X_i < \theta_0\} < \frac{n}{2} \quad \text{and} \quad \text{card}\{i \in [n] \mid X_i > \theta_0\} < \frac{n}{2}.
\]
Letting \( A_i = 1 \{ X_i < \theta \} \) and \( B_i = 1 \{ X_i > \theta \} \), using the notation \( \overline{A}_n = \frac{1}{n} \sum_{i=1}^n A_i \) and similarly for \( B_i \), we have \( \mathbb{E}[\overline{A}_n] = P(X_1 < \theta) \leq \frac{1}{2} - p_0 \) and \( \mathbb{E}[\overline{B}_n] \leq \frac{1}{2} - p_0 \). By Hoeffding’s inequality, then,

\[
P\left( \overline{A}_n < \frac{1}{2} \right) = P\left( \overline{A}_n - \mathbb{E}[A_1] < \frac{1}{2} - \mathbb{E}[A_1] \right) \geq P\left( \overline{A}_n - \mathbb{E}[A_1] < p_0 \right) \geq 1 - e^{-2n p_0^2}.
\]

Using an identical result for \( \overline{B}_n \), we obtain by the union bound that

\[
P(\text{Med}(P_n) = \theta_0) = P\left( \overline{A}_n < \frac{1}{2}, \overline{B}_n < \frac{1}{2} \right) \geq 1 - P\left( \overline{A}_n \geq \frac{1}{2}\right) - P\left( \overline{B}_n \geq \frac{1}{2}\right) \geq 1 - 2e^{-2n p_0^2}.
\]

Thus, we have for all \( N \in \mathbb{N} \) that

\[
P(\exists n \geq N \text{ s.t. Med}(P_n) \neq \theta_0) = 2 \sum_{n \geq N} e^{-2n p_0^2} = 2e^{-2N p_0^2} \sum_{n=0}^{\infty} e^{-2n p_0^2} \approx \frac{1}{1 - e^{-2p_0^2}} \approx \frac{1}{p_0^2} e^{-2N p_0^2}.
\]

In particular, for all rates \( r_n \), we have \( r_n(\hat{\theta}_n - \theta_0) = 0 \) eventually.

\[\Box\]

**Answer to 7.10**

(a) Fix \( \theta' \) such that \( \| \theta' - \theta_0 \| > r \). Then there is some \( \theta \in [\theta_0, \theta'] \) such that \( \| \theta - \theta_0 \| = r \), that is, there is a \( t \in (0, 1) \) with

\[
\theta = t\theta_0 + (1-t)\theta', \quad \text{so} \quad h(\theta) \leq th(\theta_0) + (1-t)h(\theta').
\]

Rearranging by subtracting \( h(\theta_0) \) from both sides yields \( h(\theta) - h(\theta_0) \leq (1-t)(h(\theta') - h(\theta_0)) \). Noting that \( h(\theta_0) < h(\theta) \) and that \( t \in (0, 1) \), we thus obtain

\[
0 < h(\theta) - h(\theta_0) \leq (1-t)[h(\theta') - h(\theta_0)], \quad \text{or} \quad h(\theta') > h(\theta_0).
\]

(b) This is immediate: for any (positive) measure \( \mu \), including \( P \) and \( P_n \), we have

\[
\int F(t\theta + (1-t)\theta'; x) d\mu(x) \leq \int tF(\theta; x) + (1-t)F(\theta'; x) d\mu(x).
\]

(c) The uniqueness of \( \theta_0 \) is immediate by part (b) and (a), because \( f(\theta) \geq f(\theta_0) + \frac{\nu \beta^2}{2} > f(\theta_0) \) for all \( \theta \) with \( \| \theta - \theta_0 \| = \beta \).

(d) We have that \( \mathbb{E}[\Delta(\theta, X)] = 0 \), and that

\[
|\Delta(\theta, x)| \leq |F(\theta; x) - F(\theta_0; x)| + |f(\theta) - f(\theta_0)| \leq 2L \| \theta - \theta_0 \|,
\]

that is, \( \Delta \) is bounded by \( 2L \| \theta - \theta_0 \| \). Using the standard result that a variable \( Z \in \Theta \) is \( \frac{(b-a)^2}{4} \)-sub-Gaussian, we have that \( \Delta \) is \( 16L^2 \| \theta - \theta_0 \|^2 / 4 = 4L^2 \| \theta - \theta_0 \|^2 \) sub-Gaussian.
(e) Fix $\delta \leq \beta$ and let $\Theta_\delta = \{ \theta : \| \theta - \theta_0 \| \leq \delta \}$. Suppose that $\hat{\theta}_n$ is not within $\delta$ of $\theta_0$, that is, $\| \hat{\theta}_n - \theta_0 \| \geq \delta$. Then by part (a), there must be some $\theta_\delta \in \Theta_\delta$ such that $f_n(\theta_\delta) \leq f_n(\theta_0)$. Then

$$f_n(\theta_\delta) \leq f_n(\theta_0) = f_n(\theta_0) - f(\theta_0) + f(\theta_\delta) + f(\theta_0) - f(\theta_\delta) \leq f_n(\theta_0) - f(\theta_0) + f(\theta_\delta) - \frac{\nu}{2} \| \theta_\delta - \theta_0 \|^2.$$

Rearranging, we have

$$\frac{\nu}{2} \| \theta_\delta - \theta_0 \|^2 \leq f_n(\theta_0) - f(\theta_0) + f(\theta_\delta) - f_n(\theta_\delta) \leq \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \leq \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|.$$

In particular, if we have that

$$\| \hat{\theta}_n - \theta_0 \| \geq \delta,$$

then it must be the case that

$$\frac{\nu}{2} \delta^2 \leq \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|. \quad (2)$$

Now, let us understand this last event (2). Let $\Delta'_n$ be $\Delta_n$ with the point $x_i$ swapped for $x'_i$. Then

$$\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| - \sup_{\theta \in \Theta_\delta} |\Delta'_n(\theta)| \leq \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta) - \Delta'_n(\theta)| = \frac{1}{n} \sup_{\theta \in \Theta_\delta} |(F(\theta; x_i) - f(\theta)) - (F(\theta; x'_i) - f(\theta))| \leq \frac{2L}{n} \sup_{\theta \in \Theta_\delta} \|\theta - \theta_0\| = \frac{2L}{n} \delta.$$

That is, $\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|$ satisfies bounded differences, and we have

$$\mathbb{P} \left( \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \geq \mathbb{E} \left[ \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \right] + t \right) \leq \exp \left( -\frac{nt^2}{2L^2\delta^2} \right).$$

Thus, we control the expected supremum of the errors $\Delta_n(\theta)$ over the neighborhood $\Theta_\delta$. We note by our standard symmetrization inequalities, and the fact that $\theta \mapsto \sqrt{n} \Delta_n(\theta)$ is $4L^2 \|\theta - \theta_0\|^2$-sub-Gaussian process, that

$$\mathbb{E} \left[ \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \right] \leq \frac{CL}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\Theta_\delta, \|\cdot\|, \epsilon)} d\epsilon,$$

where $N$ denotes the covering numbers of $\Theta_\delta$ in norm $\|\cdot\|$ at radius $\epsilon$ as usual. But then we have $\log N(\Theta_\delta, \|\cdot\|, \epsilon) \leq d \log (1 + \frac{\delta}{\epsilon})$ for $\epsilon < \delta$, and 0 otherwise, so that

$$\mathbb{E} \left[ \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \right] \leq \frac{CL\sqrt{d}}{\sqrt{n}} \int_0^\delta \sqrt{\log \left( 1 + \frac{\delta}{\epsilon} \right)} d\epsilon \leq C \frac{L\sqrt{d}\delta}{\sqrt{n}}.$$

That is, for some numerical constant $C$, we have

$$\mathbb{P} \left( \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \geq C \frac{L\delta}{\sqrt{n}} (\sqrt{d} + t) \right) \leq e^{-t^2} \quad (3)$$

3
for all \( t \geq 0 \).

On the event that \( \sup_{\theta \in \Theta} |\Delta_n(\theta)| \leq \frac{L\sqrt{3}}{\sqrt{n}} \delta + \frac{\sqrt{2}L}{n} t \), which occurs with probability at least \( 1 - e^{-t^2} \) by inequality (3), we have by inequality (2)

\[
\delta^2 \leq C \frac{L}{\nu \sqrt{n}} \left( \sqrt{d} + t \right) \delta,
\]

where \( C < \infty \) is an absolute constant, as long as \( \delta \leq \beta \) (where \( \beta \) is the radius of strong convexity). Setting \( \sigma = CL\sqrt{d}/\nu \sqrt{n} \), that

\[
\delta \leq \sigma (1 + t).
\]

That is,

\[
P \left( \| \hat{\theta}_n - \theta_0 \| \leq C \frac{L}{\nu \sqrt{n}} \left( \sqrt{d} + t \right) \right) \geq 1 - e^{-t^2}
\]

so long as \( \frac{L}{\nu \sqrt{n}} (\sqrt{d} + t) \leq c\beta \), where \( c > 0 \) is a numerical constant.

---

**Answer to 6.5:**

(a) Assume \( t = \mathbb{E}[|X|^p]^{1/p} > 0 \). We certainly have \( \mathbb{E}[\psi(|X|/t)] = 1 \). Any \( t' < t \) gives \( \mathbb{E}[\psi(|X|/t')] < 1 \).

(b) Question 2.7 gives the result almost immediately. Letting \( v(X) = \|X\|_\psi \), we have

\[
\mathbb{E}[\psi(|X|/t)] \leq 1 \text{ iff } t \mathbb{E}[\psi(|X|/t)] - t \leq 0,
\]

and the latter is jointly convex in \( t \) and \( X \) as it is a perspective transform. Thus

\[
v(X) = \inf_{t > 0} \{ t \mid t \mathbb{E}[\psi(|X|/t)] - t \leq 0 \},
\]

which is the optimal value function as in Q. 2.7.

(c) If \( h \) is convex, symmetric, and homogeneous, then

\[
h(x + y) = 2h \left( \frac{1}{2} x + \frac{1}{2} y \right) \leq 2h(x/2) + 2h(y/2) = h(x) + h(y),
\]

so that \( h \) satisfies the triangle inequality. Homogeneity gives that \( h(0) = 0 \), and symmetry gives that \( h(tx) = |t|h(x) \) for \( t \in \mathbb{R} \). Thus (i) implies (ii). The converse only requires that we prove \( h \) is convex (the other properties are definitions of a seminorm). We have

\[
h(\lambda x + (1 - \lambda)y) \leq h(\lambda x) + h((1 - \lambda)y) = \lambda h(x) + (1 - \lambda)h(y)
\]

by the triangle inequality.

(d) Clearly \( \|X\|_\psi \) is homogeneous, symmetric, convex, and satisfies \( \|0\|_\psi = 0 \). All we need to show is that \( \|X\|_\psi > 0 \) if \( X \) is not identically 0. Let \( \epsilon > 0 \) be such that \( \mathbb{P}(|X| \geq \epsilon) > 0 \) and let \( \mathcal{E} = \{|X| \geq \epsilon\} \), so

\[
\mathbb{E}[\psi(|X|/t)] \geq \psi(\epsilon/t) \mathbb{P}(\mathcal{E}).
\]

Taking \( t \downarrow \infty \) gives \( \psi(\epsilon/t) \to \infty \) (as \( \psi \) is convex it must grow at least linearly), so that there is some \( t_0 > 0 \) such that \( \mathbb{E}[\psi(|X|/t)] > 1 \) for all \( t < t_0 \). Thus \( \|X\|_\psi > 0 \).


Answer to 6.6:

(a) By Chernoff bounds, we have

\[ P(|X| \geq t) = P(e^{\lambda |X|^q} \geq e^{\lambda t^q}) \leq \frac{\mathbb{E}[e^{\lambda |X|^q}]}{e^{\lambda t^q}} \]

for all \( \lambda \geq 0 \). Taking \( \lambda = \|X\|^{-1}_{\psi_q} \) gives the result.

(b) We have

\[ \mathbb{E}[\max_j |X_j|^q] = \lambda \mathbb{E}[\log \exp(\max_j |X_j|^q/\lambda)] \leq \lambda \log \mathbb{E} \left[ \sum_j \exp(|X_j|^q/\lambda) \right], \]

where we used Jensen’s inequality. Take \( \lambda = \max_j \|X_j\|_{\psi_q}^q \) to get the first result. The second follows by Jensen’s (or Hölder’s) inequality, as \( \mathbb{E}[|Z|] \leq \mathbb{E}[|Z|^{1/q}] \).

(c) We use the hint, so that for \( Z \sim N(0,1) \) we have

\[ \mathbb{E}[\exp(\lambda X^2)] = \mathbb{E}[\exp(\sqrt{2\lambda}ZX)] \leq \mathbb{E}[\exp(\lambda \sigma^2 Z^2)] = \frac{1}{\sqrt{1 - 2\lambda \sigma^2}}_{+}, \]

where the inequality is because \( X \) is \( \sigma^2 \)-sub-Gaussian.

(d) Using part (a), we have

\[ \mathbb{E}[\exp(X^2/t^2)] \leq \left[ 1 - 2\sigma^2/t^2 \right]_{+}^{-1/2}, \]

and solving in \( t \) so that the right hand side is equal to 2 we take \( t^2 = 8\sigma^2/3 \). Thus \( \mathbb{E}[\exp(X^2/t^2)] \leq (1/4)^{-1/2} = 2 \).

(e) Use Lemma 5 of ?.]

Answer to 8.1:

(a) We know that \( S_\lambda(\cdot) \) is 1-Lipschitz, so that

\[ \|\hat{\theta} - \theta\|^2 = \sum_{j: \theta_j \neq 0} (\hat{\theta}_j - \theta_j)^2 + \sum_{j: \theta_j = 0} (\hat{\theta}_j - \theta_j)^2 \]

\[ \overset{(i)}{=} \sum_{j: \theta_j \neq 0} (S_\lambda(Y_j) - \theta_j)^2 + 0 \]

\[ = \sum_{j: \theta_j \neq 0} (S_\lambda(Y_j) - S_\lambda(\theta_j) + S_\lambda(\theta_j) - \theta_j)^2 \]

\[ \overset{(ii)}{\leq} 2 \sum_{j: \theta_j \neq 0} (S_\lambda(\theta_j + \epsilon_j) - S_\lambda(\theta_j))^2 + 2 \sum_{j: \theta_j \neq 0} (S_\lambda(\theta_j) - \theta_j)^2 \]

\[ \overset{(iii)}{\leq} 2 \sum_{j: \theta_j \neq 0} \epsilon_j^2 + 2k\lambda^2, \]
where equality (i) follows because if \( \theta_j = 0 \) and \( \|\varepsilon\|_\infty \leq \lambda \), then \( \hat{\theta}_j = 0 \), inequality (ii) because \((a + b)^2 \leq 2a^2 + 2b^2\) for all \(a, b\), and (iii) because \(|S_\lambda(\theta_j) - \theta_j| \leq \lambda\) and \(S_\lambda(\cdot)\) is 1-Lipschitz. As \(\sigma \|\varepsilon\|_\infty \leq \lambda\), the final sum is bounded by \(2k\lambda^2 + 2k\lambda^2 = 4k\lambda^2\) as desired.

(b) We need only show that \(\sigma \|\varepsilon\|_\infty \leq \lambda(t)\) with the desired probability. In this case, by sub-Gaussianity we have

\[
P(\sigma \|\varepsilon\|_\infty \geq \lambda) \leq 2n \exp \left( -\frac{\lambda^2}{2\sigma^2} \right) = 2n \exp \left( -\frac{2\sigma^2 \log(2n) + 2\sigma^2 t}{2\sigma^2} \right) = e^{-t},
\]

which gives the first result. Setting \(t = \log(2n)\) gives the second immediately.

\(\square\)