Stats300b Problem Set 6
Due: Thursday, February 21 at beginning of class

Exercises 7.3, 7.4, 7.5, 1.11, 7.8

Answer to 7.3:
We begin with a slightly more general statement. Let \( \|\cdot\| \) be an arbitrary norm and consider its dual norm \( \|\cdot\|_* \), where \( \|x\|_* = \sup_{\|y\| \leq 1} \langle x, y \rangle \) (and \( \|x\| = \sup_{\|y\| \leq 1} \langle x, y \rangle \) in a finite dimensional space). Then if \( F = \{ \langle \theta, x \rangle \mid \|\theta\| \leq r \} \), we have
\[
R_n(F \mid x_{1:n}) = \frac{1}{n} \mathbb{E} \left[ \sup_{\|\theta\| \leq r} \sum_{i=1}^{n} \varepsilon_i \langle x_i, \theta \rangle \right] = \frac{r}{n} \mathbb{E} \left[ \left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|_* \right].
\]

This is true for any norm. Moreover, the dual norm for the \( \ell_2 \)-norm is \( \ell_2 \), while the dual norm for the \( \ell_1 \)-norm is \( \ell_\infty \).

(a) We have by Jensen’s inequality that for any \( x_1, \ldots, x_n \in \mathbb{R}^d \),
\[
\mathbb{E} \left[ \left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|_2^2 \right] \leq \mathbb{E} \left[ \left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|_2^2 \right] = \sum_{i=1}^{n} \|x_i\|_2^2,
\]
by the independence of the random signs \( \varepsilon_i \). Thus \( R_n(F \mid x_{1:n}) \leq r n^{-1} \sqrt{\sum_{i=1}^{n} \|x_i\|_2^2} \leq M \sqrt{n} \), where we have used that \( \|x_i\|_2 \leq M \). So \( R_n(F) \leq M r \sqrt{n} \) certainly.

(b) Conditional on \( x_1, \ldots, x_n \), we must consider
\[
R_n(F \mid x_{1:n}) = \mathbb{E} \left[ \left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|_\infty \right].
\]

Let \( x_{i,j} \) denote the \( j \)th coordinate of \( x_i \in \mathbb{R}^d \). Each coordinate \( j \in \{1, \ldots, d\} \) of the vector \( \sum_{i=1}^{n} \varepsilon_i x_i \) is an \( \sum_{i=1}^{n} x_{i,j}^2 \)-sub-Gaussian random variable with mean zero, because \( \varepsilon_i \in \{-1, 1\} \) are independent signs. (Similarly, \( -\sum_{i=1}^{n} \varepsilon_i x_i \) has \( \sum_{i=1}^{n} x_{i,j}^2 \)-sub-Gaussian coordinates.) Thus, using the first homework (that maxima of sub-Gaussians have small expectations), we have
\[
\mathbb{E} \left[ \max_{j \leq d} \left\{ \sum_{i=1}^{n} \varepsilon_i x_{i,j}, -\sum_{i=1}^{n} \varepsilon_i x_{i,j} \right\} \right] \leq \sqrt{\log(2d)} \max_{j \leq d} \sqrt{\sum_{i=1}^{n} x_{i,j}^2} \leq M \sqrt{n \log(2d)}.
\]

In particular, we find that \( R_n(F) \leq \frac{M r \sqrt{\log(2d)}}{\sqrt{n}} \).

Answer to 7.4:
(a) Yes. Intuitively, it will pick \( \hat{\theta}_n \) such that (roughly) \( \langle \hat{\theta}, X_i \rangle Y_i > 0 \) for most \( i \).

(b) Let \( F = \{ f(x) = \langle \theta, x \rangle \mid \|\theta\|_2 \leq r \} \), and then by the contraction inequality, we have that
\[
R_n(\{ m_\theta \}_{\theta \in \Theta}) = R_n(\phi \circ F) \leq R_n(F) \leq \frac{M r}{\sqrt{n}},
\]
where we have used Q4. Now, we note that
\[
\sup_{\theta \in \Theta, x \in \mathcal{X}, y \in \{-1, 1\}} |\phi(y^\top x) - \phi(0)| \leq \sup_{\theta \in \Theta, x \in \mathcal{X}, y \in \{-1, 1\}} |y^\top x| = Mr.
\]

Now we use Q3, which thus implies that
\[
\mathbb{P}\left(\sup_{\theta \in \Theta} |P_n \theta - P \theta| \geq 2R_n(F) + t \right) \leq 2\exp(-cnM^2t^2)
\]
for a numerical constant $c$. In particular, taking
\[
t = \frac{Mr}{\sqrt{cn} \log \frac{1}{\delta}} \quad \text{yields} \quad \mathbb{P}\left(\sup_{\theta \in \Theta} |P_n \theta - P \theta| \geq \frac{Mr}{\sqrt{n}} \left(2 + C \log^2 \frac{1}{\delta}\right)\right) \leq \delta,
\]
where $C$ is a numerical constant. This is evidently independent of the dimension, making it much sharper than Question 7.1, which scaled with dimension $d$.

(c) We apply a completely parallel derivation to find that
\[
\mathbb{P}\left(\sup_{\theta \in \Theta} |P_n \theta - P \theta| \geq \frac{2Mr\sqrt{\log(2d)}}{\sqrt{n}} + t \right) \leq \exp\left(-cnM^2t^2\right).
\]
So we may take
\[
\epsilon_n(\delta, d, r, M) = c \frac{Mr}{\sqrt{n}} \left[\sqrt{\log(2d)} + \sqrt{\log \frac{1}{\delta}}\right]
\]
to achieve the result. Again, this scales only logarithmically in the dimension, which is substantially sharper than the results of Question 7.1.

\[\square\]

**Answer to 7.5**

(a) It is clear that for any $f(x) = 1\{x \leq t\}$ and $g(x) = 1\{g(x) \leq t'\}$, where $t' \geq t$, we have $\int (f - g)^2dP = \int 1\{t < x \leq t'\}dP(x) = P(t < X \leq t')$. So w.l.o.g. we take $\epsilon < 1$. Fix an arbitrary distribution $P$ with cdf $F(t) = P(X \leq t)$, let $K(\epsilon) = \lceil 1/\epsilon \rceil - 1$, and for $k = 1, 2, \ldots, \lceil 1/\epsilon \rceil - 1$, set
\[
t_k := F^{-1}(k\epsilon) = \inf \{t : F(t) \geq k\epsilon\} = \inf \{t : P(X \leq t) \geq k\epsilon\}.
\]
Note that for distributions with point masses, we may have $t_k = t_{k+1}$ for some $k$. We have that
\[
k\epsilon \leq P(X \leq t_k) \quad \text{and} \quad P(X < t_k) \leq k\epsilon
\]
by the right-continuity of the CDF (i.e. $t \mapsto P(X \leq t)$ is right-continuous). Now, define the collection of functions $f_{k, \leq}(x) = 1\{x \leq t_k\}$ and $f_{k, <}(x) = 1\{x < t_k\}$, of which there are evidently $2K(\epsilon)$. Then for any $t \in \mathbb{R}$, we have that either (i) $t < t_1$, (ii) $t > t_{K(\epsilon)}$, (iii) there exists $k \in \{1, \ldots, K(\epsilon)\}$ such that $t_{k-1} < t < t_k$, or (iv) we have $t = t_k$ for some $k \in \{1, \ldots, K(\epsilon)\}$. In case (i), we have
\[
\int (f - f_{1, <})^2dP = \int 1\{t \leq x < t_1\}dP(x) \leq P(X < t_1) \leq \epsilon.
\]
In case (ii), we similarly have \( \int (f - f_{K(\epsilon), \leq})^2 dP \leq P(X > t_{K(\epsilon)}) \leq \epsilon \). In case (iii), we have

\[
\int (f - f_{k-1, \leq})^2 dP = \int 1 \{t_{k-1} < x \leq t\} dP(x) \leq \int 1 \{t_{k-1} < x < t_k\} dP(x) = P(t_{k-1} < X < t_k) = P(X < t_k) - P(X \leq t_{k-1}) \leq \epsilon.
\]

Finally, in case (iv) we certainly have \( \int (f - f_{k, \leq})^2 dP = 0 \). That is, our collection \( \{f_{k, <}, f_{k, \leq}\} \leq K(\epsilon) \) satisfies

\[
\min_k \min \{\|f - f_{k, <}\|_{L_2(P)}, \|f - f_{k, <}\|_{L_2(P)}\} \leq \sqrt{\epsilon}.
\]

Replacing \( \epsilon \) with \( \epsilon^2 \), we obtain the covering number bound

\[
N(\mathcal{F}, L_2(P), \epsilon) \leq 2 \left[ \frac{1}{\epsilon^2} \right] - 2 \quad \text{for } \epsilon < 1.
\]

Taking logarithms gives the desired result.

(b) We will apply Dudley’s entropy integral. First, we note that for any two functions \( f, g : \mathbb{R} \to \mathbb{R} \), we have that for a fixed collection \( X, t \)−

\[
d\epsilon \quad \text{we have that as we vary }
\]

\[
\text{We can also show this bound directly without using chaining. Notice }
\]

\[
\text{that as we vary } x \in \mathbb{R}, \text{the tuple } (1 \{x_1 \leq t\}, \ldots, 1 \{x_n \leq t\}) \text{ can take on only } n + 1 \text{ values. If we sort } x_i \text{ as } x_{(1)} \leq \cdots \leq x_{(n)}, \text{we have } (\epsilon_{(1)}, \ldots, \epsilon_{(n)}) \overset{d}{=} (\epsilon_1, \ldots, \epsilon_n). \text{ So we get }
\]

\[
\sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i 1 \{x_i \leq t\} \right| = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_{(i)} 1 \{x_{(i)} \leq t\} \right| \overset{d}{=} \frac{1}{n} \max_{1 \leq i \leq n} |S_n|,
\]

where \( S_n \) is a symmetric random walk on \( \mathbb{Z} \). By the reflection principle, we have that for any \( t \in \mathbb{Z}_{>0}, \)

\[
\mathbb{P} \left( \max_{1 \leq i \leq n} S_n \geq t \right) \leq 2\mathbb{P}(S_n \geq t) \quad \text{and} \quad \mathbb{P} \left( \max_{1 \leq i \leq n} \{-S_n\} \leq -t \right) \leq 2\mathbb{P}(S_n \leq -t),
\]

\[
\mathbb{P} \left( \max_{1 \leq i \leq n} |S_n| \right) \leq 2\mathbb{P}(S_n \geq t) + 2\mathbb{P}(S_n \leq -t).
\]
giving us \( \mathbb{P}(\max_{1 \leq i \leq n} |S_n| \geq t) \leq 4\mathbb{P}(|S_n| \geq t) \). Applying the identity \( \mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n) \) for integer valued r.v. \( X \geq 0 \), we get
\[
\mathbb{E} \left[ \max_{1 \leq i \leq n} |S_n| \right] \leq 4\mathbb{E}(|S_n|) \leq 4 \sqrt{\mathbb{E} \left[ \left( \sum_{i=1}^{n} \varepsilon_i \right)^2 \right]} = 4\sqrt{n}.
\]

So the desired bound holds with \( C = 4 \).

(c) This is just \( C^2 \) part \( f \).

\[\square\]

Answer to \[1.11\] We have
\[
0 \leq (1 - \theta)\mathbb{E}[X] = \mathbb{E}[X - \theta\mathbb{E}[X]] \leq \mathbb{E}[1 \{X \geq \theta\mathbb{E}[X]\} (X - \theta\mathbb{E}[X])] \leq \mathbb{P}(X \geq \theta\mathbb{E}[X])^2 \mathbb{E}((X - \theta\mathbb{E}[X])^2)^2,
\]
while
\[
\mathbb{E}[(X - \theta\mathbb{E}[X])^2] = \mathbb{E}[(X - \mathbb{E}[X] + (1 - \theta)\mathbb{E}[X])^2] = \text{Var}(X) + (1 - \theta)^2\mathbb{E}[X]^2.
\]
Thus
\[
\mathbb{P}(X \geq \theta\mathbb{E}[X]) \geq \frac{(1 - \theta)^2\mathbb{E}[X]^2}{\text{Var}(X) + (1 - \theta)^2\mathbb{E}[X]^2} = (1 - \theta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X]^2 - \theta(2 - \theta)\mathbb{E}[X]^2}.
\]

\[\square\]

Answer to \[7.8\]

(a) We use the Paley-Zygmund inequality (Question \[1.11\]), which states that
\[
\mathbb{P}(\langle v, X \rangle \geq \theta\mathbb{E}[\langle v, X \rangle]) \geq (1 - \theta)^2 \frac{\mathbb{E}[\langle v, X \rangle]^2}{\mathbb{E}[\langle v, v \rangle]v^T\Sigma v - \theta(2 - \theta)\mathbb{E}[\langle v, X \rangle]^2}.
\]
Using that \( \mathbb{E}[\langle v, v \rangle] \geq \kappa \sqrt{v^T\Sigma v} \) for a constant \( \kappa > 0 \) and setting \( \theta = \frac{1}{2} \) above, we have
\[
\mathbb{P}(\langle v, X \rangle \geq \frac{\kappa}{2} \|v\|_\Sigma) \geq \frac{\kappa^2 v^T\Sigma v}{4v^T\Sigma v - (3/4)\kappa v^T\Sigma v} = \frac{\kappa^2}{4 - 3\kappa}.
\]

(b) Let \( \|v\|^2_\Sigma = v^T\Sigma v \) for shorthand, and recall that \( \lambda_{\text{min}}(\Sigma) = \inf_{v \in S^{d-1}} \sqrt{v^T\Sigma v} \). The set of half-planes in \( \mathbb{R}^d \) has VC-dimension at most \( d + 1 \), while \( \|v\|_\Sigma \geq \lambda_{\text{min}} \) for all \( v \in S^{d-1} \). Thus if we define the random variable
\[
B_i(v) = 1 \left\{ \langle v, X_i \rangle \geq \frac{\kappa}{2} \lambda_{\text{min}} \right\} + 1 \left\{ \langle v, X_i \rangle \leq -\frac{\kappa}{2} \lambda_{\text{min}} \right\},
\]
then
\[
\mathbb{E}[B_i(v)] \geq \frac{\kappa^2}{4 - 3\kappa}
\]
by the first part of the question. Using the VC-dimension bounds from class, for a numerical constant \( C \), we have
\[
\mathbb{P} \left( \exists v \in S^{d-1} \text{ s.t. } \frac{1}{n} \sum_{i=1}^{n} B_i(v) - \mathbb{E}[B_i(v)] \leq -C \sqrt{\frac{d}{n}} - t \right) \leq e^{-2nt^2}.
\]
Written differently,

\[ \Pr \left( \exists \, v \in \mathbb{S}^{d-1} \text{ s.t. } \frac{1}{n} \text{card} \{ i \in [n] \mid \langle v, X_i \rangle^2 \geq \kappa^2 \lambda_{\min}(\Sigma)/4 \} \right) \leq \frac{\kappa^2}{4 - 3\kappa} - C \sqrt{\frac{d}{n} - t} \leq e^{-2nt^2}. \]

On the complement of the event within the probability above, we have

\[ v^T \tilde{\Sigma}_n v \geq \left[ \frac{\kappa^2}{4 - 3\kappa} - C \sqrt{\frac{d}{n} - t} \right] + \frac{\kappa^2}{4} \lambda_{\min}(\Sigma) \]

for all \( v \in \mathbb{R}^d \).

\( \Box \)