UC Berkeley
Department of Electrical Engineering and Computer Science
Department of Statistics

EECS 281B / STAT 241B
ADVANCED TOPICS IN STATISTICAL LEARNING THEORY

Solutions 4
Spring 2009

Solution 4.1

By the definition of shatter coefficients, if $|\mathcal{A}| < \infty$,

$$s(\mathcal{A}, n) = \max_{z_1,\ldots,z_n} |\{A \cap \{z_1, \ldots, z_n\} \mid A \in \mathcal{A}\}| \leq \min(2^n, |\mathcal{A}|).$$

Therefore, for any n such that $s(\mathcal{A}, n) = 2^n$, we know that $n \leq \log_2 |\mathcal{A}|$. Since the VC dimension is defined as the maximum of such n, we conclude

$$V_{\mathcal{A}} \leq \log_2 |\mathcal{A}|.$$

The following trivial example shows that these upper bounds are tight. Let $\mathcal{A} = \mathcal{P}(\Omega)$ where $\Omega = \{c_1, \ldots, c_m\}$, hence $|\mathcal{A}| = 2^m$.

- If $n \leq m$, we choose $z_i = c_i, i = 1\ldots, n$, then

$$s(\mathcal{A}, n) \geq |\{A \cap \{c_1, \ldots, c_n\} \mid A \in \mathcal{A}\}| = 2^n = \min(2^n, |\mathcal{A}|).$$

- If $n > m$, we choose $z_i = c_i, i = 1, \ldots, m$ and arbitrary z_i for $i = m + 1, \ldots, n$, then

$$s(\mathcal{A}, n) \geq |\{A \cap \{c_1, \ldots, c_m, z_{m+1}, \ldots, z_n\} \mid A \in \mathcal{A}\}| = 2^m = \min(2^n, |\mathcal{A}|).$$

Moreover, $V_{\mathcal{A}} = \sup\{n \mid s(\mathcal{A}, n) = 2^n\} = m = \log_2 |\mathcal{A}|$.

Solution 4.2

(a) We will show that there exists a set of d points that can be shattered by $\mathcal{A}$, but for any $d + 1$ points there is a subset of it that can not be picked up by sets in $\mathcal{A}$.

  - Consider a set of d points that are standard basis of $\mathbb{R}^d$: $\{e_i \in \mathbb{R}^d \mid i = 1, \ldots, d\}$ where e_i has its i-th coordinate to be 1 and others to be 0s. For any subset S of these d points, we define $a_i = \mathbb{I}[e_i \in S]$ for $i = 1, \ldots, d$. It is easy to see that $A = (-\infty, a_1] \times (-\infty, a_2] \times \ldots \times (-\infty, a_d]$ picks out exactly the points (if any) in S. Hence $\mathcal{A}$ can shatter these d points.
  - Consider any $d + 1$ points $\{z^1, \ldots, z^{d+1}\}$ in $\mathbb{R}^d$, for $j = 1, \ldots, d$, define

$$j^+ = \arg\max_{i=1,\ldots,d+1} z^i_j.$$

  That is, z^{j^+} has the largest j-th coordinate. Now consider the subset $S = \{z^{j^+} \mid j = 1, \ldots, d\}$ that contains these "extreme" points. Note that $|S| \leq d$ and by construction of S all the other points are within $(-\infty, z^{1^+}_1] \times (-\infty, z^{2^+}_2] \times \ldots \times (-\infty, z^{d^+}_d]$. Therefore, there is no way to pick out exactly the points in S by sets in $\mathcal{A}$.

(b) We will show that there exists a set of $2d$ points that can be shattered by $\mathcal{A}$, but for any $2d+1$ points there is a subset of it that can not be picked up by sets in $\mathcal{A}$.

- Consider a set of $2d$ points as follows: $\{e_i^+ \in \mathbb{R}^d \mid i = 1, \ldots, d\} \cup \{e_i^- \in \mathbb{R}^d \mid i = 1, \ldots, d\}$ where e_i^+ (respectively, e_i^-) has its i-th coordinate to be 1 (respectively, -1) and others to be 0s. For any subset S of these $2d$ points, we define $\{a_i, b_i, i = 1 \ldots, d\}$ as follows (ϵ is an arbitrary positive number smaller than 1):

  $$b_i = \begin{cases} -1-\epsilon & \text{if } e_i^- \in S \\ -\epsilon & \text{otherwise} \end{cases} \qquad a_i = \begin{cases} 1 & \text{if } e_i^+ \in S \\ \epsilon & \text{otherwise} \end{cases}$$

  It is easy to see that $A = (b_1, a_1] \times (b_2, a_2] \times \ldots \times (b_d, a_d]$ picks out exactly the points (if any) in S. Hence $\mathcal{A}$ can shatter these $2d$ points.
- Consider any $2d+1$ points $\{z^1, \ldots, z^{2d+1}\}$ in $\mathbb{R}^d$, for $j = 1, \ldots, d$, define

  $$j^+ = \underset{i=1,\ldots,2d+1}{\arg\max}\ z_j^i, \qquad j^- = \underset{i=1,\ldots,2d+1}{\arg\min}\ z_j^i.$$

  That is, z^{j^+} (respectively, z^{j^-}) has the largest (respectively, smallest) j-th coordinate. Now consider the subset $S = \{z^{j^+}, z^{j^-} \mid j = 1, \ldots, d\}$ that contains these "extreme" points. Note that $|S| \leq 2d$ and by construction of S all the other points are within the closed rectangle $[z_1^{1^-}, z_1^{1^+}] \times [z_2^{2^-}, z_2^{2^+}] \times \ldots \times [z_d^{d^-}, z_d^{d^+}]$. Therefore, there is no way to pick out exactly the points in S by sets in $\mathcal{A}$.

(c) We claim $V_{\mathcal{A}} = 3$ in this case. First, we show that there exists a set of 3 points in $\mathbb{R}^2$ that can be shattered by the class of closed balls in $\mathbb{R}^2$. Figure 1 shows 8 closed balls that can pick out all subsets of $\{a, b, c\}$.

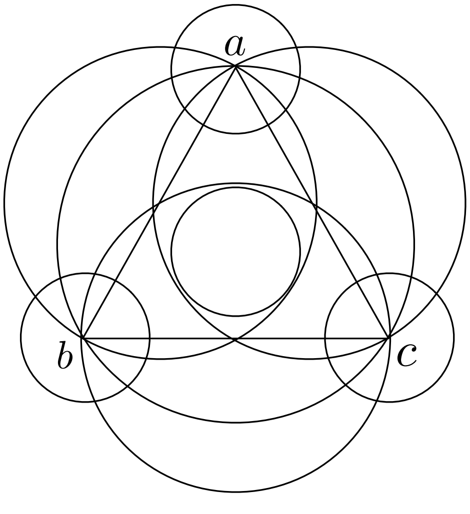

Figure 1: $\{a, b, c\}$ can be shattered by closed balls in $\mathbb{R}^2$

Next, we would like to show that no 4 points in $\mathbb{R}^2$ can be shattered by the class of all closed balls in $\mathbb{R}^2$. Radon's theorem says that any set of 4 points $\{a, b, c, d\}$ in $\mathbb{R}^2$ can be partitioned into two disjoint sets whose convex hulls don't have empty intersection. There are two cases to consider:

- WLOG, $\{a, b, c, d\}$ are partitioned into $\{a, b, c\}$ and $\{d\}$, and the convex hull of $\{a, b, c\}$ contains $\{d\}$. Hence, any closed ball containing the triple $\{a, b, c\}$ has to include the singleton $\{d\}$. There is no way to pick out $\{a, b, c\}$ without including $\{d\}$.
- WLOG, $\{a, b, c, d\}$ are partitioned into $\{a, b\}$ and $\{c, d\}$, and the line segment $a - b$ intersects with the line segment $c - d$. If there exists a pair of closed balls B_1 and B_2 such that B_1 contains only $a - b$ and B_2 contains only $c - d$, then the symmetric difference of B_1 and B_2 consists of 4 disjoint parts, each of which contains a distinct singleton from $\{a, b, c, d\}$. However, the symmetric difference of two closed balls in $\mathbb{R}^2$ can only consist of at most 2 disjoint parts. Therefore, at most one of the pairs $\{a, b\}$ and $\{c, d\}$ can be picked out by the class of closed balls.

In both cases, there exists a subset of $\{a, b, c, d\}$ that can not be picked up by the class of all closed balls. Therefore, $V_{\mathcal{A}} = 3$.

Solution 4.3

(a)

$$\begin{aligned}
\mathbb{P}[Z \geq n - k] &= \sum_{i=n-k}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n \\
&= \frac{1}{2^n} \sum_{i=n-k}^{n} \binom{n}{n-i} \\
&= \frac{1}{2^n} \sum_{i=0}^{k} \binom{n}{i}.
\end{aligned}$$

Multiplying both sides by 2^n gives the desired result.

(b) It is a straightforward calculation showing that the moment generating function of Z is

$$\mathbb{E}[e^{sZ}] = \sum_{i=0}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n (e^s)^i = \left(\frac{1 + e^s}{2}\right)^n.$$

Applying the Chernoff bound, for any $s \geq 0$,

$$\begin{aligned}
\sum_{i=0}^{k} \binom{n}{i} &= 2^n \mathbb{P}[Z \geq n - k] \\
&\leq 2^n e^{-s(n-k)} \mathbb{E}[e^{sZ}] \\
&= e^{-s(n-k)} (1 + e^s)^n \\
&= \exp\left\{n \log(1 + e^s) - s(n - k)\right\} \\
&\leq \exp\left\{-n \log(k/n) - (n - k) \log((n - k)/k)\right\} \\
&\leq \exp\left\{-k \log(k/n) - (n - k) \log((n - k)/n)\right\} \\
&= \exp\{nh(k/n)\},
\end{aligned}$$

where in the second inequality we set $s = \log((n-k)/k) \geq 0$. The last inequality holds because $k \leq n$.

Solution 4.4

(a) It is obvious that V_i is a measurable function of $Z_1, \dots, Z_i$. Moreover, for $i > 0$,

$$\begin{aligned}
\mathbb{E}[V_{i+1} \mid Z_1, \dots, Z_i] &= \mathbb{E}\left[\mathbb{E}[V \mid Z_1, \dots Z_{i+1}] - \mathbb{E}[V \mid Z_1, \dots Z_i] \mid Z_1, \dots, Z_i\right] \\
&= \mathbb{E}\left[\mathbb{E}[V \mid Z_1, \dots Z_{i+1}] \mid Z_1, \dots, Z_i\right] - \mathbb{E}[V \mid Z_1, \dots Z_i] \\
&= \mathbb{E}[V \mid Z_1, \dots Z_i] - \mathbb{E}[V \mid Z_1, \dots Z_i] \\
&= 0.
\end{aligned}$$

In the next-to-last step, we use the tower property of conditional expectation.

(b) Notice that $V_n = \mathbb{E}[V \mid Z_1, \dots Z_n] - \mathbb{E}[V \mid Z_1, \dots Z_{n-1}] = V - \mathbb{E}[V \mid Z_1, \dots Z_{n-1}]$ and $\sum_{i=1}^n V_i$ is a telescoping sum, hence $\sum_{i=1}^n V_i = V$. Another useful fact is that $\mathbb{E}[V_i] = 0$ for each $i = 1, \dots, n$. Therefore,

$$\begin{aligned}
\operatorname{var}(V) &= \operatorname{var}(\sum_{i=1}^n V_i) \\
&= \sum_{i=1}^n \operatorname{var}(V_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{cov}(V_i V_j) \\
&= \sum_{i=1}^n \operatorname{var}(V_i) + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}[V_i V_j].
\end{aligned}$$

For each pair $i < j$, $\mathbb{E}[V_i V_j] = \mathbb{E}[\mathbb{E}[V_i V_j \mid Z_1, \dots, Z_{j-1}]] = \mathbb{E}[V_i \mathbb{E}[V_j \mid Z_1, \dots, Z_{j-1}]] = 0$. The second step is valid since $V_i \in \sigma(Z_1, \dots, Z_i) \subseteq \sigma(Z_1, \dots, Z_{j-1})$ and the last step follows from the fact that $\{V_i\}$ is a MDS with respect to $\{Z_i\}$. Hence the second term in the decomposition above is 0 and $\operatorname{var}(V) = \sum_{i=1}^n \operatorname{var}(V_i)$.

(c) We claim that $\operatorname{var}[V_i \mid Z_1, \dots, Z_{i-1}] \leq c_i^2/4$. As shown in the proof of bounded difference inequality, $L_i \leq V_i \leq U_i$ and $U_i - L_i \leq c_i$ almost surely, where

$$\begin{aligned}
L_i &= \inf_t \left\{\mathbb{E}\left[f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}, t\right] - \mathbb{E}\left[f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}\right]\right\} \\
U_i &= \sup_t \left\{\mathbb{E}\left[f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}, t\right] - \mathbb{E}\left[f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}\right]\right\}.
\end{aligned}$$

Applying Jensen's inequality, we get

$$\begin{aligned}
V_i^2 &= \left[\frac{V_i - L_i}{U_i - L_i} U_i + \frac{U_i - V_i}{U_i - L_i} L_i\right]^2 \\
&\leq \frac{V_i - L_i}{U_i - L_i} U_i^2 + \frac{U_i - V_i}{U_i - L_i} L_i^2 \\
&= (U_i + L_i) V_i - U_i L_i.
\end{aligned}$$

Now take the expectation on both sides (conditioning on $Z_1, \dots, Z_{i-1}$) and notice that both L_i and U_i are measurable functions of $Z_1, \dots, Z_{i-1}$,

$$\begin{aligned}
\mathbb{E}[V_i^2 \mid Z_1, \dots, Z_{i-1}] &\leq \mathbb{E}[(U_i + L_i)V_i - U_i L_i \mid Z_1, \dots, Z_{i-1}] \\
&= (U_i + L_i)\mathbb{E}[V_i \mid Z_1, \dots, Z_{i-1}] - U_i L_i \\
&= -U_i L_i \\
&\leq \frac{(U_i - L_i)^2}{4} \\
&\leq \frac{c_i^2}{4}.
\end{aligned}$$

Therefore $\mathrm{var}[V_i \mid Z_1, \dots, Z_{i-1}] = \mathbb{E}[V_i^2 \mid Z_1, \dots, Z_{i-1}] \leq \frac{c_i^2}{4}$, which implies $\mathrm{var}(V_i) \leq \frac{c_i^2}{4}$. Plugging the bound into part (b) gives the desired bound.

Solution 4.5

To simply the notation, we denote $a = (a^1, \dots, a^k) \in (\mathbb{R}^d)^k$ and $b = (b^1, \dots, b^k) \in (\mathbb{R}^d)^k$.

(a)

$$M(a) - \inf_{b \in (\mathbb{R}^d)^k} M(b) = \left(M(a) - \widehat{M_n}(a) \right) + \left(\widehat{M_n}(a) - \inf_{b \in (\mathbb{R}^d)^k} M(b) \right)$$

The first term $M(a) - \widehat{M_n}(a) \leq \sup_{b \in (\mathbb{R}^d)^k} \left| \widehat{M_n}(b) - M(b) \right|$. For the second term, we notice that for each $b \in (\mathbb{R}^d)^k$,

$$\widehat{M_n}(a) - M(b) \leq \widehat{M_n}(b) - M(b) \leq \sup_{b \in (\mathbb{R}^d)^k} \left| \widehat{M_n}(b) - M(b) \right|.$$

Taking the infimum over b on the left hand gives us

$$\widehat{M_n}(a) - \inf_{b \in (\mathbb{R}^d)^k} M(b) \leq \sup_{b \in (\mathbb{R}^d)^k} \left| \widehat{M_n}(b) - M(b) \right|.$$

Combining the bounds for both terms yields the desired inequality.

(b) Define $f_b(x) = \min_{j=1,\dots,k} \|x - b^j\|_2^2$. Note that since x is supported on $[-B, B]^d$ for some $B < +\infty$, $0 \leq f(x) \leq d(2B)^2 = 4dB^2$. Now we can write

$$\begin{aligned}
\sup_{b \in (\mathbb{R}^d)^k} \left| \widehat{M_n}(b) - M(b) \right| &= \sup_{b \in (\mathbb{R}^d)^k} \left| \frac{1}{n} \sum_{i=1}^n f_b(X^{(i)}) - \mathbb{E}[f_b(X)] \right| \\
&= \sup_{b \in (\mathbb{R}^d)^k} \left| \int_0^\infty \left(\frac{1}{n} \sum_{i=1}^n \mathbb{I}[f_b(X^{(i)} > t] - \mathbb{P}[f_b(X) > t] \right) dt \right| \\
&\leq 4dB^2 \sup_{b \in (\mathbb{R}^d)^k, 0 \leq t \leq 4dB^2} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{I}[f_b(X^{(i)} > t] - \mathbb{P}[f_b(X) > t] \right|.
\end{aligned}$$

The last inequality comes from the fact that $f(x) \leq 4dB^2$ almost surely.

Now we define the collection of sets

$$\mathcal{A} = \{A_{b,t} \mid b \in (\mathbb{R}^d)^k, 0 \le t \le 4dB^2\},$$

where $A_{b,t} \subset \mathbb{R}^d$ is defined as $A_{b,t} = \{x \in \mathbb{R}^d \mid f_b(x) > t\}$. Then the above inequality could be written as

$$\sup_{b\in(\mathbb{R}^d)^k} \left|\widehat{M_n}(b) - M(b)\right| \le 4dB^2 \sup_{A_{b,t}\in\mathcal{A}} \left|\widehat{\mathbb{P}}_n(A_{b,t}) - \mathbb{P}(A_{b,t})\right|.$$

Applying the general Glivenko-Cantello theorem,

$$\begin{aligned}\mathbb{P}\left[\sup_{b\in(\mathbb{R}^d)^k} \left|\widehat{M_n}(b) - M(b)\right| > \epsilon\right] &\le \mathbb{P}\left[\sup_{A_{b,t}\in\mathcal{A}} \left|\widehat{\mathbb{P}}_n(A_{b,t}) - \mathbb{P}(A_{b,t})\right| > \epsilon/4dB^2\right] \\ &\le 8s(\mathcal{A}, n) \exp\left\{-\frac{n\epsilon^2}{512d^2B^4}\right\}.\end{aligned}$$

The above technique of transforming from maximizing over a class of functions to maximizing over a collection of sets (which we are familiar with) can be found in Devroye et al's book (Lemma 29.1 and Corollary 29.1). Now it remains to bound the shattering coefficient of $\mathcal{A}$. Notice that each $A_{b,t} \in \mathcal{A}$ is the complement of union of k closed balls centered at b^j with radius $t^{1/2}$. See figure 2 for illustration of a particular $A_{b,t}$ in $\mathbb{R}^2$.

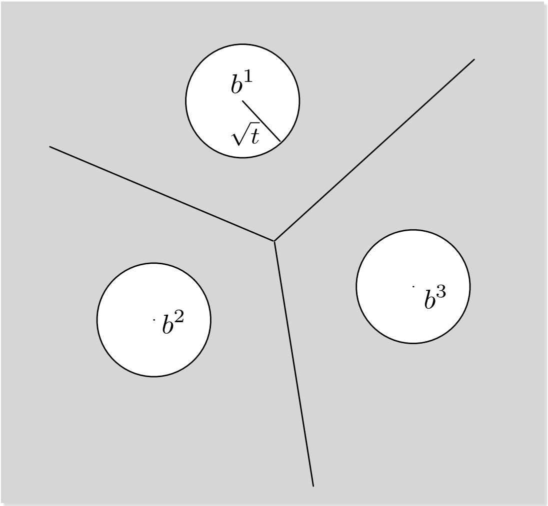

Figure 2: Voronoi diagram induced by b^1, b^2, b^3. The gray area is $A_{b,t}$.

To bound the shattering coefficient of $\mathcal{A}$, we need the following lemma:

- The class $\mathcal{B}$ of all closed balls in $\mathbb{R}^d$ has VC dimension no greater than $d + 2$ (Corollary 13.2 in Devroye et al.). Therefore, $s(\mathcal{B}, n) \le (n + 1)^{d+2}$.
- For $\mathcal{A} = \{A_1 \cup A_2 \mid A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}$, $s(\mathcal{A}, n) \le s(\mathcal{A}_1, n)s(\mathcal{A}_2, n)$ (Theorem 13.5(iv) in Devroye et al.). Therefore, the class of union of k closed balls $\mathbb{R}^d$ has shattering coefficient upper bounded by $(n + 1)^{k(d+2)}$.

  - For a class $\mathcal{A}_c$ defined as $\mathcal{A}_c = \{A^c \mid A \in \mathcal{A}\}$, then $s(\mathcal{A}, n) = s(\mathcal{A}_c, n)$ (Theorem 13.5(ii) in Devroye et al.).

  Combining these three lemmas, we conclude that $s(\mathcal{A}, n) \leq (n+1)^{k(d+2)} \leq (n+1)^{2k(d+1)}$. Hence,

$$\mathbb{P}\left[\sup_{b\in(\mathbb{R}^d)^k}\left|\widehat{M_n}(b) - M(b)\right| > \epsilon\right] \leq 8(n+1)^{2k(d+1)} \exp\left\{-\frac{n\epsilon^2}{512d^2B^4}\right\}.$$

(c) Combining part(a) and (b),

$$\begin{aligned}\mathbb{P}\left[M(a) - \inf_{b\in(\mathbb{R}^d)^k} M(b) > \epsilon\right] &\leq \mathbb{P}\left[\sup_{b\in(\mathbb{R}^d)^k}\left|\widehat{M_n}(b) - M(b)\right| > \epsilon/2\right] \\ &\leq 8(n+1)^{2k(d+1)} \exp\left\{-\frac{n(\epsilon/2)^2}{512d^2B^4}\right\} \\ &\to 0,\end{aligned}$$

as $n \to +\infty$. Therefore $M(a) \xrightarrow{p} \inf_{b\in(\mathbb{R}^d)^k} M(b)$.

Solution 4.6
Z_n is a function of $X_1, \ldots, X_n$. Changing one X_i can change Z_n by at most $1/n$. Hence Z_n satisfies the bounded difference condition with $c_i = 1/n$ for $i = 1, \ldots, n$. Applying the bounded difference inequality, for any fixed $\epsilon > 0$,

$$\mathbb{P}[|Z_n - \mathbb{E}[Z_n]| \geq \epsilon] \leq 2e^{-2n\epsilon^2},$$

which implies that

$$\sum_{n=1}^{\infty}\mathbb{P}[|Z_n - \mathbb{E}[Z_n]| \geq \epsilon] \leq \sum_{n=1}^{\infty} 2e^{-2n\epsilon^2} < +\infty.$$

By the first Borel-Cantelli lemma, $Z_n - \mathbb{E}[Z_n] \xrightarrow{a.s.} 0$. Hence $\mathbb{E}[Z_n] - Z_n = o_p(1)$. Because $Z_n = o_p(1)$, $\mathbb{E}[Z_n] = (\mathbb{E}[Z_n] - Z_n) + Z_n \longrightarrow 0$. Therefore, $Z_n = (Z_n - \mathbb{E}[Z_n]) + \mathbb{E}[Z_n] \xrightarrow{a.s} 0$.