STATS 214 Autumn 2021 Homework 2

SUNet ID: 06009508

Name: Brandon McKinzie

Collaborators: N/A

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1(a) Relation between covering and packing number

We need to show that $\forall x \in \Omega$, $\exists x' \in P$ such that $\rho(x, x') \leq \epsilon$. We can proceed with a proof by contradiction:

1. Assume $\exists x \in \Omega$ such that $\nexists x' \in P$ with $\rho(x, x') \leq \epsilon$.
2. Let $P' = P \cup \{x\}$.
3. By construction, P' must also be an ϵ-packing of Ω.
4. This contradicts the premise that P was a maximal ϵ-packing, though, since P' strictly contains P.
5. Therefore, it must be true that $\forall x \in \Omega$, $\exists x' \in P$ such that $\rho(x, x') \leq \epsilon$, and thus P is also an ϵ-cover of Ω.

Problem 1(b) Packing number upper bound

Show that

$$|B_\epsilon| \leq \left(1 + \frac{2}{\epsilon}\right)^d \tag{1}$$

We can relate the cardinality of B_ϵ to B_2^2, followed by relating it to the volume of a d-dimensional hypercube to obtain the desired result. First note that since $B_\epsilon \subset B_2^d$,

$$|B_\epsilon| \leq |B_2^d| \tag{2}$$

where B_2^d is the d-dimensional unit hypersphere. Next, note that the volume of a d-dimensional hyper*cube* is larger than the volume of a d-dimensional hypersphere. The maximal number of points we can pack into a d-dimensional hypercube, accomplished by arranging them in a d-dimensional grid with spacing ϵ, is $(\lfloor 1 + \frac{2}{\epsilon} \rfloor)^d$. Therefore,

$$\begin{aligned} |B_\epsilon| &\leq \mathrm{Vol}[B_2^d] && (3) \\ &< \left(\left\lfloor 1 + \frac{2}{\epsilon} \right\rfloor\right)^d && (4) \\ &\leq \left(1 + \frac{2}{\epsilon}\right)^d && (5) \end{aligned}$$

Problem 1(c) Covering number upper bound

Argue that for $\epsilon < 1$, we have $N(B_2^d, \epsilon) \leq (1 + 2/\epsilon)^d$.

1. Let B'_ϵ be a maximal packing of B_2^d.
2. From part (a), we know that B'_ϵ is also an ϵ-cover.
3. Also, since B_ϵ is an ϵ-packing, from part (b) we know that $B'_\epsilon \leq (1 + 2/\epsilon)^d$.
4. If we were to remove any $x \in B'_\epsilon$, then that point x would be a point in B_2^d for which

$$\rho(x, x') \geq \epsilon \qquad (\forall x' \in \{B'_\epsilon / x\}) \tag{6}$$

   which means $\{B'_\epsilon / x\}$ would not be an ϵ-covering.

5. Therefore, $|B'_\epsilon| = N(B_2^d, \epsilon) \leq (1 + 2/\epsilon)^d$.

Problem 1(d) Covering ℓ_1 ball

For any $d \geq 1$ and $1 > \epsilon > 0$, show that the ϵ-covering number of B_1^d wrt ℓ_2 distance is at most

$$\min\left\{(10d)^{\frac{5}{\epsilon^2}},\ \left(\frac{10}{\epsilon}\right)^{2d}\right\}$$

First, note that since $B_1^d \subseteq B_2^d$, we know from part (c) that $N(B_1^d, \epsilon) \leq (1+2/\epsilon)^d$. Furthermore, since $0 \leq \epsilon \leq 1$,

$$1 + \frac{2}{\epsilon} = \frac{\epsilon + 2}{\epsilon} \leq \frac{10}{\epsilon} \tag{7}$$

we have that $N(B_1^d, \epsilon) \leq \left(\frac{10}{\epsilon}\right)^{O(d)}$.

We now consider the case where $\epsilon > \sqrt{5/d}$ and derive the term on the left inside the min. Let $t := \lceil 5/\epsilon^2 \rceil$ and

$$S = \left\{\left(\frac{k_1}{t}, \ldots, \frac{k_d}{t}\right) \in B_1^d \middle| (k_i \in \mathbb{Z})\right\} \tag{8}$$

$$S' = \left\{\left(\frac{k_1}{t}, \ldots, \frac{k_d}{t}\right) \in \mathbb{R}^d \middle| \sum_{i=1}^{d} k_i \leq t\ (k_i \in \mathbb{N})\right\} \tag{9}$$

Note that $|S| \leq 2^t |S'|$, since for all $x' \in S'$, there are 2^t pre-images ($x \in S$) for a mapping from $S \to S'$. This leads to

$$|S| \leq 2^t \binom{d+t}{d} \leq 2^t (d+t)^t \leq (2d)^t \leq (2d)^{5/\epsilon^2} \leq (10d)^{5/\epsilon^2} \tag{10}$$

Therefore, if we can show that S is an ϵ-cover of B_1^d wrt ℓ_2, then we'll have shown the other half of the desired result and thus completed the proof. For any given $x \in B_1^d$, let

$$x' := \left(\frac{\lfloor x_1 t \rfloor}{t}, \ldots, \frac{\lfloor x_d t \rfloor}{t}\right) \tag{11}$$

Note that x' has the same form as the elements of S, but with $k_i := \lfloor x_i t \rfloor$. Since $x \in B_1^d$, these still sum to a number less than or equal to t. By construction of x', along with Holder's inequality, we can state

$$||x - x'||_2^2 \leq ||x - x'||_\infty ||x - x'||_1 \tag{12}$$

$$\leq \left(\frac{1}{t}\right)(1 - (-1)) \tag{13}$$

$$||x - x'||_2 \leq \sqrt{2/t} \leq \epsilon \tag{14}$$

Therefore, for $\epsilon \geq \sqrt{5/d}$, S is an ϵ-cover of B_1^d and thus

$$N(B_1^d, \epsilon) \leq |S| \leq (10d)^{5/\epsilon^2} \tag{15}$$

Combining these two bounds for N yields the desired result.

Problem 2(a) Risk Concentrates for Good Predictors

Suppose we have a fixed predictor h that achieves $L(h) \leq E$. Show that

$$Pr\left[\hat{L}(h) - L(h) \geq \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2(E+\epsilon/3)}\right) \tag{16}$$

As we've seen throughout the course, the empirical risk $\hat{L}$ can be viewed as an average of i.i.d. random variables ℓ_i[1], and that $L(h) = \mathbb{E}_{(x,y)\sim p^*}\left[\hat{L}\right]$ by definition. Therefore, we can apply Bernstein's inequality

$$\Pr\left[\hat{L}(h) - L(h) \geq \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2(\sigma^2 + (b-a)\epsilon/3)}\right) \tag{17}$$

Since we're told that $\ell(y,p) \in [0,1]$, it follows that $b-a \leq 1$. Similarly, since $L(h) = \mathbb{E}\left[\ell_i\right] \leq E$, and $L(h) \leq 1$, we know that $\mathbb{E}\left[\ell_i^2\right] \leq \mathbb{E}\left[\ell_i\right] \leq E$. Therefore

$$\sigma^2 = \mathbb{E}\left[\ell_i^2\right] - L(h)^2 \leq E - L(h)^2 \leq E \tag{18}$$

Plugging these inequalities back in yields the desired result:

$$\begin{aligned}\Pr\left[\hat{L}(h) - L(h) \geq \epsilon\right] &\leq \exp\left(\frac{-n\epsilon^2}{2(\sigma^2 + (b-a)\epsilon/3)}\right) &(19)\\ &\leq \exp\left(\frac{-n\epsilon^2}{2(E+\epsilon/3)}\right) &(20)\end{aligned}$$

[1] Throughout the homework, I'll use the shorthand ℓ_i to denote the loss on the ith training example under the current model h.

Problem 2(b) Bad Predictors Look Bad

Suppose that instead we now have another fixed predictor h' with expected risk at least $E' + e$:

$$L(h') \geq E' + \epsilon \tag{21}$$

Show that

$$Pr\left[\hat{L}(h') \leq E'\right] \leq \exp\left(\frac{-n\epsilon^2}{2(E' + 4\epsilon/3)}\right) \tag{22}$$

It will be easiest if we first derive a bound in terms of $\epsilon' = L(h') - E'$ using the same kind of logic in part (a).

$$\begin{aligned}
\Pr\left[\hat{L}(h') \leq E'\right] &= \Pr\left[\hat{L}(h') \leq L(h') - \epsilon'\right] && (23)\\
&= \Pr\left[\hat{L}(h') - L(h') \leq -\epsilon'\right] && (24)\\
&= \Pr\left[-\hat{L}(h') + L(h') \geq \epsilon'\right] && (25)
\end{aligned}$$

We can use Bernstein's inequality here since the random variable $(-\hat{L}(h'))$ can still be represented as a sum over i.i.d. bounded random variables $-\ell_i$. Not that these random variables have the same σ^2 as they did for $\hat{L}(h')$, since $\mathrm{Var}[X] = \mathrm{Var}[-X]$.Recall from part (a) that $\sigma^2 \leq \mathbb{E}\left[\ell_i^2\right] \leq \mathbb{E}\left[\ell_i\right] = L(h')$. Bernstein's inequality gives us

$$\begin{aligned}
\Pr\left[-\hat{L}(h') + L(h') \geq \epsilon'\right] &\leq \exp\left(\frac{-n\epsilon'^2}{2(\sigma^2 + \epsilon'/3)}\right) && (26)\\
&\leq \exp\left(\frac{-n(L(h') - E')^2}{2(L(h') + (L(h') - E')/3)}\right) && (27)
\end{aligned}$$

Notice that, if we can show the following inequality is true, we'd achieve the desired result:

$$\frac{(L(h') - E')^2}{L(h') + (L(h') - E')/3} \geq \frac{\epsilon^2}{E' + 4\epsilon/3} \tag{28}$$

Let

$$\begin{aligned}
g(x) &= \frac{(x - E')^2}{x + (x - E')/3} && (29)\\
\frac{d}{dx}g(x) &= \frac{2(x - E')\left(x + (x - E')/3\right) - \frac{4}{3}(x - E')^2}{\left(x + (x - E')/3\right)^2} && (30)\\
&= \frac{x - E'}{\left(x + (x - E')/3\right)^2}\left(2x + 2(x - E')/3 - \frac{4}{3}(x - E')\right) && (31)\\
&= \frac{x - E'}{\left(x + (x - E')/3\right)^2}\left(\frac{1}{3}\left(4x + 2E'\right)\right) && (32)
\end{aligned}$$

Note that this derivative is positive, and increases with $|x - E'|$. For our constraint that $x = L(h') \geq E' + \epsilon$, this means the minimum of g is achieved when $L(h') = E' + \epsilon$. In other words,

$$\min_{x \geq E'+\epsilon} g(x) = g(E' + \epsilon) = \frac{\epsilon^2}{E' + 4\epsilon/3} \tag{33}$$

Plugging this back in yields the desired result:

$$\Pr\left[\hat{L}(h') \leq E'\right] = \Pr\left[-\hat{L}(h') + L(h') \geq \epsilon'\right] \leq \exp\left(\frac{-n\epsilon^2}{2(E' + 4\epsilon/3)}\right) \tag{34}$$

Problem 2(c) Bounding Excess Risk

Suppose finite $\mathcal{H}$. Use the preceding parts to conclude that

$$Pr\left[L(\hat{h}) - L(h^*) \geq 2\epsilon\right] \leq 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2(E + 7\epsilon/3)}\right) \tag{35}$$

Recapping what we learned from parts (a) and (b) for the context of this problem:

$$\textbf{(a)} \qquad (\exists h \text{ s.t. } L(h) \leq A) \implies \Pr\left[\hat{L}(h) - L(h) \geq \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2(A + \epsilon/3)}\right) \tag{36}$$

$$\textbf{(b)} \qquad (\exists h \text{ s.t. } L(h) \geq B + 2\epsilon) \implies \Pr\left[\hat{L}(h) \leq B + \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2(B + 7\epsilon/3)}\right) \tag{37}$$

We can use part (a) for the case of $h := h^*$, since $L(h^*) = E$, to assert unconditionally that

$$\Pr\left[\hat{L}(h^*) \geq E + \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2(E + \epsilon/3)}\right) \tag{38}$$

Next, to relate with the result for part (*b*), let $\mathcal{Q} \subset \mathcal{H}$ denote the hypotheses that satisfy $L(h) \geq E + 2\epsilon$. Note that this set has at most $|\mathcal{H}| - 1$ members, since we know that $h^* \notin \mathcal{Q}$.

$$\begin{aligned}
\Pr\left[\exists h \in \mathcal{Q} \text{ s.t. } \hat{L}(h) \leq E + \epsilon\right] &\leq \sum_{h \in \mathcal{Q}} \Pr\left[\hat{L}(h) \leq E + \epsilon\right] && (39)\\
&\leq (|\mathcal{H}| - 1) \exp\left(\frac{-n\epsilon^2}{2(E + 7\epsilon/3)}\right) && (40)
\end{aligned}$$

Then we take a union over the hypothesis space consisting of $\{h^*\}$ and $\mathcal{Q}$ to obtain the desired result.

$$\begin{aligned}
\Pr\left[L(\hat{h}) - L(h^*) \geq 2\epsilon\right] &\leq \Pr\left[\hat{L}(h^*) \geq E + \epsilon\right] + \Pr\left[\exists h \in \mathcal{Q} \text{ s.t. } \hat{L}(h) \leq E + \epsilon\right] && (41)\\
&\leq \exp\left(\frac{-n\epsilon^2}{2(E + \epsilon/3)}\right) + (|\mathcal{H}| - 1) \exp\left(\frac{-n\epsilon^2}{2(E + 7\epsilon/3)}\right) && (42)\\
&\leq 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2(E + 7\epsilon/3)}\right) && (43)
\end{aligned}$$

Problem 2(d) Comparison with Hoeffding

Below are the bounds obtained in part (c) and from Hoeffding's inequality, respectively:

$$\textbf{[part (c)]} \qquad \Pr\left[L(\hat{h}) - L(h^*) \geq 2\epsilon\right] \leq 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2(E + 7\epsilon/3)}\right) \tag{44}$$

$$\textbf{[Hoeffding]} \qquad \Pr\left[L(\hat{h}) - L(h^*) \geq 2\epsilon\right] \leq 2|\mathcal{H}| \exp\left(-2n\epsilon^2\right) \tag{45}$$

$$\Delta = \frac{1}{4} - \frac{7}{6}\epsilon \tag{46}$$

When $E \leq \Delta(\epsilon)$, the RHS of 44 is less than or equal to the RHS of 45, i.e. the result for part (c) is stronger. If we consider $\epsilon \leq 0.05$, then $\Delta(\epsilon) \geq \frac{1}{20}\left(5 - 7/6\right) \approx 0.19$.

Problem 3(a) Point Mass

Suppose that $k = 1$, in which case P is a point mass at some point v. Show that

$$R_n(F) \leq \frac{1}{\sqrt{n}} \tag{47}$$

$$\begin{aligned}
R_n(\mathcal{F}) &\triangleq \mathbb{E}_{z_1,\ldots,z_n}\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right]\right] && (48)\\
&= \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(v)\right] \qquad P(v) = 1 && (49)\\
&= \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i\right|\right] \qquad f(v) \in \{-1, 1\} && (50)\\
&\leq \frac{1}{\sqrt{n}} \qquad \textbf{Sec. 4.4.3 of scribe notes} && (51)
\end{aligned}$$

where in the last step we've applied the derivation starting from equation 4.67 of the scribe notes, section 4.4.3 (Dependence of Rademacher complexity on P).

Problem 3(b) Expected Max of Sub-Gaussian Variables

Let $X_1, \ldots, X_m$ be sub-Gaussian variables with mean zero and variance proxy σ^2. Show that

$$\mathbb{E}\left[\max_{1\le i\le m} X_i\right] \le \sqrt{2\sigma^2 \log m} \tag{52}$$

We'll apply the definitions of sub-Gaussian variables, with the simplification that we'll only consider strictly positive $\lambda \in \mathbb{R}^+$:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \max_i X_i\right)\right] &\le \mathbb{E}\left[\sum_{i=1}^m \exp\left(\lambda X_i\right)\right] && (53)\\
&\le m \exp \frac{1}{2}\lambda^2\sigma^2 && (54)
\end{aligned}$$

We can use Jensen's inequality

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \max_i X_i)\right)\right] &\ge \exp\left(\mathbb{E}\left[\lambda \max_i X_i\right]\right) && (55)\\
\log \mathbb{E}\left[\exp\left(\lambda \max_i X_i)\right)\right] &\ge \mathbb{E}\left[\lambda \max_i X_i\right] && (56)
\end{aligned}$$

Next, we can use the original inequality we obtained and minimize with respect to λ:

$$\begin{aligned}
\frac{1}{\lambda}\log \mathbb{E}\left[\exp\left(\lambda \max_i X_i)\right)\right] &\le \frac{1}{\lambda}\log\left(m \exp\frac{1}{2}\lambda^2\sigma^2\right) && (57)\\
&= \frac{1}{\lambda}\log m + \frac{1}{2}\lambda\sigma^2 && (58)
\end{aligned}$$

We then compute the derivative and set to zero:

$$0 = -\frac{1}{\lambda^2}\log m + \frac{1}{2}\sigma^2 \tag{59}$$

Which yields $\lambda = \sqrt{\frac{2\log m}{\sigma^2}}$. Plugging this back in, combined with 56, yields the desired result:

$$\begin{aligned}
\mathbb{E}\left[\max_i X_i\right] &\le \frac{\sigma}{\sqrt{2\log m}}\log m + \frac{1}{2}\frac{\sqrt{2\log m}}{\sigma}\sigma^2 && (60)\\
&= \sqrt{2\sigma^2 \log m} && (61)
\end{aligned}$$

Problem 3(c) Massart's Finite Lemma

Show $\exists C > 0$ *s.t.* $\forall P$,

$$R_n(G) \triangleq \mathbb{E}\left[\sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i)\right] \leq C\sqrt{\frac{\log |G|}{n}} \tag{62}$$

Denote $A_i := \sigma_i g(z^{(i)})$.

1. Note that since $\sigma_i g(z^{(i)}) \overset{d}{=} g(z^{(i)})$, we have that $\mathbb{E}[A_i] = 0$.
2. Furthermore, since $-1 \leq A_i \leq 1$, A_i is bounded and this is sub-Gaussian with variance proxy $\sigma_i^2 = (1-(-1))^2/4 = 1$.
3. The sum of independent sub-Gaussian random variables is itself sub-Gaussian. Since $A_i \perp A_{j \neq i}$, we have that $\sum_{i=1}^n A_i$ is sub-Gaussian with variance proxy $\sigma^2 = \sum_{i=1}^n \sigma_i^2 = n$.
4. Let $\mathcal{A} = \{(x, \sigma_1, \ldots, \sigma_n) \mapsto \sum_{i=1}^n \sigma_i g(x) = \sum_{i=1}^n A_i \mid g \in G\}$. Note that, by definition, $|\mathcal{A}| \leq |G|$. We can apply the previous steps, in conjunction with the result from part (b), to obtain the desired result:

$$\begin{aligned}
R_n(G) &\triangleq \mathbb{E}\left[\sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z^{(i)})\right] && (63)\\
&= \frac{1}{n}\mathbb{E}\left[\sup_{a \in \mathcal{A}} a(z^{(i)}, \sigma_1, \ldots, \sigma_n)\right] && (64)\\
&\leq \frac{1}{n}\sqrt{2n \log |\mathcal{A}|} \qquad \textbf{[part (b)]} && (65)\\
&\leq \frac{1}{n}\sqrt{2n \log |G|} \qquad |\mathcal{A}| \leq |G| && (66)\\
&= C\sqrt{\frac{\log |G|}{n}} && (67)
\end{aligned}$$

with (at least for this derivation) $C = \sqrt{2}$.

Problem 3(d) General Discrete Distributions

Suppose $k > 1$, show that

$$R_n(F) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right] \leq C\sqrt{\frac{k}{n}} \tag{68}$$

for some universal constant $C > 0$.

Define G as a constrained version of $\mathcal{F}$, where the functions f are constrained to be applied only on the support vectors $V = \{v_i\}_{i=1}^{k}$:

$$G = \{(f(v_1), \ldots, f(v_k)) \mid f \in \mathcal{F}\} \tag{69}$$

This makes G finite, since $G \subset \{\pm 1\}^k$ (and thus $|G| \leq 2^k$). Therefore, we can use the result from part (c) to obtain the desired inequality as follows. Note that since $g \in G$ are each vectors of size k (and not functions over $\mathbb{R}^d$), I'll need to denote $g(z_i \in V)$, i.e. the element of g corresponding to $f(z_i)$, as $\sum_{v\in V} \mathbb{1}\{z_i = v\} g_v$.

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right] &= \mathbb{E}\left[\sup_{g\in G} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \sum_{v\in V}\mathbb{1}\{z_i = v\} g_v\right] && (70)\\
&\leq C\sqrt{\frac{\log|G|}{n}} \qquad \textbf{[part (c)]} && (71)\\
&\leq C\sqrt{\frac{k}{n}} \qquad [G \subset \{\pm 1\}^k] && (72)
\end{aligned}$$

Problem 3(e) Generalization Error Bound

Show that for $\delta \in (0, 1/3)$, there exists a universal constant $C > 0$ s.t. w.p. at least $1-\delta$ over the training data

$$L(\hat{h}) - L(h^*) \leq C\left(\sqrt{\frac{k}{n}} + \sqrt{\frac{\log 1/\delta}{n}}\right) \tag{73}$$

From remark 4.20 of the scribe notes, we know that $\forall h \in \mathcal{H}$,

$$L(h) - \hat{L}(h) \leq 2R_n(\mathcal{H}) + \sqrt{\frac{\log 2/\delta}{2n}} \tag{74}$$

Furthermore, we know that the excess risk is bounded by the generalization gap like

$$L(\hat{h}) - L(h^*) \leq 2\sup_{h\in\mathcal{H}}\left(L(h) - \hat{L}(h)\right) \tag{75}$$

Therefore, we can obtain the desired result by plugging in the inequality from part (d) and simplifying as follows.

$$R_n(\mathcal{H}) \leq C\sqrt{\frac{k}{n}} \qquad (C > 0) \tag{76}$$

$$L(\hat{h}) - L(h^*) \leq 4R_n(\mathcal{H}) + 2\sqrt{\frac{\log 2/\delta}{2n}} \tag{77}$$

$$\leq C_1\sqrt{\frac{k}{n}} + C_2\sqrt{\frac{\log 2/\delta}{n}} \tag{78}$$

$$\leq C_3\left(\sqrt{\frac{k}{n}} + \sqrt{\frac{\log 2/\delta}{n}}\right) \tag{79}$$

For some $C_3 = \max(C_1, C_2) > 0$.

Problem 4(a) Two Functions

Let $f : \mathcal{X} \to \mathbb{R}$ be a function, and let $\mathcal{F} := \{-f, f\}$ be a function class containing only two functions. Upper bound $R_n(\mathcal{F})$ using a function of n and $\mathbb{E}_{X\sim p^*}\left[f(X)^2\right]$, where the expectation is taken over $X \sim p^*$.

$$R_n(\mathcal{F}) = \mathbb{E}\left[\sup \frac{1}{n}\left\{\sum_{i=1}^{n} \sigma_i f(z_i), -\sum_{j=1}^{n} \sigma_j f(z_j)\right\}\right] \tag{80}$$

$$= \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right|\right] \tag{81}$$

$$\leq \left(\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right|^2\right]\right)^{1/2} \qquad \textbf{[Jensen's Ineq.]} \tag{82}$$

As usual, we can decompose this sum over σ_i like

$$\left(\sum_{i=1}^{n} \sigma_i f(z_i)\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_i \sigma_j f(z_i) f(z_j) \tag{83}$$

$$= \sum_{i=1}^{n} \sigma_i^2 f(z_i)^2 + \sum_{i=1}^{n}\sum_{j\neq i} \sigma_i \sigma_j f(z_i) f(z_j) \tag{84}$$

Noting that $\mathbb{E}\left[\sigma_i \sigma_{j\neq i}\right] = \mathbb{E}\left[\sigma_i\right]\mathbb{E}\left[\sigma_{j\neq i}\right] = 0$ since the σ are drawn i.i.d.:

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_i \sigma_j f(z_i) f(z_j)\right] = \mathbb{E}\left[\sum_{i=1}^{n} \sigma_i^2 f(z_i)^2\right] \tag{85}$$

$$= \sum_{i=1}^{n} \mathbb{E}\left[f(z_i)^2\right] \tag{86}$$

$$= n\mathbb{E}_{X\sim p^*}\left[f(X)^2\right] \tag{87}$$

Therefore

$$R_n(\mathcal{F}) \leq \frac{1}{n}\sqrt{n\mathbb{E}_{X\sim p^*}\left[f(X)^2\right]} \tag{88}$$

$$= \sqrt{\frac{\mathbb{E}_{X\sim p^*}\left[f(X)^2\right]}{n}} \tag{89}$$

Problem 4(b) Sparse Features, Dense Weights

Define the class of linear functions whose coefficients have bounded L_∞ norm:

$$\mathcal{F} = \{x \mapsto w \cdot x : ||w||_\infty \leq B\} \tag{90}$$

The domain of p^* is $\{x \in \{0,1\}^d \mid x \text{ has at most } k \text{ non-zero entries}\}$. Compute an upper bound on the Rademacher complexity $R_n\mathcal{F}$ as a function of B, k, d, n.

First, note that the dual of the ℓ_∞-norm is the ℓ_1-norm, i.e.

$$\sup_{||w||_\infty \leq B} \langle w, x\rangle = B\,||x||_1 \tag{91}$$

$$\begin{aligned}
R_n(\mathcal{F}) &= \mathbb{E}_{\substack{z^{(i)} \sim p^* \\ \sigma_i \sim \{\pm 1\}}} \left[\sup_{||w||_\infty \leq B} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle w, z^{(i)} \rangle \right] && (92)\\
&= \mathbb{E}_{\substack{z^{(i)} \sim p^* \\ \sigma_i \sim \{\pm 1\}}} \left[\sup_{||w||_\infty \leq B} \frac{1}{n} \langle w, \sum_{i=1}^n \sigma_i z^{(i)} \rangle \right] && (93)\\
&= \mathbb{E}_{\substack{z^{(i)} \sim p^* \\ \sigma_i \sim \{\pm 1\}}} \left[\frac{1}{n} B \left|\left| \sum_{i=1}^n \sigma_i z^{(i)} \right|\right|_1 \right] && (94)\\
&= \mathbb{E}_{\substack{z^{(i)} \sim p^* \\ \sigma_i \sim \{\pm 1\}}} \left[\frac{B}{n} \sum_{j=1}^d \left| \sum_{i=1}^n \sigma_i z_j^{(i)} \right| \right] && (95)\\
&\leq \mathbb{E}_{z^{(i)} \sim p^*} \left[\frac{B}{n} \sum_{j=1}^d \left| \sum_{i=1}^n z_j^{(i)} \right| \right] && (96)\\
&\leq \frac{B}{n} \mathbb{E}_{z^{(i)} \sim p^*} \left[\sum_{i=1}^n ||z^{(i)}||_1 \right] && (97)\\
&\leq \frac{B}{n} \mathbb{E}_{z^{(i)} \sim p^*} \left[\sum_{i=1}^n \min\{d, k\} \right] && (98)\\
&= B \min\{d, k\} && (99)
\end{aligned}$$

Problem 4(c) Sparse Weights, Dense Features

Now the domain of p^* is $\{z \in \mathbb{R}^d \mid ||z||_\infty \leq B\}$, and the class of linear functions is

$$\mathcal{F} = \{x \mapsto w \cdot x \mid ||w||_\infty \leq 1,\ w \text{ has at most } s \text{ non-zero entries }\} \tag{100}$$

Show that for some universal constant $c > 0$

$$R_n(\mathcal{F}) \leq cBs\sqrt{\frac{\log 2d}{n}} \tag{101}$$

$$R_n(\mathcal{F}) = \mathbb{E}_{\substack{z^{(i)} \sim p^* \\ ||z||_\infty \leq B \\ \sigma_i \sim \{\pm 1\}}} \left[\sup_{\substack{||w||_\infty \leq 1 \\ \text{at most } s \text{ non-zero}}} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle w, z^{(i)} \rangle \right] \tag{102}$$

Let $G = \{x \mapsto \langle w, x \rangle \mid ||w||_1 \leq s\}$. Note that $G \supset \mathcal{F}$. We can then apply Theorem 5.7 of the scribe notes, with $B := s$ and $C := B$, to obtain

$$R_S(G) \leq sB\sqrt{\frac{2\log 2d}{n}} \tag{103}$$

Let $c = \sqrt{2}$ to obtain the desired result.

Problem 4(d) Continuous Functions with Bounded Local Minima

Let $\mathcal{F}$ be the class of all continuous functions $f : \mathbb{R}[0,1] \to \mathbb{R}[0,1]$ with at most k local maxima. Prove that the Rademacher complexity of $\mathcal{F}$ is at most $O\left(\sqrt{\frac{k \log n}{n}}\right)$.

For bounding $R_S(\mathcal{F})$ for a function class $\mathcal{F}$ of continuous functions, we can use Dudley's theorem:

$$R_S(\mathcal{F}) \leq 12 \int_0^\infty \mathrm{d}\epsilon \frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n} \tag{104}$$

$$= 12 \int_0^1 \mathrm{d}\epsilon \frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n} \tag{105}$$

where the second step follows from the fact that the image of each f is in $[0,1]$. Since there are at most k local maxima, there are also at most $k-1$ local minima. If we have n total points z_i, then there are $\binom{n+2k-1}{n} = \binom{n+2k-1}{2k-1}$ ways to arrange the points relative the extrema.

In between each extremum, f is a monotonic (either non-increasing or non-decreasing) function. We can get a bound on the covering number for monotonic functions in $\mathcal{F}' = \{f : [a,b] \to [0,1] \mid f \in \mathcal{F}\}$.

1. Discretize the output space into $1/\epsilon$ intervals $\mathcal{Y} = \{[0,\epsilon], [\epsilon, 2\epsilon], \ldots, [(\frac{1}{\epsilon} - 1)\epsilon, 1]\}$.
2. For any given $f \in \mathcal{F}'$, note that every output $f(z_i)$ falls within an interval in $\mathcal{Y}$. Denote the upper bound of that interval as $\mathcal{Y}[z_i]$[2]. Define the piecewise function g for each of the z_i as

$$g(z_i) = \mathcal{Y}[z_i] \tag{106}$$

3. Then, by construction

$$L_2(P_n)(f,g) = \sqrt{\frac{1}{n}\sum_{i=1}^{n'} (f(z_i) - g(z_i))^2} \tag{107}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n'} (f(z_i) - \mathcal{Y}[z_i])^2} \tag{108}$$

$$\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n'} \epsilon^2} \tag{109}$$

$$= \epsilon \tag{110}$$

where $n' \leq n$ denotes the number of points z_i that are in the current monotonic interval $[a,b]$ being considered. Therefore,$\forall f \in \mathcal{F}'$, there exists *some* function g (specifically, the one defined by 106) for which $L_2(P_n)(f,g) \leq \epsilon$.

4. Therefore, we can get the covering number for monotonic functions $f : [a,b] \to [0,1]$ by counting the number of such functions g. Note that for each z_i, there are only $1/\epsilon$ unique

[2] By "upper bound of interval" here I'm referring to the value b for a given interval $[a,b]$.

possible values for $g(z_i)$. Therefore,

$$N(\epsilon, \mathcal{F}', L_2(P_n)) = O\left(n^{\frac{1}{\epsilon}}\right) \tag{111}$$

Recapping, we've now shown two main points:

1. There are $\binom{n+2k-1}{2k-1}$ possible arrangements of the n points relative to the $2k-1$ extrema of any $f \in \mathcal{F}$.
2. For each region $[a, b]$ in the input space between two extrema (minimum or maximum), the covering number for the functions in $\mathcal{F}$ evaluated over the points within that region, denoted as $\mathcal{F}'$, is

$$N(\epsilon, \mathcal{F}', L_2(P_n)) = O\left(n^{\frac{1}{\epsilon}}\right) \tag{112}$$

Therefore, the covering number over the full input space of $[0, 1]$ is

$$N(\epsilon, \mathcal{F}, L_2(P_n)) = O\left(n^{2k-1} n^{\frac{k}{\epsilon}}\right) \tag{113}$$

$$\log N(\epsilon, \mathcal{F}, L_2(P_n)) = O\left((k/\epsilon)\log n\right) \tag{114}$$

and we can now apply Dudley's theorem to obtain the desired result:

$$R_S(\mathcal{F}) \leq 12\frac{1}{\sqrt{n}}\int_0^1 \mathrm{d}\epsilon \sqrt{O\left((k/\epsilon)\log n\right)} \tag{115}$$

$$= O\left(\sqrt{\frac{k \log n}{n}}\right) \tag{116}$$