17.1 Announcements

- HW #4 due Monday, March 30
- No class on Wednesday, March 18th

17.2 Recap

Last time, we defined the covering number $N(\epsilon; \mathcal{F}, \| \cdot \|_{\infty})$ of a class $\mathcal{F}$ of functions with respect to the sup norm $\|f - g\|_{\infty} = \sup_x |f(x) - g(x)|$, and showed how the metric entropy $\log N(\epsilon; \mathcal{F}, \| \cdot \|_{\infty})$ controls the rates of convergence in binary classification. In particular, we considered a procedure based on forming an $\epsilon_n$-covering $\mathcal{F}_{\epsilon_n}$ of the original space $\mathcal{F}$, and then choosing the classifier $\hat{g}_n \in \mathcal{F}_{\epsilon_n}$ that minimizes the empirical risk. Here the accuracy $\epsilon_n$ is chosen as a function of the sample size $n$.

An important aspect of the proof from last time was the decomposition of the excess risk $R(\hat{g}_n) - R^*$ into two terms:

$$R(\hat{g}_n) - R^* = \left\{ R(\hat{g}_n) - \inf_{g \in \mathcal{F}_{\epsilon_n}} R(g) \right\} + \left\{ \inf_{g \in \mathcal{F}_{\epsilon_n}} R(g) - \inf_{g \in \mathcal{F}} R(g) \right\}$$

**Estimation error** **Approximation error**

The first term is the *estimation error*: it is a random variable that reflects the error that we incur because $\hat{g}_n$ is chosen based on minimizing an empirical quantity using $n$ samples, instead of minimizing based on the population probability of error. The second term is the *approximation error*: it is a deterministic (non-random) quantity that reflects how much we lose by searching over the cover $\mathcal{F}_{\epsilon_n}$ instead of over the original function space $\mathcal{F}$.

This type of decomposition applies also to other procedures for estimating classifiers, for instance when we use regularization, e.g., by restricting the set of allowable classifiers to some ball $\{\|f\|_H \leq R\}$ in an RKHS.

17.2.1 Data-dependent complexity: empirical covering numbers

Up until now, we have measured the complexity of our function classes in a *worst-case sense*:

- shatter coefficient $S(\mathcal{A}; n)$ of a class of sets for sample size $n$ is a worst-case over all data sets of size $n$
• covering number \( N(\epsilon; \mathcal{F}, \| \cdot \|_\infty) \) is also data-independent, and worst case

However, if we look carefully at how we need measures of function complexity for learning problems, then we see that tighter bounds are possible by measuring complexity in a data-dependent manner.

With motivation, let us define the empirical \( \ell_1 \)-covering numbers of a function class \( \mathcal{F} \) with respect to a data set \( \{ Z^{(1)}, \ldots, Z^{(n)} \} \). In particular, we define the set \( \mathcal{F}(\{ Z^{(i)} \}) \subset \mathbb{R}^n \) as follows:

\[
\mathcal{F}(\{ Z^{(i)} \}) = \{(f(Z^{(1)}), \ldots, f(Z^{(n)}) \mid f \in \mathcal{F} \}.
\]

We then define the empirical \( \ell_1 \)-covering number as \( N(\epsilon; \mathcal{F}(\{ Z^{(i)} \})) \) as the smallest number \( N \) of balls of radius \( \epsilon \) required to cover \( \mathcal{F}(\{ Z^{(i)} \}) \). In this definition, distances are measured in terms of the empirical \( \ell_1 \)-norm

\[
\| f - g \|_{\mathcal{F}(\{ Z^{(i)} \})} : = \frac{1}{n} \sum_{i=1}^{n} |f(Z^{(i)}) - g(Z^{(i)})|.
\]

**Aside:** The motivation for this definition comes from the definition of the metric in a \( L^1(\mathbb{Q}) \) space, where \( \mathbb{Q} \) is some probability distribution over \( \mathcal{X} \): it is given by \( \| f - g \|_\mathbb{Q} = \mathbb{E}_{\mathbb{Q}}[f(X) - g(X)] \). Our definition is a special case of this norm, where we choose the distribution \( \mathbb{Q} \) to be the empirical distribution that puts mass \( 1/n \) at each data point.

Note that unlike complexity measures that we have studied to date, the quantity \( N(\epsilon; \mathcal{F}(\{ Z^{(i)} \})) \) is a random variable, because it depends on the sample \( \{ Z^{(i)} \} \) that we have taken. Using this notion, we can state a uniform law of large numbers that is stronger than our previous results:

**Theorem 1** ((Pollard, 1984)). Consider a bounded class of functions \( \mathcal{F} \) such that \( \sup_x |f(x)| \leq B/2 \) for all \( f \in \mathcal{F} \). Then for any \( n \) and \( \epsilon > 0 \) with \( n \epsilon^2 \geq 8B^2 \), we have

\[
\mathbb{P}\left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z^{(i)}) - \mathbb{E}[f(Z)] \right| > \epsilon \right] \leq 2\mathbb{E}\left\{ N\left(\frac{\epsilon}{\sqrt{8}}, \mathcal{F}(\{ Z^{(i)} \}) \right) \right\} \exp\left( -\frac{n\epsilon^2}{128B^2} \right).
\]

**Remarks:**

(a) Note that this result measures function complexity in terms of the average covering number as opposed to a worst-case. Thus, it provides a sharper result than our earlier bound using the shatter coefficient from VC theory, as we will see below.

(b) From the Vapnik-Chernovenkis theorem, for a class of sets \( \mathcal{A} \), we know that

\[
\mathbb{P}\left[ \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_A(Z^{(i)}) - \mathbb{P}[Z \in A] \right| > \epsilon \right] \leq 8S(\mathcal{A}; n) \exp\left( -\frac{n\epsilon^2}{128B^2} \right),
\]

where \( S(\mathcal{A}; n) \) is the maximum number of subsets of \( n \) data points that can be distinguished by using sets in \( \mathcal{A} \). (Recall also that we showed that \( S(\mathcal{A}; n) \leq (n + 1)^{V(\mathcal{A})} \) when the class of sets \( \mathcal{A} \) has finite VC dimension \( V(\mathcal{A}) \).)
(c) To see why our new result is better, let us consider the empirical covering number in $\ell_\infty$-norm, meaning the smallest set of $N_\infty$ functions $\{g_1, \ldots, g_{N_\infty}\}$ such that for all $f \in F$, we can find a $g_i$ with

$$\max_i |f(Z^{(i)}) - g(Z^{(i)})| \leq \epsilon.$$ 

We claim that the empirical $\ell_1$-covering number $N(\epsilon; F(\{Z^{(i)}\}))$ is smaller than $N_\infty(\epsilon)$. Indeed, if $\{g_1, \ldots, g_N\}$ are a cover in empirical $\ell_\infty$-norm, then we have

$$\frac{1}{n} \sum_{i=1}^{n} |f(Z^{(i)}) - g(Z^{(i)})| \leq \max_i |f(Z^{(i)}) - g(Z^{(i)})| \leq \epsilon.$$ 

But suppose that our class of functions $F = \{\mathbb{I}_A(\cdot) \mid A \in A\}$ are the $0-1$-indicator functions for an underlying class of sets (i.e., $\mathbb{I}_A(Z) = 1$ if $Z \in A$ and 0 otherwise), as arises in the Vapnik-Chernovenkis theorem. Then for all $\epsilon \in (0, 1)$, we have

$$N_\infty(\epsilon) = \text{card}\{A \cap \{Z^{(1)}, \ldots, Z^{(n)}\} \mid A \in A\},$$

which is exactly the shattering number of the data with respect to the class of sets $A$. The shattering coefficient $S(A; n)$ is obtained by taking the worst-case over the data set, and we have shown that the empirical $\ell_1$-number is less than the shattering coefficient.

### 17.2.2 Sketch of proof

The first two steps of the proof are the same as the Vapnik-Chernovenkis theorem: we use a symmetrization argument, and then add in the random sign variables $\sigma^{(i)} \in \{-1, +1\}$ to reduce the problem to upper bounding the probability

$$\mathbb{P}\left[ \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} f(Z^{(i)}) \right| > \epsilon/4 \right].$$

(The change from $\epsilon$ to $\epsilon/4$ is a minor price that we paid in the symmetrization and random sign steps.)

Now let’s upper bound this quantity using the empirical $\ell_1$-covering. In particular, let us choose a covering of $N = N(\epsilon/8; F(\{Z^{(i)}\}))$ functions $\{g_1, \ldots, g_N\}$ such that for every $f \in F$, there is a function $g_k$ in our cover with

$$\frac{1}{n} \sum_{i=1}^{n} |f(Z^{(i)}) - g_k(Z^{(i)})| \leq \epsilon/8.$$ 

By triangle inequality, for any function $f$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} f(Z^{(i)}) \right| \leq \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} g_k(Z^{(i)}) + \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} (f(Z^{(i)}) - g_k(Z^{(i)})) \leq \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} g_k(Z^{(i)}) + \frac{1}{n} \sum_{i=1}^{n} |f(Z^{(i)}) - g_k(Z^{(i)})| \leq \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} g_k(Z^{(i)}) + \epsilon/8.$$ (17.1)
Therefore, conditioning on the data \( \{Z^{(i)}\} \) and using this inequality (17.1), we have
\[
\mathbb{P}\left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} f(Z^{(i)}) > \epsilon/4 \mid \{Z^{(i)}\} \right] \leq \mathbb{P}\left[ \max_{k=1,2,...,N} \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} g_k(Z^{(i)}) > \epsilon/8 \mid \{Z^{(i)}\} \right]
\]
This is very nice, since we have again reduced the sup over the whole function space to a maximum over our covering set (as in our last lecture), with \( N = N(\epsilon/8; \mathcal{F}(\{Z^{(i)}\})) \) elements.

Consequently, we can apply union bound to obtain that
\[
\mathbb{P}\left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} f(Z^{(i)}) > \epsilon/4 \mid \{Z^{(i)}\} \right] \leq N(\epsilon/8; \mathcal{F}(\{Z^{(i)}\})) \max_{k} \mathbb{P}\left[ \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} g_k(Z^{(i)}) > \epsilon/8 \mid \{Z^{(i)}\} \right].
\]
Recall that our functions \( g_k \) belong to \( \mathcal{F} \) and so are bounded by assumption. Therefore, conditioned on \( \{Z^{(i)}\} \), the random variables \( \sigma^{(i)} g_k(Z^{(i)}) \) are zero-mean and bounded random variables. Therefore, we can apply the Hoeffding bound to conclude that for any \( g_k \), we have
\[
\mathbb{P}\left[ \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} g_k(Z^{(i)}) > \epsilon/8 \mid \{Z^{(i)}\} \right] \leq 2 \exp\left( -\frac{ne^2}{128B^2} \right).
\]
Taking expectations over the random data \( \{\{Z^{(i)}\}\} \) completes the proof.

### 17.3 Further links between covering and VC dimension

We conclude by considering some further links between covering and VC dimension. We have defined VC dimension for classes of sets \( \mathcal{A} \), but it is naturally extended to a class of functions \( \mathcal{F} \) by considering the associated class of sets in \( \mathcal{X} \times \mathbb{R} \) given by
\[
\mathcal{A}_\mathcal{F} = \left\{ \{(x,t) \mid f(x) \leq t\}, f \in \mathcal{F} \right\}.
\]
Note that we made a similar transformation earlier in the context of binary classification, where classifiers can be defined by the subset \( \{x \mid f(x) \leq 0\} \). We use \( V_\mathcal{F} \) to denote the VC dimension of the class of sets \( \mathcal{A}_\mathcal{F} \).

Let us consider coverings in a norm of the form \( \|f - g\|_1 = \int |f(x) - g(x)\mu(x)dx \), where \( \mu(x) \) is some density function. (This includes as a special case the empirical \( \ell_1 \)-covering numbers that we have defined). Let \( N(\epsilon; \mathcal{F}) \) be the \( \epsilon \)-covering number of \( \mathcal{F} \) in such a weighted \( L_1 \)-norm. For any class of bounded functions \( \mathcal{F} \) (with \( \sup_x |f(x)| \leq B \) for all \( f \in \mathcal{F} \)), it can be shown that
\[
N(\epsilon; \mathcal{F}) \leq \left( \frac{4eB}{\epsilon} \log \frac{2eB}{\epsilon} \right)^{V_\mathcal{F}}.
\]
Therefore, taking logarithms yields that the metric entropy
\[
\log N(\epsilon; \mathcal{F}) = O(V_\mathcal{F} \log (1/\epsilon)),
\]
where we view \( B \) as a constant. Recalling from last lecture our discussion about parametric function classes, we see that classes with finite VC dimension are quite small, in that they behave like parametric classes with effective dimension \( V_\mathcal{F} \).