Symmetrization and Rademacher Complexity

Administrative/Meta:

- Daniel Hsu talk today at 1pm in Beckman 1005.
- Homework 2 this weekend? Maybe?
- Date for project presentations: December 8 at noon, or December 10 at noon.

Overview

This lecture will give **symmetrization**, a powerful argument for at the heart of many different generalization bounds. This arugment itself suggests a basic complexity measure, **Rademacher complexity**, which will be our basic complexity measure henceforth; namely, other will complexity measures we derive will then appear in expressions upper bounds Rademacher complexity.

Problems with primitive covers

Let's first see how primitive covers were inadequate. Recall that a function class $\mathcal{G}$ is a primitive cover for a function class $\mathcal{F}$ at scale $\epsilon > 0$ over some set S if:

- $\mathcal{G} \subseteq \mathcal{F}$,
- $|\mathcal{G}| < \infty$, and
- for every $f \in \mathcal{F}$ there exists $g \in \mathcal{G}$ with $\sup_{x \in S} |g(x) - f(x)| \leq \epsilon$.

Last class, we gave a generalization bound for classes with primitive covers (basically, primitive covers give discretizations, and then we apply finite class generalization).

Problems with primitive covers. It's pretty easy to run into limits of this technique.

- Consider linear predictors as before, but the points $x \in \mathbb{R}^d$ are from some unbounded distribution, for instance a Gaussian. This immediately breaks the earlier construction. One fix is to truncate the distribution: since Gaussians concentrate well, we can find an X so that $\|x\|_2 \leq X$ with probability at least $1 - \delta$ (and this X does not depend too badly on n: recall from homework 1 the analysis of the maximum of a collection of scalar Gaussian random variables). So now we can first condition away an event of probability at most δ that some points have $\|x\|_2 > X$, and then run the cover argument as before.
- Consider discontinuous function classes, for instance $w \mapsto \text{sgn}(\langle w, x \rangle)$. If $\epsilon < 2$, for any linear classifier f there must exist g_f that exactly agrees with f on every point (i.e., any $\epsilon < 2$ may as well be $\epsilon = 0$). Since for any $x \neq 0$ and $w \neq 0$, $\text{sgn}(\langle w, x \rangle) \neq \text{sgn}(\langle -w, x \rangle)$, it follows that the primitive covering number is again infinite (e.g., for any $w \neq 0$, the only vectors within $\epsilon < 2$ for this metric is the set $\{cw : c \in \mathbb{R} \setminus \{0\}\}$, so the cover must include one vector for each direction, as well as 0). There are a number of ways to fix this (including giving non-primitive covers); we will come back to it after discussing Rademacher complexity.

There is a better notion of cover that fixes these, but we'll get there through Rademacher complexity.

Symmetrization/Rademacher part 1: without concentration

We'll work in slightly more generality than before.

$$\begin{aligned} Z \quad & \text{random variable; could encode } (X,Y);\\ \mathbb{E} \quad & \text{expectation over } Z;\\ \mathbb{E}_n \quad & \text{expectation over } n \text{ i.i.d. } (Z_1,\dots,Z_n);\\ \mathbb{E}f = \mathbb{E}(f) = \mathbb{E}(f(X)) \quad & \text{shorthand;}\\ \hat{\mathbb{E}}f = \hat{\mathbb{E}}(f) = n^{-1}\sum_i f(Z_i) \quad & \text{shorthand.} \end{aligned}$$

Note that we are working with single functions f; to discuss a risk $\mathcal{R}$ in this notation, we could use the function class $\{(x,y) \mapsto \ell(-yf(x)) : f \in \mathcal{F}\}$.

Let's see how far we can get in building generalization without the use of concentration inequalities. This means that we will be controlling the expected value

$$\mathbb{E}_n\left(\sup_{f\in\mathcal{F}} \hat{\mathbb{E}}f - \mathbb{E}f\right) = \mathbb{E}_n\left(\sup_{f\in\mathcal{F}} n^{-1}\sum_i f(Z_i) - \mathbb{E}(f(Z))\right).$$

Part 2 of this analysis will invoke concentration to replace the expectation with a high probability bound.

The heart of symmetrization is to replace $\mathbb{E}f$ with $\mathbb{E}'_n f$ over a second sample $(Z'_i,\dots,Z'_n)$. In particular, define

$$\begin{aligned} (Z'_1,\dots,Z'_n) \quad & \text{second sample;}\\ \mathbb{E}'_n \quad & \text{expectation over i.i.d. } (Z'_1,\dots,Z'_n);\\ \hat{\mathbb{E}}'f = \hat{\mathbb{E}}'(f) = n^{-1}\sum_i f(Z'_i) \quad & \text{shorthand.} \end{aligned}$$

Directly, $\mathbb{E}f = \mathbb{E}'_n\hat{\mathbb{E}}'f$, thus

$$\mathbb{E}_n\left(\sup_{f\in\mathcal{F}} \mathbb{E}f - \hat{\mathbb{E}}f\right) = \mathbb{E}_n\left(\sup_{f\in\mathcal{F}} \mathbb{E}'_n\hat{\mathbb{E}}'f - \hat{\mathbb{E}}f\right) \le \mathbb{E}_n\mathbb{E}'_n\left(\sup_{f\in\mathcal{F}} \hat{\mathbb{E}}'f - \hat{\mathbb{E}}f\right).$$

(Putting the supremum inside the expectation only increases things; can be checked by choosing $\epsilon > 0$ and an f_ϵ near the supremum.)

The next piece is the magical part of the argument. For any fixed vector $\sigma \in \{-1,+1\}^n$,

$$\mathbb{E}_n\mathbb{E}'_n\left(\sup_{f\in\mathcal{F}} \hat{\mathbb{E}}'f - \hat{\mathbb{E}}f\right) = \mathbb{E}_n\mathbb{E}'_n\left(\sup_{f\in\mathcal{F}} n^{-1}\sum_i (f(Z_i) - f(Z'_i))\right) = \mathbb{E}_n\mathbb{E}'_n\left(\sup_{f\in\mathcal{F}} n^{-1}\sum_i \sigma_i(f(Z_i) - f(Z'_i))\right);$$

this follows because the distribution on $(Z_1,\dots,Z_i,\dots,Z_n,Z'_1,\dots,Z'_i,\dots,Z'_n)$ and $(Z_1,\dots,Z'_i,\dots,Z_n,Z'_1,\dots,Z_i,\dots,Z'_n)$ are the same, and the same argument holds for an arbitrary number of swaps. Said another way, we can swap data points between two random samples without changing anything. For a more explicit argument see the Shai-Shai book *[future matus: explicit ref]*. *[future matus: notatoin Z_{-i} instead of Z'_i let's me look at $Z_{\sigma i}$?.]* *[maybe also discuss it as a permutation of two data sets, and σ being a generator for that group? I discussed it from an angle like this in class.]*

Since this holds for any fixed $\sigma \in \{-1,+1\}$, it holds in expectation over σ drawn from n Rademacher random variables, meaning $\sigma \in \{-1,+1\}^n$ where $\Pr[\sigma_i = +1] = \Pr[\sigma_i = -1] = 1/2$, independently for each coordinate. Thus

$$\mathbb{E}_n\mathbb{E}'_n\left(\sup_{f\in\mathcal{F}} \hat{\mathbb{E}}'f - \hat{\mathbb{E}}f\right) = \mathbb{E}_\sigma\mathbb{E}_n\mathbb{E}'_n\left(\sup_{f\in\mathcal{F}} \hat{\mathbb{E}}'f - \hat{\mathbb{E}}f\right) = \mathbb{E}_\sigma\mathbb{E}_n\mathbb{E}'_n\left(\sup_{f\in\mathcal{F}} n^{-1}\sum_i \sigma_i(f(Z_i) - f(Z'_i))\right).$$

By properties of suprema and linearity of expectation, we can split this expression, giving

$$\begin{aligned}\mathbb{E}_\sigma \mathbb{E}_n \mathbb{E}'_n \left(\sup_{f\in\mathcal{F}} n^{-1} \sum_i \sigma_i (f(Z_i) - f(Z'_i)) \right) &\le \mathbb{E}_\sigma \mathbb{E}_n \mathbb{E}'_n \left(\sup_{f\in\mathcal{F}} \sup_{f'\in\mathcal{F}} n^{-1} \sum_i \sigma_i (f(Z_i) - f'(Z'_i)) \right) \\ &= 2\mathbb{E}_\sigma \mathbb{E}_n \left(\sup_{f\in\mathcal{F}} n^{-1} \sum_i \sigma_i f(Z_i) \right).\end{aligned}$$

This final expression gives us **Rademacher complexity**: namely, given a sample $S := (Z_1, \ldots, Z_n)$, define $\mathrm{Rad}(\mathcal{F}_{|S})$ as

$$\mathrm{Rad}(\mathcal{F}_{|S}) = \mathbb{E}_\sigma \left(\sup_{v\in\mathcal{F}_{|S}} \frac{1}{n} \sum_{i=1}^n \sigma_i v_i \right), \quad \text{where } \mathcal{F}_{|S} := \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\}.$$

(It is useful to define Rademacher complexity for vectors, and then define restriction classes $\mathcal{F}_{|S}$ separately.)

The above derivation gave the following

Theorem.

$$\begin{aligned}\mathbb{E}_n \left(\sup_{f\in\mathcal{F}} \mathbb{E} f - \hat{\mathbb{E}} f \right) &\le \mathbb{E}_n (\mathrm{Rad}(\mathcal{F}_{|S})) \le \sup_S \mathrm{Rad}(\mathcal{F}_{|S}), \\ \mathbb{E}_n \left(\sup_{f\in\mathcal{F}} \mathbb{E} f - \hat{\mathbb{E}} f \right) &\le \mathbb{E}_n (\mathrm{Rad}(\mathcal{F}_{|S})) \le \sup_S \mathrm{Rad}(\mathcal{F}_{|S}).\end{aligned}$$

Proof. The inequality in the first line was derived above. The second line follows by working with the function class $-\mathcal{F} := \{-f : f \in \mathcal{F}\}$ (by simply replacing f with $-f$ in the first line) and noting $\mathrm{Rad}(\mathcal{F}_{|S}) = \mathrm{Rad}(-\mathcal{F}_{|S})$. □

Remark.

- Observe that $\mathrm{Rad}(\mathcal{F}_{|S}) = 0$ whenever $|\mathcal{F}| = 1$; this may seem trivial, but it is a useful sanity check that $\mathrm{Rad}(\cdot)$ is really measuring some sort of complexity of $\mathcal{F}$. The original definition of Rad was $\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} |n^{-1} \sum_i \sigma_i f(x_i)|$, and the absolute value meant that the "complexity" of even $\mathcal{F}$ consisting of a single constant mapping could be arbitrary large.
- $\mathrm{Rad}(\mathcal{F}_{|S})$ can be interpreted as: the ability of a function class to fit random sign sequences.

Rademacher part 2: generalization with concentration

We will now combine the symmetrization/Rademacher approach with concentration results to get the bound we want, namely a high probability upper bound on $\sup_{f\in\mathcal{F}} \mathbb{E} f - \hat{\mathbb{E}} f$.

Notice that this random quantity is not quite amenable to Hoeffding or Azuma. We need something a little more powerful.

Theorem (McDiarmid's inequality). Let a function $f : \mathbb{R}^n \to \mathbb{R}$ be given with the *bounded differences* property: for every $i \in \{1, \ldots, n\}$, there exists c_i so that

$$\sup_{z_1,\ldots,z_i,\ldots,z_n,z'_i} |f(z_1, \ldots, z_i, \ldots, z_n) - f(z_1, \ldots, z'_i, \ldots, z_n)| \le c_i.$$

Then with probability at least $1 - \delta$ over a draw of independent random variables $(Z_1, \ldots, Z_n)$,

$$\mathbb{E}(f(Z_1, \ldots, Z_n)) \le f(Z_1, \ldots, Z_n) + \sqrt{\frac{\sum_i c_i^2}{2} \ln\left(\frac{1}{\delta}\right)}.$$

Remark.

- The proof obtaining the constants in the statement is similar to the proof of Azuma, but with the bounded differences property. It is possible to prove a version of the statement with worse constants via Azuma. Specifically, let $\sigma_i := \sigma(Z_1, \ldots, Z_i)$ be the σ-algebra generated from $(Z_1, \ldots, Z_i)$. Define $Z_i := \mathbb{E}(f(Z_1, \ldots, Z_n)|\sigma_i)$. Note that

$$\mathbb{E}(Z_i - Z_{i-1}|\sigma_{i-1})\mathbb{E}(Z_i - Z_{i-1}|\sigma_{i-1}) = \mathbb{E}(f(Z_1, \ldots, Z_n)|\sigma_{i-1}) - \mathbb{E}(f(Z_1, \ldots, Z_n)|\sigma_{i-1}) = 0.$$

  From here, the bounded difference property implies $|Z_i - Z_{i-1}| \leq 2c_i$, which allows Azuma to be applied; the "2" is the source of the degraded constants. A full proof of McDiarmid with proper constants can be found in Maxim's book *[future matus: proper reference]*.

- Note that McDiarmid implies Hoeffding. If $Z_i \in [a_i, b_i]$ with probability 1, then $n^{-1}\sum_i x_i$ satisfies bounded differences with $c_i := (b_i - a_i)/n$. Plugging this into McDiarmid recovers Hoeffding exactly.

We can now apply McDiarmid to $\sup_{f \in \mathcal{F}} \mathbb{E}f - \hat{\mathbb{E}}f$ and $\text{Rad}(\mathcal{F}_{|S})$ in order to obtain our full desired bounds.

Theorem. Let function class $\mathcal{F}$ be given, and suppose $|f(x)| \leq c$ with probability 1.

1. With probability at least $1 - \delta$ over the draw of a sample $S := (Z_1, \ldots, Z_n)$,

$$\sup_{f \in \mathcal{F}} \mathbb{E}f - \hat{\mathbb{E}}f \leq \mathbb{E}_n \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i f(Z_i) - \mathbb{E}(f(Z_1)) \right) + c\sqrt{\frac{2}{n} \ln\left(\frac{1}{\delta}\right)}.$$

2. With probability at least $1 - \delta$ over the draw of a sample $S := (Z_1, \ldots, Z_n)$,

$$\mathbb{E}_n(\text{Rad}(\mathcal{F}_{|S})) \leq \text{Rad}(\mathcal{F}_{|S}) + c\sqrt{\frac{2}{n} \ln\left(\frac{1}{\delta}\right)}.$$

3. With probability at least $1 - \delta$ over the draw of a sample $S := (Z_1, \ldots, Z_n)$, every $f \in \mathcal{F}$ satisfies

$$\mathbb{E}f \leq \hat{\mathbb{E}}f + 2\text{Rad}(\mathcal{F}_{|S}) + 3c\sqrt{\frac{2}{n} \ln\left(\frac{2}{\delta}\right)}.$$

Proof.

1. It suffices to check the bounded differences property. Observe

$$\begin{aligned}
&\sup_{Z_1,\ldots,Z_i,Z_i',\ldots,Z_n} \left| \sup_{f \in \mathcal{F}} \left(\mathbb{E}f - \hat{\mathbb{E}}f\right) - \sup_{f \in \mathcal{F}} \left(n^{-1}(f(Z_i') + \sum_{i \neq i'} f(Z_i)) - \mathbb{E}f \right) \right| \\
&= \sup_{Z_1,\ldots,Z_i,Z_i',\ldots,Z_n} \left| \sup_{f \in \mathcal{F}} \left(\mathbb{E}f - \hat{\mathbb{E}}f\right) - \sup_{f \in \mathcal{F}} \left(n^{-1}(-f(Z_i') + f(Z_i)) + \mathbb{E}f - \hat{\mathbb{E}}f \right) \right| \\
&\leq \sup_{Z_1,\ldots,Z_i,Z_i',\ldots,Z_n} \left| \sup_{f \in \mathcal{F}} \left(\mathbb{E}f - \hat{\mathbb{E}}f\right) - \sup_{f' \in \mathcal{F}} \left(n^{-1}(f'(Z_i) - f'(Z_i')\right) - \sup_{f \in \mathcal{F}} \left(\mathbb{E}f - \hat{\mathbb{E}}f\right) \right| \\
&\leq \sup_{Z_1,\ldots,Z_i,Z_i',\ldots,Z_n} |0| + \sup_{f' \in \mathcal{F}} \left| n^{-1}(f'(Z_i') - f'(Z_i)\right| \\
&\leq 2cn^{-1}.
\end{aligned}$$

   The result now follows by McDiarmid's inequality with bounded differences constant $2cn^{-1}$.

2. Similarly,

$$
\begin{aligned}
&\sup_{Z_1,\ldots,Z_i,Z_i',\ldots,Z_n} \left| \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} n^{-1} \sum_i \sigma_i f(x_i) - \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} n^{-1}(\sigma_i f(x_i') + \sum_{i\neq i'} \sigma_i f(x_i)) \right| \\
&\leq \sup_{Z_1,\ldots,Z_i,Z_i',\ldots,Z_n} \left| \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} n^{-1} \sum_i \sigma_i f(x_i) - \mathbb{E}_\sigma \sup_{f'\in\mathcal{F}} n^{-1} \sigma_i (f'(x_i') - f'(x_i)) - \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} n^{-1} \sum_{i\neq i'} \sigma_i f(x_i) \right| \\
&\leq 2cn^{-1}.
\end{aligned}
$$

3. This last follows by combining the pieces together with the earlier theorem on $\mathrm{Rad}(\mathcal{F}_{|S|})$.

Remark. *[We discussed a bunch of other stuff here but I don't remember what it was. Maybe I have some notes somewhere ...]*

References