Lecture 19. (Sketch.)

- No class this Wednesday, November 7!
- I’ll be in my office today 5-8pm if anyone wants to discuss course project.
- Please sign up for project proposal meetings — you don’t get full credit without it.
- Homework 2 should go out later today.

1. Recap from past two lectures.

Hoeffding lets us control a single random variable: with probability at least $1 - \delta$ over an iid draw of $(Z_1, \ldots, Z_n)$ with $Z_i \in [a, b]$ a.s.,

$$\mathbb{E}Z \leq \frac{1}{n} \sum_i Z_i + (b - a) \sqrt{\frac{\ln(1/\delta)}{2n}}.$$ 

- From here, we can bound a single function’s risk by defining $Z_i := \ell(\hat{f}(x_i), y_i)$.
- If $\hat{f}$ depends on $((x_i, y_i))_{i=1}^n$ (e.g., it is the output of a training algorithm), then $(Z_1, \ldots, Z_n)$ as defined above are no longer necessarily iid!

The standard fix in learning theory is a **uniform deviation bound** over a class of functions $\mathcal{F}$: e.g., a bound of the form

$$\Pr \left[ \sup_{f \in \mathcal{F}} \mathcal{R}_\ell(f) - \hat{\mathcal{R}}_\ell(f) > \epsilon \right] \leq \text{some function of } \mathcal{F}, \ell, \epsilon, n.$$ 

So far, we have a bound based on $|\mathcal{F}|$:

- Let $\mathcal{F}, \ell$, and a probability distribution be given so that $\ell(f(x), y) \in [a, b]$ almost surely. With probability at least $1 - \delta$, for every $f \in \mathcal{F}$,

$$\mathcal{R}_\ell(f) \leq \hat{\mathcal{R}}_\ell(f) + (b - a) \sqrt{\frac{\ln |\mathcal{F}| + \ln(1/\delta)}{2n}}.$$ 

- If $|\mathcal{F}| = \infty$, we can still use this via discretization. The most naive discretization (“primitive cover” from last class) requires a finite subset $G$ so that $\forall f \in \mathcal{F}, \exists g \in G,$ $\sup_x |g(x) - f(x)| \leq \epsilon$. If $\mathcal{F}$ denotes linear classifiers, and $\epsilon < 2$, then $|G| = \infty$ is necessary!

2. Generalization without concentration: symmetrization.

The standard approach has two key steps. Some notation:

- $Z$ r.v.; e.g., $(x, y)$,
- $\mathcal{F}$ functions; e.g., $f(Z) = \ell(g(X), Y)$,
- $\mathbb{E}$ expectation over $Z$,
- $\mathbb{E}_n$ expectation over $(Z_1, \ldots, Z_n)$,
- $\hat{\mathbb{E}}f = \mathbb{E}f(Z)$,
- $\hat{\mathbb{E}}_n f = \frac{1}{n} \sum_i f(Z_i)$.

In this notation, $\mathcal{R}_\ell(g) = \mathbb{E} \ell \circ g$ and $\hat{\mathcal{R}}_\ell(g) = \hat{\mathbb{E}} \ell \circ g$. 

- Is there some way to work with only the behavior on the training set, rather than all possible points?
**First key step:** introduce another sample (“ghost sample”). Let \((Z'_1, \ldots, Z'_n)\) be another iid draw from \(Z\); define \(E'_n\) and \(\hat{E}'_n\) analogously.

**Lemma 1.** \(E_n \left( \sup_{f \in \mathcal{F}} E f - \hat{E}_n f \right) \leq E_n E'_n \left( \sup_{f \in \mathcal{F}} \hat{E}'_n f - \hat{E}_n f \right)\).

**Proof.** Fix any \(\epsilon > 0\) and apx max \(f_i \in \mathcal{F}\); then

\[
E_n \left( \sup_{f \in \mathcal{F}} E f - \hat{E}_n f \right) \leq E_n \left( E f_i - \hat{E}_n f_i \right) + \epsilon
\]

\[
= E_n \left( E' f_i - \hat{E}_n f_i \right) + \epsilon
\]

\[
= E'_n E_n \left( \hat{E}'_n f_i - \hat{E}_n f_i \right) + \epsilon
\]

\[
\leq E'_n E_n \left( \sup_{f \in \mathcal{F}} \hat{E}'_n f - \hat{E}_n f \right) + \epsilon
\]

Result follows since \(\epsilon > 0\) arbitrary.

**Key step 2:** swap points between the two samples; a magic trick with random signs boils this down into a manageable quantity.

Fix a vector \(\epsilon \in \{-1, +1\}^n\) and define a r.v. \((U_i, U'_i) := (Z_i, Z'_i)\) if \(\epsilon = 1\) and \((U_i, U'_i) = (Z'_i, Z_i)\) if \(\epsilon = -1\). Then

\[
E_n E'_n \left( \sup_{f \in \mathcal{F}} \hat{E}'_n f - \hat{E}_n f \right) = E_n E'_n \left( \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \left( f(Z'_i) - f(Z_i) \right) \right)
\]

\[
= E_n E'_n \left( \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(U'_i) - f(U_i) \right) \right).
\]

Here's the big trick: since \((Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n)\) and \((U_1, \ldots, U_n, U'_1, \ldots, U'_n)\) have same distribution, and \(\epsilon\) arbitrary, then (with \(Pr[\epsilon_i = +1] = 1/2\) iid “Rademacher”)

\[
E_n E'_n \left( \sup_{f \in \mathcal{F}} \hat{E}'_n f - \hat{E}_n f \right) = E_n E'_n \left( \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(U'_i) - f(U_i) \right) \right)
\]

\[
= E_n E'_n \left( \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(Z'_i) - f(Z_i) \right) \right).
\]

Since similarly replacing \(\epsilon_i\) and \(-\epsilon_i\) doesn't change \(E\),

\[
E_n E'_n \left( \sup_{f \in \mathcal{F}} \hat{E}'_n f - \hat{E}_n f \right)
\]

\[
= E_n E'_n \left( \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(Z'_i) - f(Z_i) \right) \right)
\]

\[
\leq E_n E'_n \left( \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(Z'_i) - f'(Z_i) \right) \right)
\]

\[
= E_n E'_n \left( \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(Z'_i) \right) \right) + E_n E'_n \left( \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( -f(Z'_i) \right) \right)
\]

\[
= 2E_n \frac{1}{n} \sup_{f \in \mathcal{F}} n \sum_i \epsilon_i \left( f(Z'_i) \right) = 2E_n \frac{1}{n} URad(\mathcal{F}_S) = 2E_n Rad(\mathcal{F}_S),
\]

where \(URad(\mathcal{F}_S)\) and \(Rad(\mathcal{F}_S)\) respectively denote the unnormalized Rademacher complexity and (normalized) Rademacher complexity.
Specifically, define unnormalized Rademacher complexity $URad(V)$ as

$$URad(V) := \mathbb{E} \sup_{u \in V} \langle \varepsilon, u \rangle, \quad Rad(V) := \frac{1}{n} URad(V).$$

Typically, we’ll have a sample $S = (Z_1, \ldots, Z_n)$, and invoke this with vectors

$$\mathcal{F}|_S := \{ (f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F} \}.$$

Summarizing our derivations:

**Lemma 2.** $\mathbb{E}_n \mathbb{E}'_n \sup_{f \in \mathcal{F}} (\hat{\mathbb{E}}'_n f - \hat{\mathbb{E}}_n f) \leq \frac{2}{n} \mathbb{E}_n URad(\mathcal{F}|_S)$.

**Remarks.**
- Can flip $\hat{\mathbb{E}}'_n$ and $\hat{\mathbb{E}}_n$ using $-\mathcal{F} := \{ -f : f \in \mathcal{F} \}$.
- Rademacher complexity arose as its own concept in early 2000s (the work of Bartlett, Mendelson, Koltchinskii, \ldots); the expressions and derivations go back decades. “Stop the proof in the middle and draw a box” – Bartlett.
- Can view this as fitting $\mathcal{F}|_S$ to random signs, but usually we work with $\mathcal{F} = \ell \circ \mathcal{G}$.
- Note that $URad(\{ u \}) = 0$, $URad(\mathcal{V} + \{ c \}) = Rad(\mathcal{V})$; fails for original definition $\mathbb{E} \sup_{u \in \mathcal{V}} \langle \varepsilon, u \rangle / n$.
- Rademacher complexity is not perfect: e.g., hard to prove $1/n$ rates, and I don’t know how to use it to prove best deep net generalization. But it and its lemmas are still very convenient!
- Other texts all use Rad; I like URad.
- Both lemmas in the section are called symmetrization.

### 3. Generalization with concentration.

We controlled expected uniform deviations: $\mathbb{E}_n \sup_{f \in \mathcal{F}} Ef - \hat{\mathbb{E}}_n f$.

High probability bounds will follow via concentration inequalities.

**Theorem** (McDiarmid). Suppose $F : \mathbb{R}^n \to \mathbb{R}$ satisfies “bounded differences”: $\forall i \in \{1, \ldots, n\} \exists c_i$,

$$\sup_{z_1, \ldots, z_n, z_i} \left| F(z_1, \ldots, z_i, \ldots, z_n) - F(z_1, \ldots, z'_i, \ldots, z_n) \right| \leq c_i.$$

With $p \geq 1 - \delta$,

$$\mathbb{E}_n F(Z_1, \ldots, Z_n) \leq F(Z_1, \ldots, Z_n) + \sqrt{\frac{\sum_i c_i^2}{2}} \ln(1/\delta).$$

**Remarks.**
- Proof: analyze MGF, apply Chernoff technique. (Proof with worst constants: corollary of Azuma.)
- Hoeffding follows by setting $F(\bar{Z}) = \sum_i Z_i / n$ and verifying bounded differences $c_i := (b_i - a_i) / n$.

**Theorem.** Let $\mathcal{F}$ be given with $f(z) \in [a, b]$ a.s..

1. With probability $\geq 1 - \delta$, 

$$\sup_{f \in \mathcal{F}} Ef - \hat{\mathbb{E}}_n f \leq \mathbb{E}_n \left( \sup_{f \in \mathcal{F}} Ef - \hat{\mathbb{E}}_n f \right) + (b - a) \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

2. With probability $\geq 1 - \delta$, 

$$\mathbb{E}_n URad(\mathcal{F}|_S) \leq URad(\mathcal{F}|_S) + (b - a) \sqrt{\frac{n \ln(1/\delta)}{2}}.$$

3. With probability $\geq 1 - \delta$, 

$$\sup_{f \in \mathcal{F}} Ef - \hat{\mathbb{E}}_n f \leq \frac{2}{n} URad(\mathcal{F}|_S) + 3(b - a) \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

**Proof (sketch).** McDiarmid and our symmetrization lemmas.