## 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.,

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

- 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

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

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 $f \in \mathcal{F}$,

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

- 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 $\forall f \in \mathcal{F}$, $\exists g \in G$, $\sup_x |g(x) - f(x)| \le \epsilon$. If $\mathcal{F}$ denotes linear *classifiers*, and $\epsilon < 2$, then $|G| = \infty$ is necessary!
    - Is there some way to work with only the behavior on the

## 2. Generalization *without* concentration: symmetrization.

The standard approach has two key steps. Some notation:

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

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

**First key step:** introduce another sample ("ghost sample"). Let $(Z'_1, \ldots, Z'_n)$ be another iid draw from $Z$; define $\mathbb{E}'_n$ and $\hat{\mathbb{E}}'_n$ analogously.

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

**Proof.** Fix any $\epsilon > 0$ and apx max $f_\epsilon \in \mathcal{F}$; then

$$\begin{aligned}
\mathbb{E}_n \left( \sup_{f\in\mathcal{F}} \mathbb{E}f - \hat{\mathbb{E}}_n f \right) &\le \mathbb{E}_n \left( \mathbb{E}f_\epsilon - \hat{\mathbb{E}}_n f_\epsilon \right) + \epsilon \\
&= \mathbb{E}_n \left( \mathbb{E}'_n \hat{\mathbb{E}}'_n f_\epsilon - \hat{\mathbb{E}}_n f_\epsilon \right) + \epsilon \\
&= \mathbb{E}'_n \mathbb{E}_n \left( \hat{\mathbb{E}}'_n f_\epsilon - \hat{\mathbb{E}}_n f_\epsilon \right) + \epsilon \\
&\le \mathbb{E}'_n \mathbb{E}_n \left( \sup_{f\in\mathcal{F}} \hat{\mathbb{E}}'_n f - \hat{\mathbb{E}}_n f \right) + \epsilon
\end{aligned}$$

Result follows since $\epsilon > 0$ arbitrary.

**Remarks.**

- Notice we are working only *in expectation* for now. In the subsequent section, we'll get high probability bounds. But $\sup_{f\in\mathcal{F}} \mathbb{E}f - \mathbb{E}'_n f$ is a random variable; can describe it in many other ways too! (E.g., "asymptotic normality".)
- This lemma says we can instead work with two samples. Working with two samples could have been our starting point: by itself it is a meaningful and interpretable quantity!

**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

$$\begin{aligned}
\mathbb{E}_n \mathbb{E}'_n \left( \sup_{f\in\mathcal{F}} \hat{\mathbb{E}}'_n f - \hat{\mathbb{E}}_n f \right) &= \mathbb{E}_n \mathbb{E}'_n \left( \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_i \left( f(Z'_i) - f(Z_i) \right) \right) \\
&= \mathbb{E}_n \mathbb{E}'_n \left( \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(U'_i) - f(U_i) \right) \right).
\end{aligned}$$

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")

$$\begin{aligned}
\mathbb{E}_\epsilon \mathbb{E}_n \mathbb{E}'_n \left( \sup_{f\in\mathcal{F}} \hat{\mathbb{E}}'_n f - \hat{\mathbb{E}}_n f \right) &= \mathbb{E}_\epsilon \mathbb{E}_n \mathbb{E}'_n \left( \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(U'_i) - f(U_i) \right) \right) \\
&= \mathbb{E}_\epsilon \mathbb{E}_n \mathbb{E}'_n \left( \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(Z'_i) - f(Z_i) \right) \right).
\end{aligned}$$

Since similarly replacing $\epsilon_i$ and $-\epsilon_i$ doesn't change $\mathbb{E}_\epsilon$,

$$\begin{aligned}
&\mathbb{E}_\epsilon \mathbb{E}_n \mathbb{E}'_n \left( \sup_{f\in\mathcal{F}} \hat{\mathbb{E}}'_n f - \hat{\mathbb{E}}_n f \right) \\
&= \mathbb{E}_\epsilon \mathbb{E}_n \mathbb{E}'_n \left( \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(Z'_i) - f(Z_i) \right) \right) \\
&\le \mathbb{E}_\epsilon \mathbb{E}_n \mathbb{E}'_n \left( \sup_{f,f'\in\mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(Z'_i) - f'(Z_i) \right) \right) \\
&= \mathbb{E}_\epsilon \mathbb{E}'_n \left( \sup_{f'\in\mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( f(Z'_i) \right) \right) + \mathbb{E}_\epsilon \mathbb{E}'_n \left( \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_i \epsilon_i \left( -f'(Z_i) \right) \right) \\
&= 2\mathbb{E}_n \frac{1}{n} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_i \epsilon_i \left( f(Z_i) \right) = 2\mathbb{E}_n \frac{1}{n} \mathrm{URad}(\mathcal{F}_{|S}) = 2\mathbb{E}_n \mathrm{Rad}(\mathcal{F}_{|S}),
\end{aligned}$$

where $\mathrm{URad}(\mathcal{F}_{|S})$ and $\mathrm{Rad}(\mathcal{F}_{|S})$ respectively denote the **unnormalized Rademacher complexity** and (normalized) **Rademacher complexity**.

Specifically, define unnormalized Rademacher complexity $\text{URad}(V)$ as

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

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.** $\mathbb{E}_n \mathbb{E}'_n \sup_{f \in \mathcal{F}} \left(\hat{\mathbb{E}}'_n f - \hat{\mathbb{E}}_n f\right) \le \frac{2}{n} \mathbb{E}_n \text{URad}(\mathcal{F}_{|S})$.

**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, ...); 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 $\text{URad}(\{u\}) = 0$, $\text{URad}(V + \{c\}) = \text{Rad}(V)$; fails for original definition $\mathbb{E}_\epsilon \sup_{u \in V} |\langle \epsilon, u \rangle / n|$.
- 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: $\mathbb{E}_n \sup_{f \in \mathcal{F}} \mathbb{E} f - \hat{\mathbb{E}}_n f$.

High probability bounds will follow via concentration inequalities.

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

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

With pr $\ge 1 - \delta$,

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

**Remarks.**

- Proof: analyze MGF, apply Chernoff technique. (Proof with worst constants: corollary of Azuma.)
- Hoeffding follows by setting $F(\vec{Z}) = \sum_i Z_i / n$ 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 $\ge 1 - \delta$,
$$\sup_{f \in \mathcal{F}} \mathbb{E} f - \hat{\mathbb{E}}_n f \le \mathbb{E}_n \left( \sup_{f \in \mathcal{F}} \mathbb{E} f - \hat{\mathbb{E}}_n f \right) + (b - a) \sqrt{\frac{\ln(1/\delta)}{2n}}.$$
2. With probability $\ge 1 - \delta$,
$$\mathbb{E}_n \text{URad}(\mathcal{F}_{|S}) \le \text{URad}(\mathcal{F}_{|S}) + (b - a) \sqrt{\frac{n \ln(1/\delta)}{2}}.$$
3. With probability $\ge 1 - \delta$,
$$\sup_{f \in \mathcal{F}} \mathbb{E} f - \hat{\mathbb{E}}_n f \le \frac{2}{n} \text{URad}(\mathcal{F}_{|S}) + 3(b - a) \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

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