Recall that in order to deal with the independence issue (between algorithm output and training set used to compute training error), we switched to studying

\[ \sup_{f} R(f) - \hat{R}(f), \]

a uniform deviation.

Recall our basic covering bound.

**Theorem (abbreviated, see last time).** With \( \delta = 2 \sigma \),

\[ \sup_{f \in F} R(f) - \hat{R}(f) \leq \inf_{\epsilon \geq 0} \left( 2\epsilon + (b-\alpha) \sqrt{ \frac{\ln(\mathcal{N}(F, \epsilon, 1, 11\|\mu\|)) + \ln(1/\delta)}{2n} } \right) \]

where \( \mathcal{Z} \) is input space, \( \epsilon \) of \( \mathcal{E} \) for all \( \mathcal{E}_i \) over \( \mathcal{Z} \).

The key difficulty here is the uniform norm \( \| \cdot \|_\infty \): we're requiring cover \( \mathcal{G} \) to satisfy

\[ \sup_{f \in \mathcal{G}} \sup_{g \in \mathcal{G}} \sup_{(x,y) \in \mathcal{Z}} | l(f(x), y) - l(g(x), y) | \leq \epsilon. \]
Example: Let's revisit linear predictors (under assumption 1.3).

Suppose $Z$ imposes $\|x\| \leq x^T eW_X$, $\alpha \in (-1, 1)$:

$$f = Z \times 1 <w, x> : x \in \ell W_X$$

where $\ell$ is Lipschitz, $\ell(0, y) = \ell(-<w, xy>)$.

Then (as before) let $eW_X = \ell \alpha - eW_X$ over $Z$.

Moreover, suppose we have a cover of $\ell$ so that:

$$\forall w, \exists w' \text{ with } ||w - w'|| \leq \frac{\alpha}{eX}. \text{ Then}$$

$$||\ell(<w, xy>) - \ell(<w', xy>)|| \leq e \cdot ||w - w'|| \leq e \cdot \frac{\alpha}{eX} \leq \epsilon.$$

This cover exists, has size $O\left(\frac{eW_X}{\epsilon}\right)$. They have bound

$$\sup_{w} R(w) - \widehat{R}(w) \leq \inf_{\epsilon > 0} \left(2\epsilon + 2eW_X \sqrt{\frac{d \cdot 0(\epsilon eW_X) + \ln(\epsilon \alpha)}{2n}} \right)$$

by choosing

$$\epsilon = \frac{1}{2} \epsilon < \frac{1}{2} \epsilon (\text{why...})$$

$$\leq \frac{2}{n} + 2eW_X \sqrt{\frac{d \cdot 0(\ln(eW_X)) + d \ln(\alpha) + \ln(\frac{1}{n})}{2n}}.$$

Remarks:

* Direct sort had no "d"! We'll get rid of it in a few lectures...

* Covering is due to Kolmogorov-Tikhomirov; basic refs: book/lecture, etc.

* We discussed "why $\ln(\frac{1}{R})$" and some union bound! Without it here, we'd be expanding in dimension!
A bigger issue.

Suppose again \( \mathcal{F} := \{ x \mapsto \langle w,x \rangle : \|w\| \leq 1 \} \).

Then \( \|x\| \leq \frac{2}{\varepsilon} \|w\| \) whereby \( \|x\| \leq \varepsilon \|w\| \leq 1 \).

Consider any \( w \) \in \mathcal{F} \) with \( w \neq w' \), \( \|w\| = 1 \).

Thus, to cover \( \mathcal{F} \) with this loss (so \( \mathcal{F} \) is in \( \mathbb{R}^H \)), need a representative for each line (when \( \varepsilon < 2 \))!

I.e., any cover is infinite!

\[ \langle x, w \rangle = \frac{1}{2} (\|w\|^2 - \langle w, w' \rangle) > 0 \Rightarrow \text{sgn} (\langle x, w \rangle) = 1 \]

\[ \langle x, w' \rangle = \frac{1}{2} (\langle w, w' \rangle - \|w'\|^2) < 0 \Rightarrow \text{sgn} (\langle x, w' \rangle) = -1. \]

\rightarrow Our setup has a **trivial** issue!

We are working too hard. **Trying** to handle an enitre **uncountable set**?

Can we restrict attention to *just* the sample?
Generalization without concentration: Symmetrization.

We'll now give a way to handle $\sup_{f \in \mathcal{F}} R(f) - \hat{R}(f)$ which allows us to only consider behavior on a finite sample.

There are two key parts to this.

First some brief notation.

$\mathbf{Z}$ our r.v. could be $(X, Y)$

$\mathbf{f}$ our functions

$L$ (loss is bounded in $L$)

$\mathbb{E}$ expectation over $\mathbb{E}$

$\mathbb{E}_n$ expectation over i.i.d. $(Z_1, \ldots, Z_n)$.

$\mathbf{f}$ shorthand $\mathbb{E}(f(Z))$

$\hat{\mathbf{f}}_n$ empirical average $\frac{1}{n} \sum_i f(Z_i)$.

**First key step**: Introduction of a second sample.

Let $\tilde{Z}_1, \ldots, \tilde{Z}_n$ be another i.i.d. draw from $\mathbf{Z}$.

(Similarly define $\hat{\mathbf{f}}'_n$, $\hat{\mathbf{f}}''_n$).

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

(proof on next page.)
Proof. Fix any \( \varepsilon > 0 \), let \( \tau_0 \) be some \( \varepsilon \)-approximation.

\[
E_n \left( \sup_{f \in \mathcal{F}} E_t - \hat{E}_{n,t} \right) \leq E_n \left( E_t - \hat{E}_{n,t} \right) + \varepsilon \\
= E_n \left( E_t \cap \hat{E}_{n,t} - \hat{E}_{n,t} \right) + \varepsilon \\
= E_n \left( \hat{E}_{n,t} - \hat{E}_{n,t} \cap E_t \right) + \varepsilon \\
\leq E_n \left( \sup_{f \in \mathcal{F}} \hat{E}_{n,t} - \hat{E}_{n,t} \right) + \varepsilon.
\]

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

Remarks.

\# This lemma reduces consideration to behavior on two independent samples.

Question: why don't we make that our goal and not \( \mathbb{E}(M) - \mathbb{E}(E) \)?

I.e., what is it more meaningful to bet on bound gaps on two samples, one sampled at an expectation?

This question is cultural.

And similarly, this section is "without concentration"; we are bounding expectations.

The desire to focus on "All w/ prob. 2 \( \rightarrow \) is cultural; there are many other ways to specify the distribution of \( \sup_{f \in \mathcal{F}} E_t - \hat{E}_{n,t} \)."
**Key step 2:** Scanning points between two samples; a magic trick with random signs.

Fix a vector \( \xi \in \mathbb{E}^{-1}, \mathbb{H}^{3n} \). Then

\[
\begin{align*}
\mathbb{E}_n \mathbb{E}_n' \sup_{f \in \mathbb{H}^n} \hat{\mathbb{E}}^n f \cdot \hat{\mathbb{E}}^n f = \mathbb{E}_n \mathbb{E}_n' \sup_{f \in \mathbb{H}^n} \frac{1}{n} \sum_i (f(Z_i) - f(Z'_i)) \\
= \mathbb{E}_n \mathbb{E}_n' \sup_{f \in \mathbb{H}^n} \frac{1}{n} \sum_i \xi_i (f(Z_i) - f(Z'_i))
\end{align*}
\]

Since \((Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n)\) has the same distribution as

\[
(U_1, \ldots, U_n, U'_1, \ldots, U'_n)
\]

where

\[
U_i = \begin{cases} 
Z_i & \xi_i = +1 \\
Z'_i & \xi_i = -1
\end{cases} \quad U'_i = \begin{cases} 
Z'_i & \xi_i = +1 \\
Z_i & \xi_i = -1
\end{cases}
\]

(Noisy, both are i.i.d. from \(Z\) and \(Z'\) of size \(2n/4\).)

Since this holds for any \(\xi \in \mathbb{E}^{-1}, \mathbb{H}^{3n}\), it holds for

\[
\text{distributing over } \xi. \text{ Pick Rademacher. Now } \Pr \{ \xi_i = \pm 1 \} = \frac{1}{2} \text{ for all } i \leq 13^3 \text{ dec둡 }
\]

independently for all coordinates. Then

\[
\mathbb{E}_n \mathbb{E}_n' \sup_{f \in \mathbb{H}^n} \hat{\mathbb{E}}^n f \cdot \hat{\mathbb{E}}^n f = \mathbb{E}_n \mathbb{E}_n' \sup_{f \in \mathbb{H}^n} \frac{1}{n} \sum_i (f(Z_i) - f(Z'_i)) \\
\leq \mathbb{E}_n \mathbb{E}_n' \sup_{f \in \mathbb{H}^n} \frac{1}{n} \sum_i \xi_i (f(Z_i) - f(Z'_i)) \\
= 2 \mathbb{E}_n \mathbb{E}_n' \sup_{f \in \mathbb{H}^n} \frac{1}{n} \sum_i \xi_i (f(Z_i) - f(Z'_i))
\]

Next page.
Then
\[ \sup_{f \in \mathcal{F}} \left( \hat{\varepsilon}_n + f - \hat{\varepsilon}_n f \right) = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n z_i (f(\xi_i) - f(\xi'_i)) \]
\[ \leq \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n z_i (f(\xi_i) - f(\xi'_i)) \]
\[ = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n z_i f(\xi_i) + \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n z_i f(\xi'_i) \]

**Lemma.**
\[ \sup_{f \in \mathcal{F}} \left( \hat{\varepsilon}_n + f - \hat{\varepsilon}_n f \right) \leq \frac{2}{n} \sup_{\alpha \in \mathcal{H}} \text{Rad}(\mathcal{F}_s) \]
\[ \leq \frac{2}{n} \sup_{\alpha \in \mathcal{H}} \text{Rad}(\mathcal{F}_s) \]

**Proof.** As above; second line by considering \(-f = \hat{\varepsilon} - f : f \in \mathcal{F}_s\).
Remarks.

- Rademacher complexity appeared as a concept "in its own right" in early 2000s;
  papers by Bartlett, Mendelson, Kolthofer.

The expressions & derivations go back decades; took all that time to isolate it!

"Stop proof in middle & draw line" — Bartlett.

Can view it as fitting 4 is to random signs, but complicated because typically we apply 4 to \((E^3)\) is.

- It has sanity check that
  \[
  \text{Rad} (\mathcal{E} + 3) = 0, \quad \text{Rad} (\mathcal{V} + 3\mathcal{E}) = \text{Rad} (\mathcal{V}).
  \]
  The original definition \(E \sup_{f \in \mathcal{X}} \sum_{i \in [n]} |f(i)| \) fails this and adds other complications.

- Rademacher complexity is not perfect; e.g., it is hard to prove \(\frac{1}{n} \) rates with it.
  Even so we'll prove other styles of bounds (VC & covering) \(\leq\) Rademacher.

- Everyone else includes factor \(1/n\). I'm experimenting...

- People refer to both key steps as "symmetrization".
We only controlled the expectation:

$$\mathbb{E} \sup \limits_{f \in \mathcal{F}} R(f) - \bar{R}(f).$$

We can easily convert this to a high probability bound via the following concentration inequality:

**Theorem (McDiarmid).** Suppose $f : \mathbb{R}^n \to \mathbb{R}$ satisfies bounded differences:

$$\forall i \in \{1, \ldots, n\} \exists c_i \quad \forall z, z' \quad |f(z_1, \ldots, z_i, \ldots, z_n) - f(z_1, \ldots, z'_i, \ldots, z_n)| \leq c_i.$$

Then

$$\sup \limits_{z_1, \ldots, z_n} |f(z_1, \ldots, z_n) - f(Z_1, \ldots, Z_n)| \leq \sum \frac{c_i^2}{2} \ln \left( \frac{1}{\delta} \right) + \sqrt{\sum \frac{c_i^2}{2} \ln \left( \frac{1}{\delta} \right)}.$$

**Rem.**

A brute force w/ degendered construct follows from Azuma. Full proof requires its own application of Chernoff bounding & mgf manipulation.

McDiarmid implies Hoeffding: $Z_i \in [a_i, b_i], Z \Rightarrow \frac{1}{n} \sum \frac{c_i \cdot Z_i}{n}.$
Theorem. Let functions $f$ be given with $f(z) \in (a,b)$ a.s.

1. w/ $pr \geq 1-\delta$,
   \[
   \sup_{f \in F} \mathbf{E}|f - \hat{f}| \leq \mathbf{E}_n \left( \sup_{f \in F} |f - \hat{f}| \right) + \left( b - a \right) \sqrt{\frac{n \ln(n\delta)}{2n}},
   \]

2. w/ $pr \geq 1-\delta$,
   \[
   \mathbf{E}_n \left( \text{Rod} \left( \mathbf{F}_n \right) \right) \leq \text{Rod} \left( \mathbf{F}_1 \right) + \left( b - a \right) \sqrt{\frac{n \ln(n\delta)}{2n}}.
   \]

3. w/ $pr \geq 1-\delta$,
   \[
   \sup_{f \in F} \mathbf{E}|f - \hat{f}| \leq \frac{2}{n} \text{Rod} \left( \mathbf{F}_1 \right) + 3(b-a) \sqrt{\frac{\ln(n\delta)}{2n}}.
   \]

Remark. Not hard to establish bounded differences; key of proof is instead the symmetrization stuff; details might be a homework problem?