Lecture 21. (Sketch.)

- Project proposal meetings today!

Rademacher recap (same slide as last time!).

Concentration controlled one function at a time. To control many functions, out main tool is (unnormalized) Rademacher complexity:

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

Given data $S := (Z_1, \ldots, Z_n)$ and functions $\mathcal{F}$, define vectors

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

Our main generalization tool involves $\mathrm{URad}(\mathcal{F}_{|S})$, and is a consequence of our two symmetrization lemmas and McDiarmid's inequality.

Theorem. Let $\mathcal{F}$ be given with $f(z) \in [a, b]$ a.s.. With probability $\geq 1 - \delta$,

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

We proved we can peel off Lipschitz losses.

Corollary. Suppose ℓ is ρ-lipschitz and $\ell \circ \mathcal{F} \in [a, b]$ a.s.. With probability $\geq 1 - \delta$, every $f \in \mathcal{F}$ satisfies

$$\mathcal{R}_\ell(f) \leq \widehat{\mathcal{R}}_\ell(f) + \frac{2\rho}{n}\mathrm{URad}(\mathcal{F}_{|S}) + 3(b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Now suppose we want to control misclassifications:

$$\Pr[\mathrm{sgn}(f(X)) \neq Y] = \mathcal{R}_z(f) \leq ?$$

We'll give two approaches:

- VC ("Vapnik-Chernvonenkis") theory: RHS based on $\widehat{\mathcal{R}}_z$. Seems easier to get bounds based on combinatorial properties of $\mathcal{F}$.
- Margin theory: RHS based on *margin loss*. Seems easier to get bounds based on real-valued properties of $\mathcal{F}$.

2. VC Theory.

First, some definitions. First, the zero-one/classification risk/error:

$$\mathcal{R}_z(\mathrm{sgn}(f)) = \Pr[\mathrm{sgn}(f(X)) \neq Y], \quad \widehat{\mathcal{R}}_z(\mathrm{sgn}(f)) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}[\mathrm{sgn}(f(x_i)) \neq y_i]$$

The earlier Rademacher bound will now have

$$\mathrm{URad}\left(\{(x, y) \mapsto \mathbb{1}[\mathrm{sgn}(f(x)) \neq y] : f \in \mathcal{F}\}_{|S}\right).$$

This is at most 2^n; we'll reduce it to a combinatorial quantity:

$$\mathrm{sgn}(U) := \{(\mathrm{sgn}(u_1), \ldots, \mathrm{sgn}(u_n)) : u \in V\},$$

$$\mathrm{Sh}(\mathcal{F}_{|S}) := \left|\mathrm{sgn}(\mathcal{F}_{|S})\right|,$$

$$\mathrm{Sh}(\mathcal{F}; n) := \sup_{\substack{S \in ? \\ |S| \leq n}} \left|\mathrm{sgn}(\mathcal{F}_{|S})\right|,$$

$$\mathrm{VC}(\mathcal{F}) := \sup\{i \in \mathbb{Z}_{\geq 0} : \mathrm{Sh}(\mathcal{F}; i) = 2^i\}.$$

Remarks.

- Sh is "shatter coefficient", VC is "VC dimension".
- Both quantities are criticized as being too tied to their worst case; bounds here depend on (empirical quantity!) $\mathrm{URad}(\mathrm{sgn}(\mathcal{F}_{|S}))$, which can be better, but throws out the labels.

Theorem ("VC Theorem"). With probability at least $1-\delta$, every $f \in \mathcal{F}$ satisfies

$$\mathcal{R}_z(\mathrm{sgn}(f)) \le \widehat{\mathcal{R}}_z(\mathrm{sgn}(f) + \frac{2}{n}\mathrm{URad}(\mathrm{sgn}(\mathcal{F}_{|S})) + 3\sqrt{\frac{\ln(2/\delta)}{2n}},$$

and

$$\mathrm{URad}(\mathrm{sgn}(\mathcal{F}_{|S})) \le \sqrt{2n \ln \mathrm{Sh}(\mathcal{F}_{|S})},$$
$$\ln \mathrm{Sh}(\mathcal{F}_{|S}) \le \ln \mathrm{Sh}(\mathcal{F}; n) \le \mathrm{VC}(\mathcal{F}) \ln(n+1).$$

Remarks.

- Need $\mathrm{Sh}(\mathcal{F}_{|s}) = o(n)$ "in order to learn".
- $\mathrm{VC}(\mathcal{F}) < \infty$ suffices; many considered this a conceptual breakthrough, namely "learning is possible"!
- The quantities (VC, Sh) appeared in prior work (not by V-C). Symmetrization apparently too, though I haven't dug this up.

First step of proof: pull out the zero-one loss.

Lemma.
$\mathrm{URad}(\{(x,y) \mapsto \mathbb{1}[\mathrm{sgn}(f(x)) \ne y] : f \in \mathcal{F}\}_{|S}) \le \mathrm{URad}(\mathrm{sgn}(\mathcal{F}_{|S}))$.

Proof. For each i, define

$$\ell_i(z) := \max\left\{0, \min\left\{1, \frac{1 - y_i(2z-1)}{2}\right\}\right\},$$

which is 1-Lipschitz, and satisfies

$$\ell_i(\mathrm{sgn}(f(x_i))) = \mathbb{1}[\mathrm{sgn}(f(x_i)) \ne y_i].$$

(Indeed, it is the linear interpolation.) Then

$$\begin{aligned}
&\mathrm{URad}(\{(x,y) \mapsto \mathbb{1}[\mathrm{sgn}(f(x)) \ne y] : f \in \mathcal{F}\}_{|S}) \\
&= \mathrm{URad}(\{(\ell_1(\mathrm{sgn}(f(x_1))), \ldots, \ell_n(\mathrm{sgn}(f(x_n)))) : f \in \mathcal{F}\}_{|S}) \\
&= \mathrm{URad}(\ell \circ \mathrm{sgn}(\mathcal{F})_{|S}) \\
&\le \mathrm{URad}(\mathrm{sgn}(\mathcal{F})_{|S}).
\end{aligned}$$

Plugging this into our Rademacher bound: w/ pr $\ge 1-\delta$, $\forall f \in \mathcal{F}$,

$$\mathcal{R}_z(\mathrm{sgn}(f)) \le \widehat{\mathcal{R}}_z(\mathrm{sgn}(f)) + \frac{2}{n}\mathrm{URad}(\mathrm{sgn}(\mathcal{F})_{|S}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Our next step is a general Rademacher bound for finite sets.

Theorem (Massart finite lemma).
$\mathrm{URad}(V) \le \sup_{u \in V} \|u\|_2 \sqrt{2 \ln |V|}$.

Remarks.

- $\ln |V|$ is what we expect from union bound.
- $\|\cdot\|_2$ (rather than arbitrary geometry) is kindof annoying and intrinsic to these tools (subgaussian, hoeffding, ...).

We'll prove this via a few lemmas.

Lemma. If $(X_1, \ldots, X_n)$ are c^2-subgaussian, then $\mathbb{E}\max_i X_i \le c\sqrt{2\ln(n)}$.

Proof. Similar to homework 2.

Lemma. If $(X_1, \ldots, X_n)$ are c_i^2-subgaussian and independent, $\sum_i X_i$ is $\|\vec{c}\|_2^2$-subgaussian.

Proof. We did this in the concentration lecture, but here it is again:

$$\mathbb{E}\exp(t\sum_i X_i) = \prod_i \mathbb{E}\exp(tX_i) \le \prod_i \exp(t^2 c_i^2/2) = \exp(t^2\|\vec{c}\|_2^2/2).$$

Proof (of Massart finite lemma).

Let $\vec{\epsilon}$ be iid Rademacher and fix $u \in V$. Define $X_{u,i} := \epsilon_i u_i$ and $X_u := \sum_i X_{u,i}$.

By Hoeffding lemma, $X_{u,i}$ is $(u_i - -u_i)^2/4 = u_i^2$ -subgaussian, thus (by Lemma) X_u is $\|u\|_2^2$-subgaussian. Thus

$$\mathrm{URad}(V) = \mathbb{E}_\epsilon \max_{u\in V} \langle \epsilon, u\rangle = \mathbb{E}_\epsilon \max_{u\in V} X_u \le \max_{u\in V}\|u\|_2\sqrt{2\ln|V|}.$$

Plugging this into our bound gives

$$\mathrm{URad}(\mathrm{sgn}(\mathcal{F}_{|S})) \le \sqrt{2n\mathrm{Sh}(\mathcal{F}_{|S})}.$$

One last lemma remains for the proof.

Lemma (Sauer-Shelah? Vapnik-Chervonenkis? Warren? ...)

Let $\mathcal{F}$ be given, and define $V := \mathrm{VC}(\mathcal{F})$. Then

$$\mathrm{Sh}(\mathcal{F}; n) \le \begin{cases} 2^n & \text{when } n \le V, \\ \left(\frac{en}{V}\right)^V & \text{otherwise.} \end{cases}$$

Moreover, $\mathrm{Sh}(\mathcal{F}; n) \le n^V + 1$.

(**Proof.** Omitted. Exists in many standard texts.)

Remarks. (on the VC theorem.)

- Minimizing $\widehat{\mathcal{R}}_z$ is NP-hard in many trivial cases, but those require noise and neural networks can often get $\widehat{\mathcal{R}}_z(\mathrm{sgn}(f)) = 0$.
- Recent work prefers real-valued / scale-sensitive complexity measures, where it is easier (?) to depend on things like weight matrix norms in neural networks.

3. Margin bounds.

- Rather than looking at just $\operatorname{sgn}(f(x))$, let's evaluate the *magnitude* of f.
- These bounds will be sensitive to real-valued (rather than combinatorial) properties of $\mathcal{F}$, and also to the labels (encoded via a "margin assumption" implicit in assuming the training margin risk $\widehat{\mathcal{R}}_\gamma$ is small).

Define $\ell_\gamma(z) := \max\{0, \min\{1, 1 + z/\gamma\}\}$,
$\mathcal{R}_\gamma(f) := \mathcal{R}_{\ell_\gamma}(f) = \mathbb{E}\ell_\gamma(-Yf(X))$.

Theorem. With probability $\geq 1-\delta$, $\forall f \in \mathcal{F}$,

$$\mathcal{R}_z(f) \leq \mathcal{R}_\gamma(f) \leq \widehat{\mathcal{R}}_\gamma(f) + \frac{2}{n\gamma}\mathrm{URad}(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Proof. Since

$$\mathbb{1}[\operatorname{sgn}(f(x)) \neq y] \leq \mathbb{1}[-f(x)y \geq 0] \leq \ell_\gamma(-f(x)y),$$

then $\mathcal{R}_z(f) \leq \mathcal{R}_\gamma(f)$. The bound between $\mathcal{R}_\gamma$ and $\widehat{\mathcal{R}}_\gamma$ follows from the fundamental Rademacher bound, and by peeling the $1/\gamma$-Lipschitz function ℓ_γ.

Remark.

- Started with Bartlett '96 "For valid generalization, the size of the weights is more important than the size of the network". (Worst-case VCcan't handle scale: $\operatorname{sgn}(f) = \operatorname{sgn}(cf)$ for $c > 0$. Margin bounds can handle scale.)
- Intuition: can wiggle (rotate up to γ) predictor without changing output labels.
- To invoke theorem, we need to show that algorithms actually give a small $\widehat{\mathcal{R}}_\gamma$ (which is stronger than requiring small $\widehat{\mathcal{R}}_z$). We'll see in homework that we often have something like this for convex losses.
- Often these bounds are used with l_1 balls of predictors, which is the same as $\operatorname{conv}(\mathcal{F} \cup -\mathcal{F})$. (Next page gives some tools for this.)

Following properties can help apply margin bounds.

Lemma.

1. $\mathrm{URad}(V) \geq 0$.
2. $\mathrm{URad}(cV + \{u\}) \leq |c|\mathrm{URad}(V)$.
3. $\mathrm{URad}(\operatorname{conv}(V)) \leq \mathrm{URad}(V)$.
4. Let $(V_i)_{i\geq 0}$ be given with $\sup_{u\in V_i}\langle u, \epsilon\rangle \geq 0$ $\forall \epsilon \in \{-1,+1\}^n$. (E.g., $V_i = -V_i$, or $0 \in V_i$.) Then $\mathrm{URad}(\cup_i V_i) \leq \sum_i \mathrm{URad}(V_i)$.
5. $\mathrm{URad}(V) = \mathrm{URad}(-V)$.

Remarks.

- (3) is a mixed blessing: "Rademacher is insensitive to convex hulls",
- (4) is true for $\mathrm{URad}_{|\cdot|}$ directly: define $W_i := V_i \cup -V_i$, which satisfies the conditions, and note $(\cup_i V_i) \cup -(\cup_i V_i) = \cup_i W_i$. Since $\mathrm{URad}_{|\cdot|}(V_i) = \mathrm{URad}(W_i)$, then $\mathrm{URad}_{|\cdot|}(\cup_i V_i) = \mathrm{URad}(\cup_i W_i) \leq \sum_{i\geq 1}\mathrm{URad}(W_i) = \sum_{i\geq 1}\mathrm{URad}_{|\cdot|}(V_i)$.

Proof.

(1.) Fix any $u_0 \in V$; then $\mathbb{E}_\epsilon \sup_{u \in V} \langle \epsilon, v \rangle \geq \mathbb{E}_\epsilon \langle \epsilon, u_0 \rangle = 0$.

(2.) Either check directly, or use the $|c|$-Lipschitz functions $\ell_i(r) := c \cdot r + u_i$.

(4.) Using the condition,

$$\begin{aligned}\mathbb{E}_\epsilon \sup_{u \in \cup_i V_i} \langle \epsilon, u \rangle &= \mathbb{E}_\epsilon \sup_i \sup_{u \in V_i} \langle \epsilon, u \rangle \leq \mathbb{E}_\epsilon \sum_i \sup_{u \in V_i} \langle \epsilon, u \rangle \\ &= \sum_{i \geq 1} \mathrm{URad}(V_i).\end{aligned}$$

(5.) Since integrating over ϵ is the same as integrating over $-\epsilon$ (the two are equivalent distributions),

$$\mathrm{URad}(-V) = \mathbb{E}_\epsilon \sup_{u \in V} \langle \epsilon, -u \rangle = \mathbb{E}_\epsilon \sup_{u \in V} \langle -\epsilon, -u \rangle = \mathrm{URad}(V).$$

Proof (continued).

(3.) This follows since optimization over a polytope is achieved at a corner. In detail,

$$\begin{aligned}\mathrm{URad}(\mathrm{conv}(V)) &= \mathbb{E}_\epsilon \sup_{\substack{k \geq 1 \\ \alpha \in \Delta_k}} \sup_{u_1, \ldots, u_k \in V} \left\langle \epsilon, \sum_j \alpha_j u_j \right\rangle \\ &= \mathbb{E}_\epsilon \sup_{\substack{k \geq 1 \\ \alpha \in \Delta_k}} \sum_j \alpha_j \sup_{u_j \in V} \langle \epsilon, u_j \rangle \\ &= \mathbb{E}_\epsilon \left(\sup_{\substack{k \geq 1 \\ \alpha \in \Delta_k}} \sum_j \alpha_j \right) \sup_{u \in V} \langle \epsilon, u \rangle \\ &= \mathrm{URad}(V).\end{aligned}$$