Lecture 18. (Sketch.)

- No class November 7; instead, I'll hold office hours 5-8pm and you can talk to me about projects as long as you wish (and no one kicks you out).
- Project proposal is due Wednesday, November 14 at 3pm.
- See the piazza for project meeting signups.

1. Hoeffding, overfitting, and uniform deviations.

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

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

Applications for a **fixed** f:

- Set $Z_i := \mathbb{1}[f(x_i) \neq y_i] \in [0,1]$; with pr $\geq 1-\delta$,

$$\mathcal{R}_z(f) = \mathbb{E}Z_1 \leq \frac{1}{n}Z_i + \sqrt{\frac{\ln(1/\delta)}{2n}} = \widehat{\mathcal{R}}_z(f) + \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

- Set $Z_i := \ell(-f(x_i)y_i) \in [a,b]$; with pr $\geq 1-\delta$,

$$\mathcal{R}_\ell(f) = \mathbb{E}Z_1 \leq \frac{1}{n}Z_i + (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}} = \widehat{\mathcal{R}}_\ell(f) + (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Remark. For both to hold simultaneously, we need to apply union bound.

Remark. Suppose ℓ is ρ-lipschitz and $|f(x)| \leq R$. Then

$$|\ell(-f(x)y) - \ell(0)| \leq \rho \cdot |-f(x)y - 0| \leq \rho R.$$

Thus $\ell(-f(x)y) \in [\ell(0) - \rho R, \ell(0) + \rho R]$.

So we could have instead said this:

- Suppose $|f| \leq R$ and ℓ is ρ-Lipschitz; with pr $\geq 1-\delta$,

$$\mathcal{R}_\ell(f) \leq \widehat{\mathcal{R}}_\ell(f) + \rho R\sqrt{\frac{2\ln(1/\delta)}{n}}.$$

Why are we saying "fixed f"?

Indeed, why are we fixing it **before** the randomization?

- **Example.** Consider a classifier $\hat{f}$ which memorizes training data S, and outputs -1 otherwise:

$$\hat{f}(x) := \begin{cases} y_i & x = x_i, x_i \in S, \\ -1 & \text{otherwise.} \end{cases}$$

  Consider two situations with $\Pr[Y = +1] = 1$.

  - Suppose marginal on X has finite support. Eventually, this support is memorized and $\widehat{\mathcal{R}}_z(\hat{f}) = 0 = \mathcal{R}_z(\hat{f})$.
  - Suppose marginal on X is continuous. With probability 1, $\widehat{\mathcal{R}}_z(\hat{f}) = 0$ but $\mathcal{R}_z(\hat{f}) = 1$!

What broke Hoeffding's inequality (and its proof)?

- $\hat{f}$ is a *random variable* depending on $S = ((x_i, y_i))_{i=1}^n$. Even if $((x_i, y_i))_{i=1}^n$ are independent, the new random variables $Z_i := \mathbb{1}[\hat{f}(x_i) \neq y_i]$ are not !

These are bad examples of **overfitting**: $\widehat{\mathcal{R}}(\hat{f})$ is small, but $\mathcal{R}(\hat{f})$ is large.

Remarks.

- Can't we fix independence with **two samples** (train $\hat{f}$ with S_1, estimate $\widehat{\mathcal{R}}(\hat{f})$ with S_2)?
  - Yes, but we're using half as much data. **(Project idea.)** Look into (cross-)validation, for which there is still little theory.
- In SGD, didn't we have this correlation issue? Yes, but we still got a bound by (a) restricting the way the algorithm interacts with the data, (b) using a corresponding refined concentration inequality (Azuma for martingales).

Standard fix in learning theory: prove

$$\Pr[\sup_{f\in\mathcal{F}} \mathcal{R}(f) - \widehat{\mathcal{R}}(f) > \epsilon] \le \dots .$$

- This is a **uniform deviation** or **generalization** bound: it controls the random variable $\sup_{f\in\mathcal{F}} \mathcal{R}(f) - \widehat{\mathcal{R}}$, namely it controlls the **devations** $(\mathcal{R}(f) - \widehat{\mathcal{R}}(f))_{f\in\mathcal{F}}$ *uniformly* over $\mathcal{F}$.

Remarks.

- This bound will therefore hold for not just the output of the algorithm but everything else in $\mathcal{F}$.
- **This may seem brutal and loose.** Sometimes it is! To do this properly, the choice of $\mathcal{F}$ should be well-adapted to the algorithm and how it interacts with data; then it *can* be tight.
- There are other approaches: stability (Bousquet and Elisseeff 2002), custom analyses within the algorithm (SGD/Azuma, ordinary least squares, ...).
- **Measure theory note:** that uniform r.v. is broken...

2. Finite classes and primitive covers.

Theorem. Let $\mathcal{F}$ be given, and suppose $\ell(f(x), y) \in [a, b]$ for all $f \in \mathcal{F}$. With probability at least $1-\delta$, every $f \in \mathcal{F}$ satisfies

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

Proof. Suppose $|\mathcal{F}| < \infty$, since otherwise bound is immediate. Define $\delta' := \delta/|\mathcal{F}|$ and $\epsilon := (b-a)\sqrt{\ln(1/\delta')/(2n)}$; for any fixed $f \in \mathcal{F}$,

$$\Pr\left[\mathcal{R}_\ell(f) - \widehat{\mathcal{R}}_\ell(f) \ge \epsilon\right] \le \delta'.$$

Thus ("by union bound")

$$\begin{aligned}\Pr\left[\exists f \in \mathcal{F} \centerdot \mathcal{R}_\ell(f) - \widehat{\mathcal{R}}_\ell(f) \ge \epsilon\right] &\le \sum_{f\in\mathcal{F}} \Pr\left[\mathcal{R}_\ell(f) - \widehat{\mathcal{R}}_\ell(f) \ge \epsilon\right] \\ &\le |\mathcal{F}|\delta' = \delta.\end{aligned}$$

Remarks.

- We can be **adaptive** even here by choosing non-uniform δ_f with $\sum_{f\in\mathcal{F}} \delta_f = \delta$.
- When is this bound tight? Just like the Venn Diagram: when the failure events inhabit different parts of the sample space.

Finite classes are most often invoked by first discretizing or **covering** the function class.

Definition. $\mathcal{G}$ is a **primitive ϵ-cover** of $\mathcal{F}$ over S if: for all $f \in \mathcal{F}$, there exists $g_f \in \mathcal{G}$ so that $\sup_{z \in S} |f(z) - g_f(z)| \le \epsilon$.

Remark.

- So: we take an infinite $\mathcal{F}$, and work with its discretation/cover G.
- Later we'll get to "real" covers, which have much better bounds.
- These primitive covers are **improper:** we do not require $\mathcal{G} \subseteq \mathcal{F}$; we could be covering decision trees with neural networks!

Define $\ell \circ \mathcal{F} := \{(x, y) \mapsto \ell(f(x), y) : f \in \mathcal{F}\}$; we'll often work with covers of $\ell \circ \mathcal{F}$.

Theorem (primitive bound for primitive convers). Suppose $\ell \circ \mathcal{F}$ has primitive ϵ-covers of cardinality N_ϵ over a set S, and $\ell \circ f \in [a, b]$ over S. For any $\epsilon > 0$, with probability $\ge 1 - \delta$ over an iid draw from a distribution supported on S,

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

Proof. Let G_ϵ denote a minimal primitive ϵ-cover with cardinality $\le N_\epsilon$. For any $g \in G_\epsilon$, there must exist $h := \ell \circ f$ with $\sup_{z \in S} |h(z) - g(z)| \le \epsilon$, since otherwise g isn't contributing to the ϵ-cover and G_ϵ is not minimal; therefore

$$\begin{aligned}
&\sup_{z, z' \in S} |g(z) - g(z')| \\
&\le \sup_{z, z' \in S} |g(z) - h(z)| + |h(z) - h(z')| + |h(z') - g(z')| \\
&\le 2\epsilon + (b - a).
\end{aligned}$$

Remarks.

- If ℓ is Lipschitz, we can convert between covers of $\mathcal{F}$ and $\ell \circ \mathcal{F}$ easily. Indeed, if $\mathcal{F}$ is linear with l_2 norm 1, S has l_2 norm 1, and ℓ is 1-Lipschitz,

  $$|\ell(\langle w, -xy\rangle) - \ell(\left\langle w', -xy \right\rangle)| \le |\langle w, -xy\rangle - \left\langle w', xy \right\rangle| \le \|w - w'\|.$$

  Consequently, $N_\epsilon = \mathcal{O}(1/\epsilon^d)$ suffices, and $\ln N_\epsilon = d\mathcal{O}(\ln(1/\epsilon))$. With other tools, we will later remove the dimension dependence.
- If ℓ is not Lipschitz, if for instance it is discontinuous, catastrophically bad things can happen. E.g., if $\ell(f(x), y) = \mathbb{1}[f(x) \ne y]$, then in the above setting the only primitive ϵ-cover with $\epsilon < 2$ has cardinality equal to $\mathbb{R}$, and $\ln N_\epsilon = \infty$!
- We'll fix these issues in subsequent lectures (with "real" covers and other tools as well).

References.

Bousquet, Olivier, and André Elisseeff. 2002. "Stability and Generalization." *JMLR* 2: 499–526.