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.

Remark. Suppose $\ell$ is $\rho$-lopschitz 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 $\ell$ is $\rho$-Lipschitz; with $\Pr \geq 1 - \delta$,

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

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 $Z_i := 1[f(x_i) \neq y_i] \in [0, 1]$; with $\Pr \geq 1 - \delta$,

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

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

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

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

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 $\mathcal{R}_z(\hat{f}) = 0 = \mathcal{R}_z(\hat{f})$.

- Suppose marginal on $X$ is continuous. With probability 1, $\mathcal{R}_z(\hat{f}) = 0$ 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 $Z_i := 1[\hat{f}(x_i) \neq y_i]$ are not!
These are bad examples of overfitting: $\hat{R}(\hat{f})$ is small, but $R(\hat{f})$ is large.

Remarks.

▶ Can’t we fix independence with two samples (train $\hat{f}$ with $S_1$, estimate $\hat{R}(\hat{f})$ 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 $\Pr[\sup_{f \in F} R(f) - \hat{R}(\hat{f}) > \epsilon] \leq \ldots$.

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

Remarks.

▶ We can be adaptive even here by choosing non-uniform $\delta_f$ with $\sum_{f \in F} \delta_f = \delta$.

▶ When is this bound tight? Just like the Venn Diagram: when the failure events inhabit different parts of the sample space.

2. Finite classes and primitive covers.

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

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

**Proof.** Suppose $|F| < \infty$, since otherwise bound is immediate.

Define $\delta' := \delta/|F|$ and $\epsilon := (b - a)\sqrt{\ln(1/\delta')/(2n)}$; for any fixed $f \in F$,  

$$\Pr[R_\ell(f) - \hat{R}_\ell(f) \geq \epsilon] \leq \delta'.$$

Thus (“by union bound”)

$$\Pr[\exists f \in F \cdot R_\ell(f) - \hat{R}_\ell(f) \geq \epsilon] \leq \sum_{f \in F} \Pr[R_\ell(f) - \hat{R}_\ell(f) \geq \epsilon]$$

$$\leq |F|\delta' = \delta.$$
Finite classes are most often invoked by first discretizing or covering the function class.

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

**Remark.**

- So: we take an infinite \( F \), and work with its discretization/cover \( G \).
- Later we’ll get to “real” covers, which have much better bounds.
- These primitive covers are improper: we do not require \( G \subseteq F \); we could be covering decision trees with neural networks!

**Theorem** (primitive bound for primitive covers). Suppose \( \ell : F \to [a, b] \) and \( \ell \circ f \in [a, b] \) over \( S \). For any \( \epsilon > 0 \), with probability \( \geq 1 - \delta \) over an iid draw from a distribution supported on \( S \),

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

**Proof.** Let \( G_\epsilon \) denote a minimal primitive \( \epsilon \)-cover with cardinality \( \leq N_\epsilon \). For any \( g \in G_\epsilon \), there must exist \( h := \ell \circ f \) with \( \sup_{z \in S} |h(z) - g(z)| \leq \epsilon \), since otherwise \( g \) isn’t contributing to the \( \epsilon \)-cover and \( G_\epsilon \) is not minimal; therefore
\[
\begin{align*}
\sup_{z, z' \in S} |g(z) - g(z')| \\
\leq \sup_{z, z' \in S} |g(z) - h(z)| + |h(z) - h(z')| + |h(z') - g(z')| \\
\leq 2\epsilon + (b - a).
\end{align*}
\]

**Remarks.**

- If \( \ell \) is Lipschitz, we can convert between covers of \( F \) and \( \ell \circ F \) easily. Indeed, if \( F \) is linear with \( l_2 \) norm 1, \( S \) has \( l_2 \) norm 1, and \( \ell \) is 1-Lipschitz,

\[
|\ell((w, -xy)) - \ell((w', -xy'))| \leq |\langle w, -xy \rangle - \langle w', xy \rangle| \leq \|w - w'\|.
\]

Consequently, \( N_\epsilon = O(1/\epsilon^d) \) suffices, and \( \ln N_\epsilon = dO(\ln(1/\epsilon)) \).

With other tools, we will later remove the dimension dependence.

- If \( \ell \) is not Lipschitz, if for instance it is discontinuous, catastrophically bad things can happen. E.g., if \( \ell(f(x), y) = 1[f(x) \neq y] \), then in the above setting the only primitive \( \epsilon \)-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.**