Intro to generalization; finite classes, primitive covering

Administrative/Meta:

- Weierstrass's original proof of his approximation theorem was via a smoothing/diffusion argument! And the Bernstein polynomial version we discussed is a discrete time approximation!
- Daniel Hsu talk this week.
- Homework 2 this weekend? Maybe?
- *[poll on project presentation options.]*

Overview

Using the tools from the last class, we obtained the following control on a single predictor $f : \mathbb{R}^d \to \mathbb{R}$: with probability at least $1 - \delta$ over an i.i.d. draw of $((x_i, y_i))_{i=1}^n$ from distribution (X, Y),

$$\mathcal{R}_{\mathrm{z}}(\mathrm{sgn}(f)) \leq \widehat{\mathcal{R}_{\mathrm{z}}}(\mathrm{sgn}(f)) + \sqrt{\frac{1}{2n} \ln\left(\frac{1}{\delta}\right)},$$

where

$$\mathcal{R}_{\mathrm{z}}(g) := \Pr[g(X) \neq Y] \qquad \text{and} \qquad \widehat{\mathcal{R}_{\mathrm{z}}}(g) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[g(x_i) \neq y_i].$$

Our goal is to control not a single predictor, but the output of a training algorithm. The next two lectures will establish our basic tools here as follows.

- **Independence issue / overfitting.** First we will investigate a technical issue will require us to study deviation properties of the random variable $\sup_{f \in \mathcal{F}} \mathcal{R}(f) - \widehat{\mathcal{R}}(f)$ rather than $\mathcal{R}(\hat{f}) - \widehat{\mathcal{R}}(\hat{f})$ where $\hat{f}$ is output by an algorithm.
- **Finite classes and primitive covering numbers.** These are easy to derive and serve as our first generalization bounds.
- **Rademacher complexity / symmetrization.** Then we will introduce a powerful technique, symmetrization, which will form the basis of our approach to generalization. Symmetrization suggests a complexity notion, Rademacher complexity, which will be our basic notion as it lower bounds (expressions involving) many of the other complexity types we will consider (e.g., VC, covering, finite, etc.).

Independence issue / overfitting

Suppose we obtain examples $((x_i, y_i))_{i=1}^n$, feed them to an algorithm, and obtain a predictor $\hat{f}$. What prevents us from applying the earlier proof technique?

Earlier, we defined random variables $Z_i := \mathbf{1}[\mathrm{sgn}(\hat{f}(X_i)) \neq Y_i]$ and $W_i := Z_i - \mathbb{E}(Z_i)$, and then applied Hoeffding's inequality to $(W_1, \ldots, W_n)$. The issue is that $(W_1, \ldots, W_n)$ must be independent, but $\hat{f}$ is a random variable depending on $((X_i, Y_i))_{i=1}^n$: each W_i in general depends on every $((X_i, Y_i))_{i=1}^n$.

Remark. We analyzed stochastic gradient descent in the first lecture, where we showed that the averaged iterate $\bar{w} := (n+1)^{-1} \sum_i w_i$ satisfies, with probability at least $1 - \delta$,

$$f(\bar{w}) - \inf_{\|w\|_2 \leq R} \leq \mathcal{O}\left(\frac{RL\sqrt{\ln(1/\delta)}}{\sqrt{n}}\right)$$

This proof uses the output of an algorithm and seems to sidestep the above independence issues. What gives?

The algorithm invokes Azuma's ienquality, and not Hoeffding, to the martingale difference sequence

$$Z_i := \left\langle \hat{g}_i - \nabla f(w_{i-1}, w_{i-1} - \bar{w} \right\rangle,$$

where stochastic gradient satisfies $\mathbb{E}(\hat{g}_i) = \nabla f(w_{i-1})$; conditioned on $((x_j, y_j)_{j=1}^{i-1}$, $\nabla f(w_{i-1})$ and $\hat{g}_i$ are independent, so we can apply Azuma's inequality to $X_i := \sum_{j \leq i} Z_i$.

As such, this analysis and algorithm *barely* dodged the earlier independence issue. Recall that we also stated an open problem, that repeating even a single example breaks the sgd analysis. So the independence issue is there as well.

Finite classes and uniform covering numbers

One resolution to the preceding independence issues is to be very careful about the way the algorithm uses random samples, as in the martingale analysis of sgd.

Another way, which is what we'll use as the basis for the "generalization" part of the course, is to study the behavior of the random variable

$$\sup_{f \in \mathcal{F}} \mathcal{R}(f) - \widehat{\mathcal{R}}(f).$$

This way, if we run an to pick some $\hat{f} \in \mathcal{F}$, the preceding blanket statement will also imply a guarantee on $\hat{f}$, since we made the guarantee without seeing or depending on the data in any way.

Remarks.

- Whereas controlling a single function gave a deviation of the form $\mathcal{O}(\text{scaling}\sqrt{\ln(1/\delta)/n})$, handling many functions will add a term capturing a notion of "complexity" of $\mathcal{F}$. If the complexity is not chosen appropriately to the number of examples n, then the gap between $\mathcal{R}(f)$ and $\widehat{\mathcal{R}}(f)$ grows; this is usually called overfitting.
- Note that we stopped talking about $\mathcal{R}_{\mathrm{z}}$ and $\widehat{\mathcal{R}_{\mathrm{z}}}$, instead we are discussing

  $$\mathcal{R}(f) := \mathbb{E}(\ell(f(X), Y)) \qquad \text{and} \qquad \widehat{\mathcal{R}}(f) := \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i)).$$

  If for instance $\ell(f(x), y)$ is bounded for every triple (x, y, f) that we care about, then we can still apply Hoeffding and the whole earlier discussion goes through; the only property of the loss $\ell(f(X), Y) = \mathbf{1}[f(X) \neq Y]$ which we used was boundedness. Most of the remaining discussion will be for arbitrary losses.
- For now we're only discussing the one-sided error $\sup_{f \in \mathcal{F}} \mathcal{R}(f) - \widehat{\mathcal{R}}(f)$. Going back to the decomposition of error from the start of this "generalization" part of the couse, this was the guarantee we needed on the output of the algorithm; there was also a step controlling the optimum for $\mathcal{R}$, but there we have no independence issue and can use a direct Chernoff bound. *[future matus: maybe give some refs where controlling only one side is nice, for instance the small ball stuff exploits this? also point to texts that do control the two-sided error.]*

- **Technical note (measure theory).** The random variable $\sup_{f\in\mathcal{F}} \mathcal{R}(f) - \widehat{\mathcal{R}}(f)$ is not measurable in general. There are many technical ways around this but most are not satisfactory, it is a real issue. Ping me if interested. *[future matus: maybe include some refs? That Ramon von Handel paper, then the old paper where Shai Ben-David pointed out the bug, and maybe the usual countability fix used by, say Talagrand?]*

As a first foray into this setting, let's suppose $|\mathcal{F}|$ is finite.

Theorem. Suppose $|\mathcal{F}| < \infty$, and $\ell(f(x), y) \in [0, b]$ for every (x, y, f) with probability 1. Then with probability at least $1 - \delta$ over the i.i.d. draw of $((x_i, y_i))_{i=1}^n$,

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

Proof. Set $\epsilon := b\sqrt{\ln(|\mathcal{F}|/\delta)/(2n)}$. By Hoeffding's inequality, for any $f \in \mathcal{F}$,

$$\Pr[\mathcal{R}(f) - \widehat{\mathcal{R}}(f) \ge \epsilon] \le \frac{\delta}{2|\mathcal{F}|}.$$

By the union bound,

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

□

Finite classes rarely arise directly; instead, we usually discretize some infimite class. This discretization will make it clear that the having $\ln(|\mathcal{F}|)$ (as opposed to $\text{poly}(|\mathcal{F}|)$) is nice.

Given a function class $\mathcal{F}$, a (possibly infinite) set of inputs $\mathcal{Z}$, and a precision level ϵ, Say that a finite subset $\mathcal{G} \subseteq \mathcal{F}$ is a **primitive cover** of $\mathcal{F}$ if for every $f \in \mathcal{F}$ there exists $g_f \in \mathcal{G}$ so that $|g_f(z) - f(z)| \le \epsilon$ for every $z \in \mathcal{Z}$. The **primitive covering number** $\mathcal{N}(\epsilon, \mathcal{F}, \mathcal{Z})$ is infinite if no primitive covers exist, and otherwise it is the size of the smallest primitive cover.

Theorem. Consider losses of the form $\ell : \mathbb{R} \to \mathbb{R}_+$, meaning $\mathcal{R}(f) = \mathbb{E}(\ell(-Yf(X)))$, and ℓ is L-lipschitz. Suppose $Y \in \{-1, +1\}$, and $X \in S$ for some S with probability 1. Suppose $|f(x)| \le b$ for $X \in S$ and $\ell(0) \le Lb$. Then with probability at least $1 - \delta$ over the draw of $((x_i, y_i))_{i=1}^n$,

$$\sup_{f\in\mathcal{F}} \mathcal{R}(f) - \widehat{\mathcal{R}}(f) \le \inf_{\epsilon\ge 0} 2Lb\left(\epsilon + \sqrt{\frac{\ln|\mathcal{N}(b\epsilon, \mathcal{F}, S)| + \ln(1/\delta)}{2n}}\right).$$

Remark.

- The handling of ℓ is awkward here (in particular, the need to worry about $\ell(0)$. We'll get a better version with Rademacher complexity.
- Note that rather than covering $\mathcal{F}$, it is more direct to cover $\{(x, y) \mapsto \ell(-yf(x)) : f \in \mathcal{F}\}$. In particular, every time we decouple things, we add a place for bounds to become loose. Relatedly, there is a notion of *improper covers*, where we drop the requirement that the cover G satisfies $G \subseteq \mathcal{F}$.
- A reasonable choice is $\epsilon := \mathcal{O}(1/\sqrt{n})$, whereby the entire bound becomes $\tilde{\mathcal{O}}(1/\sqrt{n})$.

Proof. Fix any $\epsilon > 0$, suppose $\mathcal{N}(b\epsilon, \mathcal{F}, S) < \infty$ (the bound is for free otherwise), and fix any minimal cover $\mathcal{G}$. For any $g \in \mathcal{G}$, with probability 1, every (X, Y) satisfy

$$|\ell(-Yg(X))| \le |\ell(-Yg(X)) - \ell(0)| + |\ell(0)| \le L|-Yg(X) - 0| + Lb \le 2Lb.$$

Applying the finite class generalization bound to $\mathcal{G}$ gives

$$\sup_{g\in\mathcal{G}} \mathcal{R}(g) - \widehat{\mathcal{R}}(f) \le 2Lb\sqrt{\frac{\ln(\mathcal{N}(\epsilon, \mathcal{F}, S)) + \ln(1/\delta)}{2n}}.$$

Now fix any $f \in \mathcal{F}$, taking $g \in \mathcal{G}$ to denote the approximating element,

$$|\widehat{\mathcal{R}}(f) - \widehat{\mathcal{R}}(g)| \leq \frac{1}{n}\sum_{i=1}^{n} \left|\ell(-yf(x)) - \ell(-yg(x))\right| \leq \frac{L}{n}\sum_{i=1}^{n} \left|(-y)(f(x) - g(x))\right| \leq Lb\epsilon.$$

Similarly, $|\mathcal{R}(f) - \mathcal{R}(g)| \leq \epsilon$. Thus

$$\begin{aligned}\mathcal{R}(f) - \widehat{\mathcal{R}}(f) &= (\mathcal{R}(f) - \mathcal{R}(g)) + (\mathcal{R}(g) - \widehat{\mathcal{R}}(g)) + (\widehat{\mathcal{R}}(g) - \widehat{\mathcal{R}}(f)) \\ &\leq 2Lb\epsilon + 2Lb\sqrt{\frac{\ln(\mathcal{N}(b\epsilon, \mathcal{F}, S)) + \ln(1/\delta)}{2n}}.\end{aligned}$$

Since the bound held for every $\epsilon \geq 0$, it holds over their infimum. $\square$

Example. Consider the case of linear prediction, meaning functions of the form $\mathcal{F} := \{x \mapsto \langle w, x\rangle : \|w\|_2 \leq W\}$, and suppose moreover that $\|x\|_2 \leq X$ with probability 1. Pick a set S of weight vectors $w \in \mathbb{R}^d$ so that for every $\|w\| \leq W$, there exsts $w' \in S$ with $\|w' - w\|_2 \leq \epsilon/X$. A standard estimate is that $|S| \leq \mathcal{O}((XW/\epsilon)^d)$. Moreover, for any $\|x\|_2 \leq X$,

$$|\langle w, x\rangle - \langle w', x\rangle| \leq \|x\|_2\|w - w'\|_2 \leq \epsilon.$$

Thus $\mathcal{N}(\epsilon, \mathcal{F}, \{x \in \mathbb{R}^d : \|x\|_2 \leq X\}) = \mathcal{O}((XW/\epsilon)^d)$, and thus the preceding bound gives (for any 1-Lipschitz loss for simplicity)

$$\begin{aligned}\sup_{f\in\mathcal{F}} \mathcal{R}(f) - \widehat{\mathcal{R}}(f) &\leq \mathcal{O}\left(\inf_{\epsilon>0} XW\epsilon + XW\sqrt{\frac{d\ln(XW/\epsilon) + \ln(2/\delta)}{n}}\right) \\ &= \mathcal{O}\left(XW\sqrt{\frac{d\ln(nXW) + \ln(2/\delta)}{n}}\right),\end{aligned}$$

where the final step simplified via $\epsilon := 1/\sqrt{n}$. By contrast, the sgd bound lacked the extra log term! We'll see how to drop this with Rademacher complexity.

Remark.

- Covers are the original generalization technique; see "metric entropy" of Kolmogorov & Tikhomorov. *[future matus: how can you love Kolmogorov but have not read this paper?!]*
- *[future matus: mention occam bounds, countable classes, pac-bayes?]*

References