Lecture 16

Sample Complexity via Rademacher Averages

Sasha Rakhlin

Oct 31, 2018

Recap

One way to get an upper bound on $\mathbb{E}\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f_{\mathcal{F}})$ for ERM $\widehat{f}_n$ over $\mathcal{F}$ is via uniform deviations:

$$\mathbb{E} \max_{f \in \mathcal{F}} \left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f) \right].$$

(more mathematical name: expected maximum of empirical process)

At this point, there is not algorithm $\widehat{f}_n$ in the picture. Purely a question about $\mathcal{F}$ (and, perhaps, P). If expected maximum is small (as a function of n), we can conclude that $\mathcal{F}$ is "learnable" by ERM.

Recap

To shorten the notation, we introduced $z = (x, y)$, $g = \ell \circ f$, $\mathcal{G} = \ell \circ \mathcal{F}$.

Then we write

$$\mathbb{E} \max_{g \in \mathcal{G}} \left[\mathbb{E} g(Z) - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right].$$

Perhaps we can make things even more transparent by writing

$$\mathbb{E} \max_{g \in \mathcal{G}} U_g$$

where $U_g \triangleq \mathbb{E} g(Z) - \frac{1}{n} \sum_{i=1}^{n} g(Z_i)$ is a *zero-mean* random variable indexed by g (with typical fluctuations $O(1/\sqrt{n})$ due to CLT).

Key point: the larger $\mathcal{G}$ is, the more likely it is that one of U_g takes on a higher value (as in *multiple hypothesis testing*). In particular, if $\mathcal{G}$ is "too large," we cannot control the maximum any longer. *This is the reason we split the learning problem analysis into estimation-approximation tradeoff, so that we can control statistical fluctuations on a smaller set.*

Recap

Again, if $\mathcal{G} = \{g_0\}$, then expected supremum is zero. If $\mathcal{G}$ contains two "different enough" functions, it is $\Theta(1/\sqrt{n})$. How about for countable $\mathcal{G}$? Uncountable $\mathcal{G}$? How about correlations of functions in $\mathcal{G}$? Perhaps not all variables U_g are uncorrelated?

What is the right measure of complexity of $\mathcal{G}$?

Complexity

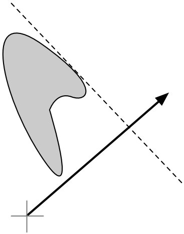
We start by looking at a simpler problem and then relate to above.

Question: given a set $G \subseteq [-1, 1]^n$, what is its "complexity"? Of course, this is an ill-posed question, but let's brainstorm anyway.

Attempt 1: complexity = count elements of G. Not good for uncountable G.

Attempt 2: complexity = volume of G if uncountable. Bad: if G is thin in one dimension, volume goes to zero.

Attempt 3: complexity = average size of projection onto a random Gaussian vector



For a random vector v (say, uniform on unit sphere S^{n-1}), measure

$$\max_{g \in G} \langle v, g \rangle$$

The expected maximum measures an average "width" of G over all directions:

$$\mathbb{E} \max_{g \in G} \langle v, g \rangle$$

If v is a multivariate normal, this quantity is called "Gaussian width."

If v is a vector of independent $\{\pm 1\}$'s (prob $1/2$ each), this quantity is called "Rademacher averages."

We will focus on Rademacher averages as a measure of complexity. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ be sequence of i.i.d. Rademacher random variables (unbiased coin flips with values ± 1).

$$\widehat{\mathscr{R}}_n(G) = \frac{1}{n}\mathbb{E}_{\epsilon_{1:n}} \max_{g \in G} \langle \epsilon, g \rangle = \mathbb{E}_{\epsilon_{1:n}} \max_{g \in G} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i g_i.$$

Verify:

$$\widehat{\mathscr{R}}_n(\{g_0\}) = 0$$

and

$$\widehat{\mathscr{R}}_n(\{-1, 1\}^n) = 1$$

How about

$$\widehat{\mathscr{R}}_n(\{-\mathbf{1}, \mathbf{1}\})$$

where $\mathbf{1}$ is a vector of 1's?

If G is finite,

$$\widehat{\mathscr{R}}_n(G) \leq c\sqrt{\frac{\log|G|}{n}}$$

for some constant c.

This bound can be lose, as it does not take into account "overlaps"/correlations between vectors.

A few properties of Rademacher averages:

- **Convex hull property:**

$$\widehat{\mathscr{R}}_n(G) = \widehat{\mathscr{R}}_n(\mathrm{conv}(G))$$

  where $\mathrm{conv}(G)$ is convex hull of G.

- **Scaling property:** for a constant c,

$$\widehat{\mathscr{R}}_n(c \cdot G) = |c|\widehat{\mathscr{R}}_n(G)$$

- **Subset Property:**

$$G \subseteq F \quad \Rightarrow \quad \widehat{\mathscr{R}}_n(G) \leq \widehat{\mathscr{R}}_n(F)$$

- **Contraction:** if $\phi : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz then

$$\widehat{\mathscr{R}}_n(\phi(G)) \leq L\widehat{\mathscr{R}}_n(G)$$

  where $\phi(G) = \{(\phi(g_1), \ldots, \phi(g_n)) : g \in G\}$, ϕ acting coordinate-wise.

Let B_p^n be a unit ball in $\mathbb{R}^p$:

$$\mathsf{B}_p^n = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p \leq 1\}$$

where

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |\mathbf{x}_i|^p\right)^{1/p}.$$

Then

$$n\widehat{\mathscr{R}}_n(\mathsf{B}_2^n) = \mathbb{E} \max_{\|g\|_2 \leq 1} \langle \epsilon, g \rangle = \mathbb{E} \|\epsilon\|_2 = \mathbb{E}\left(\|\epsilon\|_2^2\right)^{1/2} \leq \left(\mathbb{E} \|\epsilon\|_2^2\right)^{1/2} = \sqrt{n}$$

Hence,

$$\widehat{\mathscr{R}}_n(\mathsf{B}_2^n) \leq \frac{1}{\sqrt{n}}.$$

Show that

$$\widehat{\mathscr{R}}_n(\mathsf{B}_1^n) = \frac{1}{n}.$$

Clearly, $\log|G|$ gives a loose bound here, as random variables $\{\frac{1}{n}\langle \epsilon, e_j\rangle : j = 1, \ldots, n\}$ produce values close to 0.

Homework: for $p \in [1, \infty]$, find upper bound on

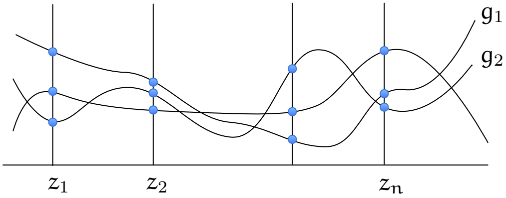
$$\widehat{\mathscr{R}}_n(\mathsf{B}_p^n).$$

Symmetrization

What do these Rademacher averages have to do with our problem of bounding uniform deviations?

Let

$$\mathcal{G}|_{z_{1:n}} = \{(g(z_1), \ldots, g(z_n)) : g \in \mathcal{G}\} \subset \mathbb{R}^n$$



Symmetrization Lemma

Lemma:

$$\mathbb{E} \max_{g \in \mathcal{G}} \left[\mathbb{E} g(Z) - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right] \leq 2 \mathbb{E} \widehat{\mathscr{R}}_n (\mathcal{G}|_{Z_{1:n}})$$

In fact, this is also a lower bound.

Message: to understand uniform deviations, enough to understand richness of sets $\mathcal{G}|_{Z_{1:n}}$.

Equivalent way of writing Rademacher averages on previous slide is to directly write

$$2\mathbb{E}\widehat{\mathscr{R}}_n(\mathcal{G}|_{Z_{1:n}}) = 2\mathbb{E}\max_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(Z_i)$$

where the expectation is both over $Z_{1:n}$ and $\epsilon_{1:n}$.

(make sure this simple rewriting is clear to you)

Symmetrization

Looks like we shifted the difficulty from uniform deviations to the difficulty of estimating Rademacher averages.

The key gain in this step is that we can reason conditionally on $Z_1, \ldots, Z_n$. This is a crucial point that makes the analysis simple in many cases.

To illustrate the last point, consider $\mathcal{G} = \{z \mapsto \mathbf{I}\{z \geq \theta\} : \theta \in \mathbb{R}\}$, a class of thresholds on $\mathbb{R}$. This class is uncountable.

Question: is

$$\mathbb{E} \max_{\theta \in \mathbb{R}} \left[\mathbb{E}\mathbf{I}\{Z \geq \theta\} - \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}\{Z_i \geq \theta\} \right]$$

small? *Not at all clear how to do this directly!*

However, consider the Rademacher averages, conditionally on $Z_1, \ldots, Z_n$:

$$\widehat{\mathscr{R}}_n(\mathcal{G}|_{Z_{1:n}}) = \mathbb{E}_\epsilon \max_{\theta \in \mathbb{R}} \left[\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathbf{I}\{Z_i \geq \theta\} \right]$$

How many distinct vectors are in $\widehat{\mathscr{R}}_n(\mathcal{G}|_{Z_{1:n}})$? Answer: $n + 1$. Hence,

$$\widehat{\mathscr{R}}_n(\mathcal{G}|_{Z_{1:n}}) \leq c \sqrt{\frac{\log(n+1)}{n}}.$$

That was super easy! A more careful analysis removes $\log(n+1)$. This is a version of Kolmogorov's result on uniform closeness of CDF and empirical CDF (a quantified version of Glivenko-Cantelli Theorem).

Binary case

Suppose $\mathcal{G}$ is a class of $\{-1, 1\}$-valued functions. Then $G = \mathcal{G}|_{Z_1,\ldots,Z_n} \subseteq \{-1, +1\}^n$, a subset of n-dimensional hypercube.

Vapnik-Chervonenkis theory says that cardinality of G is at most $O(n^d)$ whenever $n > \text{vc-dim}(\mathcal{G})$.

On the other hand, if $\text{vc-dim}(\mathcal{G}) = \infty$, then for any n there exist $Z_1, \ldots, Z_n$ such that $|G| = 2^n$ and, hence, upper bound via uniform deviations is vacuous.

However, we might not care about existence of these $Z_1, \ldots, Z_n$ if distribution P is 'nice'.

… wait, where is the loss function

So far, we dealt with abstract functions $g \in \mathcal{G}$. But in the learning problem, we take $g = \ell \circ f$ for a fixed loss function and $f \in \mathcal{F}$.

Using contraction property of Rademacher averages, it is easy to remove any L-Lipschitz loss and claim that $\widehat{\mathscr{R}}_n(\ell \circ \mathcal{F})$ are at most $L \cdot \widehat{\mathscr{R}}_n(\mathcal{F})$. For zero-one loss (which is not Lipschitz), we can do an easy direct computation (homework).

Conclusion: to analyze performance of ERM, we can shift focus to uniform deviations, and then to Rademacher averages. There are a variety of techniques for upper bounding Rademacher averages (covering numbers, chaining, scale-sensitive dimensions / VC dimension). We will do some of these calculations when studying neural nets.

Proof of Symmetrization (only for those interested)

Let $\mathscr{S} = \{Z_1, \ldots, Z_n\}$ and $\mathscr{S}' = \{Z'_1, \ldots, Z'_n\}$ (another n i.i.d. datapoints).

$$\mathbb{E}_{\mathscr{S}} \max_{g \in \mathcal{G}} \left[\mathbb{E}_Z g(Z) - \frac{1}{n} \sum_{i=1}^n g(Z_i) \right] = \mathbb{E}_{\mathscr{S}} \max_{g \in \mathcal{G}} \left[\mathbb{E}_{\mathscr{S}'} \left\{ \frac{1}{n} \sum_{i=1}^n g(Z'_i) \right\} - \frac{1}{n} \sum_{i=1}^n g(Z_i) \right]$$

$$\leq \mathbb{E}_{\mathscr{S}, \mathscr{S}'} \max_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{i=1}^n \left\{ g(Z'_i) - g(Z_i) \right\} \right]$$

For any sequence of signs $\epsilon_1, \ldots, \epsilon_n$, distribution of $\frac{1}{n} \sum_{i=1}^n \{g(Z'_i) - g(Z_i)\}$ is the same as distribution of $\frac{1}{n} \sum_{i=1}^n \epsilon_i \{g(Z'_i) - g(Z_i)\}$. Hence, last expression is equal to

$$\mathbb{E}_{\mathscr{S}, \mathscr{S}', \epsilon} \max_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{i=1}^n \epsilon_i \left\{ g(Z'_i) - g(Z_i) \right\} \right]$$

Using $\sup A + B \leq \sup A + \sup B$ and symmetry of random signs ϵ_i, we get upper bound of

$$2\mathbb{E}_{\mathscr{S}, \epsilon} \max_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{i=1}^n \epsilon_i g(Z_i) \right] = 2\mathbb{E}_{\mathscr{S}} \widehat{\mathscr{R}}_n(\mathcal{G}|_{Z_{1:n}})$$

NB: We've been writing uniform deviations and Rademacher averages with a "max" but it should really be "sup".