COMS 4773: Rademacher complexity

Daniel Hsu

March 11, 2024

1 Uniform convergence, again

Recall the uniform convergence theorem.

Theorem 1. *Let $\mathcal{F} \subset \{-1,1\}^{\mathcal{X}}$ have VC dimension $d < \infty$. Let μ be a probability distribution on $\mathcal{X}$, and let μ_n be the empirical distribution based on an iid sample from μ of size n. Then for any $\epsilon > 0$,*

$$\Pr\left(\sup_{f \in \mathcal{F}} \mu_n f - \mu f \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{8}\right) + 2\binom{n}{\leq d}\exp\left(-\frac{n\epsilon^2}{32}\right).$$

This implies that, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{f \in \mathcal{F}} \mu_n f - \mu f \leq O\left(\sqrt{\frac{d \log n + \log(1/\delta)}{n}}\right).$$

A different way to prove Theorem 1 starts by using McDiarmid's inequality. Let $X_1, \dots, X_n$ be an iid sample from μ, and let $\mathcal{F} \subset \{-1,1\}^{\mathcal{X}}$ be a function class. For any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{f \in \mathcal{F}} \mu_n f - \mu f \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \mu_n f - \mu f\right] + \sqrt{\frac{2\log(1/\delta)}{n}}.$$

McDiarmid's inequality applies because the random variable on the left-hand side satisfies the $(c_1, \dots, c_n)$-bounded differences property with $c_i = 2/n$ for all i. So the main task is to bound the expectation on the right-hand side.

Let μ_n' be the empirical distribution on an independent iid sample of size n, $X_1', \dots, X_n'$ (the ghost sample). Instead of using conditional expectation notations, we shall write $\mathbb{E}$ for expectation with respect to $X_{1:n}$, and we write $\mathbb{E}'$ for expectation with respect to $X_{1:n}'$. Then $\mu f = \mathbb{E}'[\mu_n' f]$, and therefore

$$\sup_{f \in \mathcal{F}} \mu_n f - \mu f = \sup_{f \in \mathcal{F}} \mathbb{E}'\left[\mu_n f - \mu_n' f\right] \leq \mathbb{E}'\left[\sup_{f \in \mathcal{F}} \mu_n f - \mu_n' f\right]$$

where the inequality follows by Jensen's inequality and the convexity of the supremum of affine functions. So we have

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \mu_n f - \mu f\right] \leq \mathbb{E}\,\mathbb{E}'\left[\sup_{f \in \mathcal{F}} \mu_n f - \mu_n' f\right].$$

Letting $\sigma = (\sigma_1, \ldots, \sigma_n)$ be a Rademacher random vector, we also have

$$\mathbb{E}\,\mathbb{E}'\left[\sup_{f\in\mathcal{F}} \mu_n f - \mu_n' f\right] = \mathbb{E}\,\mathbb{E}'\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(f(X_i) - f(X_i')\big)\right]$$

where $\mathbb{E}_\sigma$ denotes expectation with respect to σ. Since $X_{1:n}$ and $X_{1:n}'$ have the same distribution, and using the symmetry of σ,

$$\begin{aligned}
\mathbb{E}\,\mathbb{E}'\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(f(X_i) - f(X_i')\big)\right] &\le \mathbb{E}\,\mathbb{E}'\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) + \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (-\sigma_i) f(X_i')\right] \\
&= 2\,\mathbb{E}\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right] \\
&= 2\,\mathbb{E}\,\mathbb{E}_\sigma\left[\sup_{v\in\mathcal{F}(X_{1:n})} \langle \sigma, v\rangle_n\right].
\end{aligned}$$

(Recall our notations for empirical inner product $\langle u, v\rangle_n = \frac{1}{n}\sum_{i=1}^n u_i v_i$ and empirical norm $\|v\|_n = \sqrt{\langle v, v\rangle_n}$.) Since each $v \in \mathcal{F}(X_{1:n})$ has $\|v\|_n = 1$, it follows by Massart's lemma that

$$\mathbb{E}_\sigma\left[\sup_{v\in\mathcal{F}(X_{1:n})} \langle \sigma, v\rangle_n\right] \le \sqrt{\frac{2\log|\mathcal{F}(X_{1:n})|}{n}}.$$

So, if $\mathcal{F}$ has VC dimension $d < \infty$, we conclude by Sauer's lemma that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mu_n f - \mu f\right] \le 2\sqrt{\frac{2\log\binom{n}{\le d}}{n}}.$$

2 Rademacher complexity

Going back a few steps in this development, we have the inequality

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mu_n f - \mu f\right] \le 2\,\mathbb{E}[\mathrm{Rad}_n(\mathcal{F}(X_{1:n}))]. \tag{1}$$

where, for any $V \subseteq \mathbb{R}^n$,

$$\mathrm{Rad}_n(V) = \mathbb{E}_\sigma\left[\sup_{v\in V} \langle \sigma, v\rangle_n\right].$$

The quantity $\mathbb{E}[\mathrm{Rad}_n(\mathcal{F}(X_{1:n}))]$ is called the *(one-sided) Rademacher complexity*[1] of $\mathcal{F}$ (which also depends on μ and n).[2] The Rademacher complexity measures how well functions from $\mathcal{F}$ can be used to "correlate" the sample $X_{1:n}$ with random signs.

[1]The one-sided Rademacher complexity is somewhat non-standard, but it is more convenient in some technical respects and sufficient for our purposes. The usual (two-sided) Rademacher complexity of $\mathcal{F}$ is defined with $|\langle \sigma, v\rangle_n|$ in place of $\langle \sigma, v\rangle_n$ in the definition of Rad_n.

[2]Sometimes $\mathrm{Rad}_n(\mathcal{F}(X_{1:n}))$ itself is called the *empirical Rademacher complexity* of $\mathcal{F}$.

- A "complex" class is one that is able to make this correlation $\mathrm{Rad}_n(\mathcal{F}(X_{1:n}))$ large (in expectation with respect to $X_{1:n}$). For example, the set of *all* $\{-1, 1\}$-valued functions on $\mathcal{X}$ has $\mathrm{Rad}_n(\mathcal{F}(X_{1:n})) = 1$.
- A class that contains only a single function (which should be considered "simple" by any measure...) has $\mathrm{Rad}_n(\mathcal{F}(X_{1:n})) = 0$.

Note that Rademacher complexity is well-defined not just for $\{-1, 1\}$-valued functions, but for any class of functions real-valued functions (although some normalization is needed to make the quantity meaningful). Another feature of Rademacher complexity is that it is sensitive to the data distribution μ. These two "features" of Rademacher complexity distinguish it from VC dimension.

3 Properties of (empirical) Rademacher complexity

Proposition 1. *Let A and B be subsets of $\mathbb{R}^n$. Then the following hold.*

1. *If $A \subseteq B$, then* $\mathrm{Rad}_n(A) \leq \mathrm{Rad}_n(B)$.
2. $\mathrm{Rad}_n(A + B) = \mathrm{Rad}_n(A) + \mathrm{Rad}_n(B)$.
3. $\mathrm{Rad}_n(cA) = |c| \mathrm{Rad}_n(A)$.
4. $\mathrm{Rad}_n(\mathrm{conv}(A)) = \mathrm{Rad}_n(A)$.
5. *(Lipschitz contraction.) Let $\phi_1, \ldots, \phi_n$ be L-Lipschitz $\mathbb{R}$-valued functions on a domain $D \subseteq \mathbb{R}$: i.e., for each $i \in [n]$,*
$$\phi_i(t) - \phi_i(t') \leq L|t - t'| \quad \textit{for all } t, t' \in D.$$
Define
$$\phi(A) = \{(\phi_1(a_1), \ldots, \phi_n(a_n)) : (a_1, \ldots, a_n) \in A\}.$$
If $A \subseteq D^n$, then
$$\mathrm{Rad}_n(\phi(A)) \leq L \, \mathrm{Rad}_n(A).$$

Proof.

1. Since $A \subseteq B$, we have $\{\langle \sigma, v \rangle_n : v \in A\} \subseteq \{\langle \sigma, v \rangle : v \in B\}$; since a supremum over a set can only stay the same or increase by adding more vectors to the set, it follows that $\sup_{v \in A} \langle \sigma, v \rangle_n \leq \sup_{v \in B} \langle \sigma, v \rangle_n$ for all σ, and so $\mathrm{Rad}_n(A) \leq \mathrm{Rad}_n(B)$.
2. Since $A + B = \{a + b : a \in A, b \in B\}$, it follows that $\sup_{v \in A+B} \langle \sigma, v \rangle_n = \sup_{a \in A} \sup_{b \in B} \langle \sigma, a + b \rangle_n = \sup_{a \in A} \langle \sigma, a \rangle_n + \sup_{b \in B} \langle \sigma, b \rangle_n$ for all σ, so $\mathrm{Rad}_n(A + B) = \mathrm{Rad}_n(A) + \mathrm{Rad}_n(B)$.
3. If $c \geq 0$, then $\sup_{v \in cA} \langle \sigma, v \rangle_n = \sup_{a \in A} \langle \sigma, ca \rangle_n = \sup_{a \in A} |c| \langle \sigma, a \rangle_n = |c| \sup_{a \in A} \langle \sigma, a \rangle_n$ for all σ. If $c < 0$, then $\sup_{v \in cA} \langle \sigma, v \rangle_n = \sup_{a \in A} \langle \sigma, ca \rangle_n = \sup_{a \in A} |c| \langle -\sigma, a \rangle_n = |c| \sup_{a \in A} \langle -\sigma, a \rangle_n$ for all σ. In either case, σ and $-\sigma$ have the same distribution, so we conclude that $\mathrm{Rad}_n(cA) = |c| \mathrm{Rad}_n(A)$.
4. By definition, for any $v \in \mathrm{conv}(A)$, we can write $v = c_1 v_1 + \cdots + c_k v_k$ for some $k \in \mathbb{N}$, some $c = (c_1, \ldots, c_k) \in \Delta^{k-1}$, and some $v_1, \ldots, v_k \in A$. For such v, we have $\langle \sigma, v \rangle_n = \sum_{i=1}^k c_i \langle \sigma, v_i \rangle_n \leq \max_{i \in [n]} \langle \sigma, v_i \rangle_n$. So $\sup_{v \in \mathrm{conv}(A)} \langle \sigma, v \rangle_n \leq \sup_{a \in A} \langle \sigma, a \rangle_n$ for all σ, and therefore $\mathrm{Rad}_n(\mathrm{conv}(A)) \leq \mathrm{Rad}_n(A)$. Since $A \subseteq \mathrm{conv}(A)$, it also follows that $\mathrm{Rad}_n(A) \leq \mathrm{Rad}_n(\mathrm{conv}(A))$ by the first property above.

5. We show that by replacing $\phi_1(\cdot)$ with $L\cdot$ can only increase the Rademacher average. Write $\mathbb{E}_{\sigma_1}$ for expectation over σ_1 only (conditioning on $\sigma_2, \dots, \sigma_n$). Then

$$
\begin{aligned}
&\mathbb{E}_{\sigma_1}\left[\sup_{v\in\phi(A)} \langle\sigma, v\rangle_n\right] \\
&= \mathbb{E}_{\sigma_1}\left[\sup_{a\in A} \frac{1}{n}\sum_{i=1}^n \sigma_i\phi_i(a_i)\right] \\
&= \frac{1}{2}\left(\sup_{a\in A} \frac{1}{n}\phi_1(a_1) + \underbrace{\frac{1}{n}\sum_{i=2}^n \sigma_i\phi_i(a_i)}_{B}\right) + \frac{1}{2}\left(\sup_{a'\in A} -\frac{1}{n}\phi_1(a_1') + \underbrace{\frac{1}{n}\sum_{i=2}^n \sigma_i\phi_i(a_i')}_{B'}\right) \\
&= \sup_{a,a'\in A} \frac{1}{2n}\phi_1(a_1) - \frac{1}{2n}\phi_1(a_1') + \frac{1}{2}\big(B + B'\big) \\
&\leq \sup_{a,a'\in A} \frac{L}{2n}|a_1 - a_1'| + \frac{1}{2}\big(B + B'\big) \\
&= \sup_{a,a'\in A} \frac{L}{2n}(a_1 - a_1') + \frac{1}{2}\big(B + B'\big) \\
&= \frac{1}{2}\left(\sup_{a\in A} \frac{1}{n}La_1 + \frac{1}{n}\sum_{i=2}^n \sigma_i\phi_i(a_i)\right) + \frac{1}{2}\left(\sup_{a'\in A} -\frac{1}{n}La_1' + \frac{1}{n}\sum_{i=2}^n \sigma_i\phi_i(a_i')\right) \\
&= \mathbb{E}_{\sigma_1}\left[\sup_{a\in A} \frac{1}{n}\sigma_1(La_1) + \frac{1}{n}\sum_{i=2}^n \sigma_i\phi_i(a_i)\right].
\end{aligned}
$$

Repeat this for $\phi_2, \dots, \phi_n$ to get

$$\mathbb{E}_\sigma\left[\sup_{v\in\phi(A)} \langle\sigma, v\rangle_n\right] \leq \mathbb{E}_\sigma\left[\sup_{v\in LA} \langle\sigma, v\rangle_n\right]$$

(where $LA = \{La : a \in A\}$). Then apply the third property to prove the claim.

□

4 Lower bound

The following is a complement to (1).

Proposition 2. *For any $\mathcal{F} \subset [-1, 1]^{\mathcal{X}}$ and any probability distribution μ on $\mathcal{X}$,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mu_n f - \mu f\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \mu f - \mu_n f\right] \geq \mathbb{E}\,\mathrm{Rad}_n(\mathcal{F}(X_{1:n})) - \sup_{f\in\mathcal{F}}|\mu f|\frac{1}{\sqrt{n}}.$$

Together, (1) and Proposition 2 show that Rademacher complexity essentially characterizes distribution-specific uniform convergence.

Proof of Proposition 2.

$$\begin{aligned}
\operatorname{Rad}_n(\mathcal{F}(X_{1:n})) &= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i)-\mu f)+\frac{1}{n}\sum_{i=1}^n\sigma_i\mu f\right]\\
&\le \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i)-\mu f)\right]+\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i\mu f\right]\\
&\le \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i)-\mu f)\right]+\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}|\mu f|\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\right|\right]\\
&= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i)-\mu f)\right]+\sup_{f\in\mathcal{F}}|\mu f|\,\mathbb{E}_\sigma\left[\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\right|\right]\\
&\le \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i)-\mu f)\right]+\sup_{f\in\mathcal{F}}|\mu f|\frac{1}{\sqrt{n}}.
\end{aligned}$$

Now let $X'_1,\dots,X'_n$ be an independent iid sample from μ, and write $\mathbb{E}'$ for expectation with respect to $X'_{1:n}$. Then we have

$$\begin{aligned}
\mathbb{E}\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i(f(X_i)-\mu f)\right] &= \mathbb{E}\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i\big(f(X_i)-\mathbb{E}'\,f(X'_i)\big)\right]\\
&\le \mathbb{E}\,\mathbb{E}'\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i\big(f(X_i)-f(X'_i)\big)\right]\\
&= \mathbb{E}\,\mathbb{E}'\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i(f(X_i)-\mathbb{E}\,f(X_i))-\sigma_i\big(f(X'_i))-\mathbb{E}'\,f(X'_i)\big)\right]\\
&= \mathbb{E}\,\mathbb{E}'\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(f(X_i)-\mathbb{E}\,f(X_i))-\big(f(X'_i))-\mathbb{E}'\,f(X'_i)\big)\right]\\
&\le \mathbb{E}\,\mathbb{E}'\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(f(X_i)-\mathbb{E}\,f(X_i))+\sup_{f\in\mathcal{F}}-\big(f(X'_i))-\mathbb{E}'\,f(X'_i)\big)\right]\\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\mu_n f-\mu f\right]+\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mu f-\mu_n f\right].
\end{aligned}$$

We conclude that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mu_n f-\mu f\right]+\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mu f-\mu_n f\right]\ge \mathbb{E}\operatorname{Rad}_n(\mathcal{F}(X_{1:n}))-\sup_{f\in\mathcal{F}}|\mu f|\frac{1}{\sqrt{n}}. \qquad \square$$

A simple corollary of Proposition 2 is that, for any $\mathcal{F}\subset[-1,1]^{\mathcal{X}}$ and any probability distribution μ on $\mathcal{X}$,

$$\max\left\{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mu_n f-\mu f\right],\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mu f-\mu_n f\right]\right\}\ge\frac{1}{2}\,\mathbb{E}\operatorname{Rad}_n(\mathcal{F}(X_{1:n}))-\sup_{f\in\mathcal{F}}|\mu f|\frac{1}{2\sqrt{n}}.$$

Notice that if $\mathcal{F}$ is closed under negation (i.e., $\mathcal{F} = \mathcal{F} \cup (-\mathcal{F})$), then both terms in the max are the same. But it is not generally possible to replace the max with min.[3]

[3] To see this, consider the class $\mathcal{F}$ of all characteristic functions $f_S(x) = \mathbb{1}\{x \in S\}$ of finite subsets of $[0, 1]$, and let μ be the uniform distribution on $[0, 1]$. The Rademacher complexity of $\mathcal{F}$ is $1/2$, but $\mathbb{E}\big[\sup_{f \in \mathcal{F}} \mu f - \mu_n f\big] = 0$.