Rademacher process

COMS 4995-1 Spring 2020 (Daniel Hsu)

1 Motivation from empirical processes

Our motivation comes from the study of the supremum of an empirical process. Let $\mathcal{Z}$ be an abstract space, and $\mathcal{F}$ be a family of real-valued functions on $\mathcal{Z}$. For any $z_1, \dots, z_n \in \mathcal{Z}$, we write

$$\mathcal{F}(z_{1:n}) := \{(f(z_1), \dots, f(z_n)) : f \in \mathcal{F}\}$$

to be the behaviors of $\mathcal{F}$ on $z_{1:n} := (z_1, \dots, z_n) \in \mathcal{Z}^n$.

Theorem. Let P be a probability distribution on $\mathcal{Z}$, and let P_n be the empirical distribution on $Z_1, \dots, Z_n \sim_{\text{iid}} P$. Let $\mathcal{F}$ be a family of real-valued functions on $\mathcal{Z}$. Then

$$\mathbb{E} \sup_{f \in \mathcal{F}} |Pf - P_n f| \leq 2 \cdot \mathbb{E} \operatorname{Rad}_n(\mathcal{F}(Z_{1:n}))$$

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

$$\operatorname{Rad}_n(A) := \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{a} \in A} |\langle \boldsymbol{\sigma}, \boldsymbol{a} \rangle_n|.$$

Above, $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$ is vector of iid Rademacher random variables, $\mathbb{E}_{\boldsymbol{\sigma}}$ is expectation conditional on everything except $\boldsymbol{\sigma}$, and $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_n = \frac{1}{n} \sum_{i=1}^n u_i v_i$ is the *normalized* inner product.

1.1 A variation

In some applications, we are primarily interested in a different empirical process, namely

$$\sup_{f \in \mathcal{F}} Pf - P_n f.$$

The same proof establishes

$$\mathbb{E} \sup_{f \in \mathcal{F}} Pf - P_n f \leq 2 \cdot \mathbb{E} \operatorname{Rad}'_n(\mathcal{F}(Z_{1:n}))$$

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

$$\mathrm{Rad}'_n(A) := \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{a} \in A} \langle \boldsymbol{\sigma}, \boldsymbol{a} \rangle_n.$$

(The notation Rad'_n is non-standard.) Relative to the theorem above, the absolute values are omitted both in the empirical process and in Rad'_n. In some texts, similar notation is used for Rad_n and Rad'_n, although there are some subtle differences between the two (notably in the Contraction Lemma, below). Note that for any $A \subseteq \mathbb{R}^n$, we have

$$\mathrm{Rad}'_n(A \cup -A) = \mathrm{Rad}_n(A).$$

1.2 Use with VC classes

Recall that if $\mathcal{F}$ is a family of $\{\pm 1\}$-valued functions on $\mathcal{Z}$, then its VC dimension is the size of the largest set in $\mathcal{Z}$ that is *shattered* by $\mathcal{F}$, i.e., the largest n such that there exists $z_1, \ldots, z_n \in \mathcal{Z}$ such that $|\mathcal{F}(z_{1:n})| = 2^n$. *Sauer's lemma* states that for any $z_1, \ldots, z_n \in \mathcal{Z}$,

$$|\mathcal{F}(z_{1:n})| \leq \sum_{k=0}^{d} \binom{n}{k} =: \binom{n}{\leq d},$$

where d is the VC dimension of $\mathcal{F}$.

Therefore, if $\mathcal{F}$ has VC dimension d, then for any $z_1, \ldots, z_n \in \mathcal{Z}$,

$$\{\langle \boldsymbol{\sigma}, \boldsymbol{a} \rangle_n : \boldsymbol{a} \in \mathcal{F}(z_{1:n})\}$$

is a collection of $\binom{n}{\leq d}$ subgaussian random variables, each with variance proxy $1/n$. By *Massart's finite lemma*, we have

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{a} \in \mathcal{F}(z_{1:n})} |\langle \boldsymbol{\sigma}, \boldsymbol{a} \rangle_n| = \sqrt{\frac{2 \ln(2|\mathcal{F}(z_{1:n})|)}{n}} \leq \sqrt{\frac{2 \ln(2\binom{n}{\leq d})}{n}}.$$

Therefore,

$$\mathbb{E} \sup_{f \in \mathcal{F}} |Pf - P_n f| \leq 2 \cdot \mathbb{E}\, \mathrm{Rad}_n(\mathcal{F}(Z_{1:n})) \leq 2 \cdot \sqrt{\frac{2 \ln(2\binom{n}{\leq d})}{n}}.$$

2 Properties of Rad_n and Rad'_n

There are several properties of Rad_n that are frequently used in learning theory applications. Here are some relatively simple ones:

- If $A \subseteq B$, then $\mathrm{Rad}_n(A) \leq \mathrm{Rad}_n(B)$.
- $\mathrm{Rad}_n(A+B) \leq \mathrm{Rad}_n(A) + \mathrm{Rad}_n(B)$.
- $\mathrm{Rad}_n(cA) = |c| \, \mathrm{Rad}_n(A)$.
- $\mathrm{Rad}_n(\mathrm{absconv}(A)) = \mathrm{Rad}_n(A)$, where

$$\mathrm{absconv}(A) := \mathrm{conv}(A \cup -A).$$

All but the third property are shared by Rad'_n, and the second property can be refined:

- $\mathrm{Rad}'_n(A+B) = \mathrm{Rad}'_n(A) + \mathrm{Rad}'_n(B)$.
- $\mathrm{Rad}'_n(cA) \leq |c| \, \mathrm{Rad}'_n(A)$.

A highly non-obvious property of Rad_n and Rad'_n is given by the *Contraction Lemma*. Let $\phi_1, \ldots, \phi_n$ be L-Lipschitz $\mathbb{R}$-valued functions on $D \subseteq \mathbb{R}$, i.e.,

$$\phi_i(t) - \phi_i(t') \leq L|t - t'| \quad \forall t, t' \in D.$$

For any $\boldsymbol{a} \in D^n$, define

$$\boldsymbol{\phi}(\boldsymbol{a}) := (\phi_1(a_1), \ldots, \phi_n(a_n))$$

and for any $A \subseteq D^n$, define

$$\boldsymbol{\phi}(A) := \{\boldsymbol{\phi}(\boldsymbol{a}) : \boldsymbol{a} \in A\}.$$

Contraction Lemma. For $\phi_1, \ldots, \phi_n$ and A as above, we have

$$\mathrm{Rad}'_n(\boldsymbol{\phi}(A)) \leq L \, \mathrm{Rad}'_n(A).$$

Furthermore, if $\phi_i(0) = 0$ for all i, then

$$\mathrm{Rad}_n(\boldsymbol{\phi}(A)) \leq 2L \, \mathrm{Rad}_n(A).$$

2.1 Proof of the Contraction Lemma for Rad'_n

We just have to show that $\mathrm{Rad}'_n(\boldsymbol{\phi}(A))$ can be bounded above by the same quantity except with ϕ_1 replaced by the function $t \mapsto Lt$. Then, repeatedly doing the same replacement for all other ϕ_i, we will obtain

$$\mathrm{Rad}'_n(\boldsymbol{\phi}(A)) \leq \mathrm{Rad}'_n(LA) \leq L \cdot \mathrm{Rad}'_n(A).$$

Let use write $\mathbb{E}_{\sigma_1}$ to mean the expectation conditional on $\sigma_2, \ldots, \sigma_n$. Then

$$\begin{aligned}
&\mathbb{E}_{\sigma_1} \sup_{\boldsymbol{a} \in A} \langle \boldsymbol{\sigma}, \boldsymbol{a} \rangle_n \\
&= \frac{1}{2n}\left(\sup_{\boldsymbol{a} \in A} \phi_1(a_1) + \underbrace{\sum_{i=2}^{n} \sigma_i \phi_i(a_i)}_{S(\boldsymbol{a}_{2:n})} + \sup_{\boldsymbol{a}' \in A} -\phi_1(a_1) + \underbrace{\sum_{i=2}^{n} \sigma_i \phi_i(a'_i)}_{S(\boldsymbol{a}'_{2:n})} \right) \\
&= \frac{1}{2n}\left(\sup_{\boldsymbol{a}, \boldsymbol{a}' \in A} \phi_1(a_1) - \phi_1(a'_1) + S(\boldsymbol{a}_{2:n}) + S(\boldsymbol{a}'_{2:n}) \right) \\
&\leq \frac{1}{2n}\left(\sup_{\boldsymbol{a}, \boldsymbol{a}' \in A} L|a_1 - a'_1| + S(\boldsymbol{a}_{2:n}) + S(\boldsymbol{a}'_{2:n}) \right) \\
&\overset{(!)}{\leq} \frac{1}{2n}\left(\sup_{\boldsymbol{a}, \boldsymbol{a}' \in A} L(a_1 - a'_1) + S(\boldsymbol{a}_{2:n}) + S(\boldsymbol{a}'_{2:n}) \right) \\
&= \mathbb{E}_{\sigma_1} \sup_{\boldsymbol{a} \in A} \frac{1}{n}\left(L\sigma_1 a_1 + \sum_{i=2}^{n} \sigma_i \phi_i(a_i) \right).
\end{aligned}$$

The first inequality uses the L-Lipschitz property of ϕ_1. To see why the step marked (!) holds, we note that

$$\begin{aligned}
&\sup_{\boldsymbol{a}, \boldsymbol{a}' \in A} L|a_1 - a'_1| + S(\boldsymbol{a}_{2:n}) + S(\boldsymbol{a}'_{2:n}) \\
&= \max\left\{ \sup_{\substack{\boldsymbol{a}, \boldsymbol{a}' \in A \\ a_1 \geq a'_1}} L(a_1 - a'_1) + S(\boldsymbol{a}_{2:n}) + S(\boldsymbol{a}'_{2:n}),\ \sup_{\substack{\boldsymbol{a}, \boldsymbol{a}' \in A \\ a'_1 \geq a_1}} L(a'_1 - a_1) + S(\boldsymbol{a}_{2:n}) + S(\boldsymbol{a}'_{2:n}) \right\} \\
&= \sup_{\substack{\boldsymbol{a}, \boldsymbol{a}' \in A \\ a_1 \geq a'_1}} L(a_1 - a'_1) + S(\boldsymbol{a}_{2:n}) + S(\boldsymbol{a}'_{2:n}) \\
&\leq \sup_{\boldsymbol{a}, \boldsymbol{a}' \in A} L(a_1 - a'_1) + S(\boldsymbol{a}_{2:n}) + S(\boldsymbol{a}'_{2:n}).
\end{aligned}$$

The second equality above holds because the two terms in the max are the same after renaming. (In fact, we can go one step further and upper-bound $\sup_{\boldsymbol{a},\boldsymbol{a}'\in A} L(a_1-a_1')+S(\boldsymbol{a}_{2:n})+S(\boldsymbol{a}'_{2:n})$ by $\sup_{\boldsymbol{a},\boldsymbol{a}'\in A} L|a_1-a_1'|+S(\boldsymbol{a}_{2:n})+S(\boldsymbol{a}'_{2:n})$, which in turn shows that the step marked (!) must hold with equality.) ∎

2.2 A note about the Contraction lemma for Rad_n

The Contraction Lemma for Rad_n, given as Theorem 4.12 in *Probability in Banach Spaces* by Ledoux & Talagrand, is proved using a lot of case analysis, so we omit the proof here. (It would be fantastic if it could be simplified!)

The condition $\phi_i(0) = 0$ is not very onerous. For example, if we had wanted to apply the Contraction Lemma with $\tilde{\phi}_i$ but $\tilde{\phi}_i(0) \neq 0$, we just instead apply it with $\phi_i(t) := \tilde{\phi}_i(t) - \tilde{\phi}_i(0)$, which does satisfy the conditions for the Contraction Lemma. Then

$$
\begin{aligned}
\mathrm{Rad}_n(\tilde{\boldsymbol{\phi}}(A)) &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{a}\in A} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \tilde{\phi}_i(a_i) \right| \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{a}\in A} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi_i(a_i) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \tilde{\phi}_i(0) \right| \\
&= \mathrm{Rad}_n(\boldsymbol{\phi}(A)) + \mathbb{E}_{\boldsymbol{\sigma}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \tilde{\phi}_i(0) \right| .
\end{aligned}
$$

The second term on the right-hand side is just the expected absolute value of the sum of independent subgaussian random variables.