1 Motivation from empirical processes

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

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

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

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

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

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

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

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

1.1 A variation

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

$$\sup_{f \in F} P f - P_n f.$$  

The same proof establishes

$$\mathbb{E} \sup_{f \in F} P f - P_n f \leq 2 \cdot \mathbb{E} \text{Rad}_n'(F(Z_{1:n}))$$
where for $A \subseteq \mathbb{R}^n$,
\[
\text{Rad}'_n(A) := \mathbb{E}_\sigma \sup_{a \in A} \langle \sigma, a \rangle_n.
\]
(The notation $\text{Rad}'_n$ is non-standard.) Relative to the theorem above, the absolute values are omitted both in the empirical process and in $\text{Rad}'_n$. In some texts, similar notation is used for $\text{Rad}_n$ and $\text{Rad}'_n$, 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
\[
\text{Rad}'_n(A \cup -A) = \text{Rad}_n(A).
\]

### 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 $z_1, \ldots, z_n \in \mathcal{Z}$,
\[
|\mathcal{F}(z_{1:n})| \leq \sum_{k=0}^{d} \binom{n}{k} = : \left( \begin{array}{c} n \\ \leq d \end{array} \right),
\]
where $d$ is the VC dimension of $\mathcal{F}$.

Therefore, if $\mathcal{F}$ has VC dimension $d$, then for any $z_1, \ldots, z_n \in \mathcal{Z}$,
\[
\{ \langle \sigma, a \rangle_n : a \in \mathcal{F}(z_{1:n}) \}
\]
is a collection of $\left( \begin{array}{c} n \\ \leq d \end{array} \right)$ subgaussian random variables, each with variance proxy $1/n$. By Massart’s finite lemma, we have
\[
\mathbb{E}_\sigma \sup_{a \in \mathcal{F}(z_{1:n})} |\langle \sigma, a \rangle_n| = \sqrt{\frac{2 \ln(2|\mathcal{F}(z_{1:n})|)}{n}} \leq \sqrt{\frac{2 \ln(2\left( \begin{array}{c} n \\ \leq d \end{array} \right))}{n}}.
\]
Therefore,
\[
\mathbb{E} \sup_{f \in \mathcal{F}} |Pf - P_n f| \leq 2 \cdot \mathbb{E} \text{Rad}_n(\mathcal{F}(Z_{1:n})) \leq 2 \cdot \sqrt{\frac{2 \ln(2\left( \begin{array}{c} n \\ \leq d \end{array} \right))}{n}}.
\]
2 Properties of $\text{Rad}_n$ and $\text{Rad}'_n$

There are several properties of $\text{Rad}_n$ that are frequently used in learning theory applications. Here are some relatively simple ones:

- If $A \subseteq B$, then $\text{Rad}_n(A) \leq \text{Rad}_n(B)$.
- $\text{Rad}_n(A + B) \leq \text{Rad}_n(A) + \text{Rad}_n(B)$.
- $\text{Rad}_n(cA) = |c| \text{Rad}_n(A)$.
- $\text{Rad}_n(\text{absconv}(A)) = \text{Rad}_n(A)$, where
  \[\text{absconv}(A) := \text{conv}(A \cup -A)\].

All but the third property are shared by $\text{Rad}'_n$, and the second property can be refined:

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

A highly non-obvious property of $\text{Rad}_n$ and $\text{Rad}'_n$ 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 $a \in D^n$, define

\[\phi(a) := (\phi_1(a_1), \ldots, \phi_n(a_n))\]

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

\[\phi(A) := \{\phi(a) : a \in A\}.\]

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

\[\text{Rad}'_n(\phi(A)) \leq L \text{Rad}'_n(A).\]

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

\[\text{Rad}_n(\phi(A)) \leq 2L \text{Rad}_n(A).\]
2.1 Proof of the Contraction Lemma for $\text{Rad}_n'$

We just have to show that $\text{Rad}_n'(\phi(A))$ can be bounded above by the same quantity except with $\phi_1$ replaced by the function $t \mapsto Lt$. Then, repeatedly doing the same replacement for all other $\phi_i$, we will obtain

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

Let us write $E_{\sigma_1}$ to mean the expectation conditional on $\sigma_2, \ldots, \sigma_n$. Then

$$E_{\sigma_1} \sup_{a \in A} (\sigma, a)_n$$

$$= \frac{1}{2n} \left( \sup_{a \in A} \phi_1(a_1) + \sum_{i=2}^{n} \sigma_i \phi_i(a_i) + \sup_{a' \in A} -\phi_1(a_1) + \sum_{i=2}^{n} \sigma_i \phi_i(a'_i) \right)$$

$$= \frac{1}{2n} \left( \sup_{a, a' \in A} \phi_1(a_1) - \phi_1(a'_1) + S(a_{2:n}) + S(a'_{2:n}) \right)$$

$$\leq \frac{1}{2n} \left( \sup_{a, a' \in A} L|a_1 - a'_1| + S(a_{2:n}) + S(a'_{2:n}) \right)$$

$^{(1)}$ 

$$\leq \frac{1}{2n} \left( \sup_{a, a' \in A} L(a_1 - a'_1) + S(a_{2:n}) + S(a'_{2:n}) \right)$$

$$= E_{\sigma_1} \sup_{a \in A} \frac{1}{n} \left( L\sigma_1 a_1 + \sum_{i=2}^{n} \sigma_i \phi_i(a_i) \right).$$

The first inequality uses the $L$-Lipschitz property of $\phi_1$. To see why the step marked (!) holds, we note that

$$\sup_{a, a' \in A} L|a_1 - a'_1| + S(a_{2:n}) + S(a'_{2:n})$$

$$= \max \left\{ \sup_{a, a' \in A \atop a_1 \geq a'_1} L(a_1 - a'_1) + S(a_{2:n}) + S(a'_{2:n}), \sup_{a, a' \in A \atop a'_1 \geq a_1} L(a'_1 - a_1) + S(a_{2:n}) + S(a'_{2:n}) \right\}$$

$$= \sup_{a, a' \in A \atop a_1 \geq a'_1} L(a_1 - a'_1) + S(a_{2:n}) + S(a'_{2:n})$$

$$\leq \sup_{a, a' \in A} L(a_1 - a'_1) + S(a_{2:n}) + S(a'_{2:n}).$$

4
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 \( \sup_{a, a' \in A} L(a_1 - a'_1) + S(a_{2:n}) + S(a'_{2:n}) \) by \( \sup_{a, a' \in A} L|a_1 - a'_1| + S(a_{2:n}) + S(a'_{2:n}), \) 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

\[
\text{Rad}_n(\tilde{\phi}(A)) = \mathbb{E}_\sigma \sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \tilde{\phi}_i(a_i) \right| \\
\leq \mathbb{E}_\sigma \sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi_i(a_i) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \tilde{\phi}_i(0) \right| \\
= \text{Rad}_n(\phi(A)) + \mathbb{E}_\sigma \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \tilde{\phi}_i(0) \right|.
\]

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