Rademacher complexity properties 1: Lipschitz losses, finite class lemma

Administrative/Meta:

- Homework 2 is out.
- Project presentations are on December 8 at noon (location TBA); details forthcoming on webpage.

Overview

The culmination of Rademacher/symmetrization was the following bound.

**Theorem.** Let functions $\mathcal{F}$ be given with $|f(z)| \leq c$ almost surely for every $f \in \mathcal{F}$. With probability $\geq 1 - de|ta$ over an i.i.d. draw $S := (Z_1, \ldots, Z_n)$, every $f \in \mathcal{F}$ satisfies

$$E_f \leq \hat{E}f + 2\text{Rad}(\mathcal{F}|S) + 3c\sqrt{\frac{2}{n}\ln(2/\delta)}.$$

where

$$E_f := E(f(Z)),$$

$$\hat{E}f := \frac{1}{n} \sum_{i=1}^{n} f(Z_i),$$

$$\text{Rad}(U) := E_v \sup_{v \in V} \langle \sigma, v \rangle_n,$$

$$\langle a, b \rangle_n := \frac{1}{n} \sum_{i=1}^{n} a_i b_i,$$

$$\mathcal{F}|S := \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\}.$$

Goals for today.

1. Bounds on $\mathcal{R}_\ell(f) := E(\ell(-Yf(X)))$ when $\ell$ is Lipschitz.
2. A Rademacher bound for finite classes. We’ll use this in the next class to discuss Shatter coefficients and VC dimension.

$\mathcal{R}_\ell$ for lipschitz $\ell$

Before getting the tools to work with Lipschitz losses, let’s see how easily we can control $l_2$ bounded linear functions with Rademacher complexity; this will allow us to get bounds for logistic regression soon after.

**Lemma.** Set $X := \sup_{x \in S} \|x\|_2$. Then

$$\text{Rad}(\{x \mapsto \langle w, x \rangle\}|S : \|w\|_2 \leq W) \leq W \sqrt{\sum_{i} \|x_i\|_2^2/n} \leq WX/\sqrt{n}.$$
Proof. For any fixed $\sigma \in \{-1, +1\}^n$, setting $x_\sigma := \sum_{i=1}^n \sigma_i x_i / n$, the equality case of Cauchy-Schwarz grats
\[
\sup_{\|w\|_2 \leq W} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle w, x_i \rangle \sup_{\|w\|_2 \leq W} \langle w, x_\sigma \rangle = \begin{cases} 0 & \text{if } x_\sigma = 0; \\ \langle W x_\sigma / \|x_\sigma\|_2, x_\sigma \rangle & \text{otherwise}. \end{cases} = W \|x_\sigma\|_2.
\]
Now to handle the expectation, we invoke the only inequality in the whole proof: by Jensen’s inequality,
\[
E_\sigma \|x_\sigma\|_2 = E_\sigma \sqrt{\|x_\sigma\|_2^2} \leq \sqrt{E_\sigma \|x_\sigma\|_2^2}.
\]
The rest is again equalities: since $E_\sigma(\sigma_i \sigma_j) = 1[i = j]$,
\[
E \|x_\sigma\|_2^2 = \frac{1}{n^2} \sum_{j=1}^d \sum_{i=1}^n E_\sigma(\sum_{i=1}^n x_{i,j} \sigma_i)^2 = \frac{1}{n^2} \sum_{j=1}^d \sum_{i=1}^n E_\sigma(\sum_{i=1}^n x_{i,j}^2 \sigma_i^2 + \sum_{i \neq j} x_{i,j} x_{i,j} \sigma_i \sigma_j) = \sum_{i=1}^n \|x_i\|_2^2 / n^2
\]
\[\square\]
Remark. This proof is particularly clean for $|| \cdot ||_2$, but in the homework we’ll see a clean trick to handle other norms.

Lemma. Let functions $\ell = (\ell_i)_{i=1}^n$ be given with each $\ell_i : \mathbb{R} \to \mathbb{R}$ $L$-lipschitz, and for any vector $v \in \mathbb{R}^n$ define the coordinate-wise composition $\ell \circ v := (\ell_i(v_i))_{i=1}^n$, and similarly $\ell \circ U := \{\ell \circ v : v \in U\}$. Then
\[
\text{Rad}(\ell \circ U) \leq L \text{Rad}(U).
\]
Remark. The proof seems straightforward but it is a little magical. It uses a step akin to symmetrization, but quite different. The difficulty arises since $|\ell_i(a) - \ell_i(b)| \leq L |a - b|$ by the definition of Lipschitz, but we need to erase that absolute value.

Proof. We will show that we can replace $\ell_1$ with $L$, and the proof is complete by recursing on $\{2, \ldots, n\}$. The idea is that we need to get two terms that depend on $\ell_i$ in order to invoke Lipschitz. Proceeding from the definition of Rad,
\[
\text{Rad}(\ell \circ U) = E_\sigma \sup_{v \in U} \frac{1}{n} \sum_{i} \sigma_i \ell_i(v_i)
\]
\[
= \frac{1}{2n} E_{\sigma_2^n} \left( \sup_{v \in U} \ell_1(v_1) + \sum_{i \geq 2} \sigma_i \ell_i(v_i) + \sup_{w \in U} \ell_1(w_1) + \sum_{i \geq 2} \sigma_i \ell_i(w_i) \right)
\]
\[
\leq \frac{1}{2n} E_{\sigma_2^n} \left( \sup_{v \in U} L |v_1 - w_1| + \sum_{i \geq 2} \sigma_i (\ell_i(v_i) + \ell_i(w_i)) \right)
\]
\[
= \frac{1}{2n} E_{\sigma_2^n} \left( \sup_{v \in U, w \in U} \left( L |v_1 - w_1| + \sum_{i \geq 2} \sigma_i (\ell_i(v_i) + \ell_i(w_i)) \right) \right)
\]
\[
= \frac{1}{2n} E_{\sigma_2^n} \left( \sup_{v \in U, w \in U, v_1 \geq w_1} L (v_1 - w_1) + \sum_{i \geq 2} \sigma_i (\ell_i(v_i) + \ell_i(w_i)) \right)
\]
\[
= \frac{1}{n} E_{\sigma_2^n} \sup_{v \in U} \left( L v_1 + \sum_{i \geq 2} \sigma_i \ell_i(v_i) \right)
\]
The same technique is now applied for $i \in \{2, \ldots, n\}$. \[\square\]
This gives the following useful bound.

**Theorem.** Let functions $F$ and loss $\ell$ be given. Suppose $\ell$ is $L$-Lipschitz and $|\ell(-yf(x))| \leq c$ and $|y| \leq 1$ almost surely. Then with probability at least $1 - \delta$ over $S$, every $f \in F$ satisfies

$$R_\ell(f) \leq \tilde{R}_\ell(f) + 2L\text{Rad}(F|_S) + 3c \sqrt{\frac{2}{n} \ln \left( \frac{1}{\delta} \right)}.$$ 

**Proof.** The proof follows from the previous Rademacher rules by noting $\ell_i(z) := \ell(-yiz)$ is $L$-Lipschitz, just like $\ell$.

**Remark** (logistic regression). Combining these pieces, with $\|x\|_2 \leq X$ and functions $F := \{x \mapsto \langle w, x \rangle : \|w\|_2 \leq W\}$ and nondecreasing $L$-lipschitz loss (e.g., logistic loss $z \mapsto \ln(1 + \exp(z))$ is $1$-lipschitz) we get, with probability $\geq 1 - \delta$, that each $f \in F$ satisfies

$$R_\ell(f) \leq \tilde{R}_\ell(f) + 2WX/\sqrt{n} + 3(LWX + \ell(0)) \sqrt{\frac{2}{n} \ln \left( \frac{1}{\delta} \right)}.$$ 

We had roughly this bound with SGD, but via a very different analysis!

### Finite classes

The main Rademacher tool here is as follows. [In class we also discussed Shatter coefficients and VC dimension, but these will be in the next lecture notes.]

**Theorem** (Massart finite lemma). 

$$\text{Rad}(U) \leq \max_{v \in U} \|v\|_2 \sqrt{2 \ln(|U|)} / n.$$ 

While this has a fancy name, it’s a consequence of the following lemma from homework.

**Lemma.** If $(X_1, \ldots, X_n)$ are $c^2$-subgaussian (but not necessarily independent or identical), then

$$\mathbb{E} \max_i X_i \leq c \sqrt{2 \ln(n)}.$$ 

**Proof.** As in the homework, this follows by noting $\max_i X_i \leq \inf_{t > 0} t^{-1} \ln \sum_i \exp(t X_i)$, using the definition of $c^2$-subgaussian, and optimizing $t$. (The homework problem had $2n$ not $n$, but it controlled for $\max_i |X_i|$.)

Next we need to see how subgaussianity transfer from a random variables to sums of them.

**Lemma.** If $(Z_1, \ldots, Z_n)$ are $c^2_i$-subgaussian and independent, then $\sum_i Z_i/n$ is $\sum_i c^2_i/n^2$-subgaussian.

**Proof.** Set $\mathbf{Z} := \sum_i Z_i/n$. For any $t \in \mathbb{R}$,

$$\mathbb{E}(\exp(tZ_i)) = \prod_i \mathbb{E}(\exp(tZ_i/n)) \leq \prod_i \exp(c^2 t^2/(2n^2)) = \exp(\sum_i c^2_i t^2/2n^2).$$

These lemmas suffice to prove the Rademacher bound.

**Proof** (of Massart finite lemma). For each $v \in V$, define random variable $X_v := \langle \sigma, v \rangle_n$; crucially, the distribution of $X_v$ is determined by the distribution of $\sigma$. Moreover, $\sigma, v_i$ is $c^2_i$-subgaussian by the Hoeffding lemma, meaning $X_v$ is $\|v\|^2_2/n^2$-subgaussian by the preceding lemma, which together with the lemma on maxima of subgaussian distributions gives the bound.

**References**

3