CPSC 532D — 5. RADEMACHER COMPLEXITY

Danica J. Sutherland

University of British Columbia, Vancouver

Fall 2024

Last time (Section 4.2) was our first time showing a uniform convergence bound, one on $\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$, for an infinite $\mathcal{H}$. We can then easily turn that into a bound on the estimation error of ERM, $L_{\mathcal{D}}(\hat{h}_S) - \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$, as in (4.6).

We're now going to develop a technique that's less intuitive, but will show a better result (no $\sqrt{d \log m}$), is somewhat more general, and once you understand it can be easier to use.

We'll start with a bound on the *mean* worst-case generalization gap. That is, we'll show that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le \varepsilon(m).$$

This gives us, for instance, that if $\hat{h}_S$ is an ERM then

$$\mathbb{E}\, L_{\mathcal{D}}(\hat{h}_S) = \underbrace{\mathbb{E}\Big[L_{\mathcal{D}}(\hat{h}_S) - L_S(\hat{h}_S)\Big]}_{\le\varepsilon(m)} + \underbrace{\mathbb{E}\Big[L_S(\hat{h}_S) - L_S(h^*)\Big]}_{\le 0} + \underbrace{\mathbb{E}\Big[L_S(h^*)\Big]}_{=L_{\mathcal{D}}(h^*)} \le L_{\mathcal{D}}(h^*) + \varepsilon(m).$$

We'll use this to prove a high-probability bound on $\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$ in Section 5.3.

5.1 A G-G-G-G-GHOST (SAMPLE)

Using that $L_{\mathcal{D}}(h) = \mathbb{E}_{S\sim\mathcal{D}^m} L_S(h)$:

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m} L_{S'}(h) - L_S(h).$$

S′ here is sometimes called a "ghost sample."

Now, we'll exploit the following general fact:

Lemma 5.1. *Let f_y be a class of functions indexed by y, and X be some random variable. Then when the expectations exist,*

$$\sup_y \mathop{\mathbb{E}}_X f_y(X) \le \mathop{\mathbb{E}}_X \sup_y f_y(X).$$

This should be intuitive, once you think about it a bit: if the optimization can see what particular sample you got, it can "overfit" better than if it has to optimize on average.

Proof. For any y, we have $f_y(X) \le \sup_{y'} f_{y'}(X)$ by definition, no matter the value of X. Taking the expectation of both sides, for any y, $\mathbb{E}_X f_y(X) \le \mathbb{E}_X \sup_{y'} f_{y'}(X)$. So it's also true if we take the supremum over y. □

Applying this, we see that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le \mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m\\ S'\sim\mathcal{D}^m}} \sup_{h\in\mathcal{H}} L_{S'}(h) - L_S(h). \tag{5.1}$$

For more, visit https://cs.ubc.ca/~dsuth/532D/24w1/.

The right-hand-side of (5.1) is itself a natural thing to think about: how much does anything in $\mathcal{H}$ overfit relative to a test set?

Now, $\mathrm{S} = (z_1, \dots, z_m)$ and $\mathrm{S}' = (z_1', \dots, z_m')$ are composed of independent samples from the same distribution. So, if we decided to swap z_3 and z_3', this would still be a "valid," equally likely sample for S and S′. Rademacher complexity is based on this idea.

Watch out that σ_i has nothing to do with a standard deviation or subgaussian parameter σ; we'll refer to the vector $(\sigma_1, \dots, \sigma_m)$ as $\boldsymbol{\sigma}$, or $\vec{\sigma}$ in handwriting. Unfortunate, but no option is great here.

Notationally, let $\sigma_i \in \{-1, 1\}$ for $i \in [m]$, and define $(u_i, u_i') = \begin{cases} (z_i, z_i') & \text{if } \sigma_i = 1 \\ (z_i', z_i) & \text{if } \sigma_i = -1. \end{cases}$ Then, for any choice of $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_m)$, we have

$$\ell(h, z_i') - \ell(h, z_i) = \sigma_i(\ell(h, u_i') - \ell(h, u_i)).$$

So, for any value of S, S′, and $\boldsymbol{\sigma}$, defining $\mathrm{U} = (u_1, \dots, u_m)$ and $\mathrm{U}' = (u_1', \dots, u_m')$ accordingly, we have

$$\mathrm{L}_{\mathrm{S}'}(h) - \mathrm{L}_{\mathrm{S}}(h) = \frac{1}{m} \sum_i \sigma_i [\ell(h, u_i') - \ell(h, u_i)].$$

Since this holds for *any* choice of $\boldsymbol{\sigma}$, it also holds if we pick them at random and then take a mean over that choice. We'll choose them according to a Rademacher distribution, also written Unif(±1), which is 1 half the time and −1 the other half. Thus,

$$\mathop{\mathbb{E}}_{\mathrm{S},\mathrm{S}' \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}} \mathrm{L}_{\mathrm{S}'}(h) - \mathrm{L}_{\mathrm{S}}(h) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{\mathrm{S},\mathrm{S}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\mathrm{U},\mathrm{U}'} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i [\ell(h, u_i') - \ell(h, u_i)] \;\middle|\; \mathrm{S}, \mathrm{S}', \boldsymbol{\sigma} \right].$$

Here we're writing U and U′ as random variables, even though they're actually deterministic conditional on S, S′, and $\boldsymbol{\sigma}$. The marginal distributions of U and U′ are each exactly $\mathcal{D}^m$, though, the same as S and S′. So, it makes sense for us to switch the order of the expectations. $\boldsymbol{\sigma} \mid \mathrm{U}, \mathrm{U}'$ is still just random signs; given $\boldsymbol{\sigma}$ and U, U′, S and S′ become deterministic. This gives us

This switch is allowed by Fubini's theorem as long as $\mathbb{E}\left|\sup_h \mathrm{L}_{\mathrm{S}'}(h) - \mathrm{L}_{\mathrm{S}}(h)\right|$ is finite, which is always true e.g. for a bounded loss.)

$$\mathop{\mathbb{E}}_{\mathrm{S},\mathrm{S}' \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}} \mathrm{L}_{\mathrm{S}'}(h) - \mathrm{L}_{\mathrm{S}}(h) = \mathop{\mathbb{E}}_{\mathrm{U},\mathrm{U}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{\mathrm{S},\mathrm{S}'} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i [\ell(h, u_i') - \ell(h, u_i)] \;\middle|\; \mathrm{U}, \mathrm{U}', \boldsymbol{\sigma} \right].$$

But... S and S′ no longer appear at all, so we can forget about that expectation on the right. Continuing,

This proof technique of introducing a random sign is called symmetrization.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathrm{S},\mathrm{S}' \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}} \mathrm{L}_{\mathrm{S}'}(h) - \mathrm{L}_{\mathrm{S}}(h) &= \mathop{\mathbb{E}}_{\mathrm{U},\mathrm{U}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i [\ell(h, u_i') - \ell(h, u_i)] \\
&\le \mathop{\mathbb{E}}_{\mathrm{U},\mathrm{U}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i \ell(h, u_i') + \sup_{h' \in \mathcal{H}} \frac{1}{m} \sum_i (-\sigma_i) \ell(h', u_i) \right] \\
&= \mathop{\mathbb{E}}_{\mathrm{U},\mathrm{U}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i \ell(h, u_i') + \mathop{\mathbb{E}}_{\mathrm{U},\mathrm{U}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \sup_{h' \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i \ell(h', u_i) \\
&= 2 \mathop{\mathbb{E}}_{\mathrm{S},\mathrm{S}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i \ell(h, z_i) \\
&=: 2 \mathop{\mathbb{E}}_{\mathrm{S},\mathrm{S}' \sim \mathcal{D}^m} \operatorname{Rad}\left((\ell \circ \mathcal{H})|_{\mathrm{S}} \right).
\end{aligned}$$

$\sup_x f(x) + g(x) \le \sup_x f(x) + \sup_{x'} g(x')$

$-\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}$ have the same distribution

Renaming U *to* S

We're defining some notation at the end: $\ell \circ \mathcal{H} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}$ is a set of functions from $\mathcal{Z}$ to $\mathbb{R}$, and $\mathcal{F}|_S$ denotes $\{(f(z_1), \dots, f(z_m)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^m$, so that

$$(\ell \circ \mathcal{H})|_S = \{(\ell(h, z_1), \dots, \ell(h, z_m)) : h \in \mathcal{H}\} \subseteq \mathbb{R}^m.$$

Definition 5.2. The *Rademacher complexity* of a set $V \subseteq \mathbb{R}^m$ is given by

Many sources define Rad *with an absolute value around the sum. This is the more common modern definition, since it makes some things nicer.*

$$\mathrm{Rad}(V) = \mathop{\mathbb{E}}_{\sigma \sim \mathrm{Unif}(\pm 1)^m} \sup_{v \in V} \frac{1}{m} \sum_{i=1}^m \sigma_i v_i = \mathop{\mathbb{E}}_{\sigma \sim \mathrm{Unif}(\pm 1)^m} \sup_{v \in V} \frac{\sigma \cdot v}{m}.$$

One way to think of it is a measure of how much a set V extends in the direction of a random binary vector. $\mathrm{Rad}(\mathcal{F}|_S)$ measures how well $\mathcal{F}$ can align with random signs on the particular set S, or equivalently how well it can separate a random subset of S from the rest.

For intuition, it might be nice to compare to the closely-related *Gaussian complexity* [BM02], which uses $\sigma \sim \mathcal{N}(0, I_m)$ instead of a Rademacher vector. That's maybe more natural to see as a notion of the size of a set: "if I look in a random direction, how far do I get?" (Remember that the norm of a random Gaussian concentrates tightly in high dimensions.) For Rademacher, "looking in any direction" versus "looking along 'binary' directions" isn't so different.

Finally, notice that nothing here depended on the structure of the actual functions $z \mapsto \ell(h, z) \in \ell \circ \mathcal{H}$, and so we've proved the following result for general function classes (rather than just those of the form $\ell \circ \mathcal{H}$).

Theorem 5.3. *For any class $\mathcal{F}$ of functions $f : \mathcal{Z} \to \mathbb{R}$, and any distribution $\mathcal{D}$ over $\mathcal{Z}$ with $S = (z_1, \dots, z_m) \sim \mathcal{D}^m$, we have*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \sup_{f \in \mathcal{F}} \left(\mathop{\mathbb{E}}_{z \sim \mathcal{D}} [f(z)] - \frac{1}{n} \sum_{i=1}^m f(z_i) \right) \le 2 \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \mathrm{Rad}(\mathcal{F}|_S).$$

In particular, in our standard learning setup,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le 2 \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \mathrm{Rad}((\ell \circ \mathcal{H})|_S).$$

5.2 PROPERTIES OF RADEMACHER COMPLEXITY

First, note that

$$\mathrm{Rad}(\{v\}) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sigma \cdot v = 0 :$$

no matter the vector, a singleton set has no complexity. (In terms of generalization: any given hypothesis is equally likely to over- or under-estimate the risk.)

On the other extreme, for the vertices of a hypercube,

$$\mathrm{Rad}(\{-1, 1\}^m) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{v} \sum_{i=1}^m \sigma_i v_i = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} m = 1.$$

As we'll see later (**??**), this is highly related to considering the complexity of the hypothesis class of all possible $\{-1, 1\}$-valued functions; if we tried to do ERM in the set of "all possible classifiers," we'd get that the expected zero-one loss is ≤ 1. Exciting!

Letting $cV = \{cv : v \in V\}$ for any $c \in \mathbb{R}$, we have that

$$\mathrm{Rad}(cV) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{v \in V} \sigma \cdot (cv) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{v \in V} |c| (\mathrm{sign}(c)\sigma) \cdot v = |c| \, \mathrm{Rad}(V) \tag{5.2}$$

since $\mathrm{sign}(c)\sigma$ has the same distribution as σ.

For $V + W = \{v + w : v \in V, w \in W\}$, also called the Minkowski sum, we get

$$\mathrm{Rad}(V+W) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{\substack{v \in V \\ w \in W}} \sigma \cdot (v+w) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{v \in V} \sigma \cdot v + \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{w \in W} \sigma \cdot w = \mathrm{Rad}(V) + \mathrm{Rad}(W).$$

Combined with the fact that $\mathrm{Rad}(\{v\}) = 0$, this means that translating a set by a constant vector doesn't change its complexity.

5.2.1 *Talagrand's contraction lemma*

How do we compute $\mathrm{Rad}(\ell \circ \mathcal{H}|_S)$ for practical losses and hypothesis classes? The first key step is usually to "peel off" the loss, getting a bound in terms of $\mathrm{Rad}(\mathcal{H}|_{S_x})$. We can do that with the following lemma, which is also *very* helpful for bounding $\mathrm{Rad}(\mathcal{H})$ for $\mathcal{H}$ that are defined compositionally, like deep networks.

The major way to do that is with the following results, for Lipschitz losses (Definition 4.6). For example, recall from Lemma 4.9 that logistic loss, used in logistic regression, is 1-Lipschitz.

A 1-Lipschitz function is called a contraction*: it doesn't increase the distance between any points, but (usually) contracts at least some.*

Lemma 5.4 (Talagrand). *Let $\phi : \mathbb{R}^m \to \mathbb{R}^m$ be given by $\phi(t) = (\varphi_1(t_1), \dots, \varphi_m(t_m))$, where each ϕ_i is M-Lipschitz. Then*

$$\mathrm{Rad}(\phi \circ V) = \mathrm{Rad}(\{\phi(v) : v \in V\}) \le M \, \mathrm{Rad}(V).$$

Our proof will be based on the following special case:

Lemma 5.5. *If $\varphi : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz,* $\mathrm{Rad}(\{(\varphi(v_1), v_2, \dots, v_m) : v \in V\}) \le \mathrm{Rad}(V)$.

Proof of Lemma 5.4, assuming Lemma 5.5. First notice that "rotating" the vectors in V doesn't change its complexity, since σ has iid entries:

$$\mathrm{Rad}(\{(v_2, \dots, v_m, v_1) : v \in V\}) = \mathrm{Rad}(V).$$

Now, notice that each component of $\frac{1}{M}\phi(t) = (\frac{1}{M}\varphi_1(t_1), \dots, \frac{1}{M}\varphi_m(t_m))$ is 1-Lipschitz. So, start by applying Lemma 5.5 to V with $\frac{1}{M}\varphi_1$, then rotating, to obtain

$$\mathrm{Rad}\left(\left\{\left(v_2, \dots, v_m, \tfrac{1}{M}\varphi_1(v_1)\right) : v \in V\right\}\right) \le \mathrm{Rad}(V).$$

Repeat these steps with $\frac{1}{M}\varphi_2$, then $\frac{1}{M}\varphi_3$, and so on, until we obtain

$$\mathrm{Rad}\left(\left[\tfrac{1}{M}\phi\right] \circ V\right) \le \mathrm{Rad}(V).$$

Finally, scale by M, which by (5.2) means

$$\mathrm{Rad}(\phi \circ V) = M \, \mathrm{Rad}\left(\left[\tfrac{1}{M}\phi\right] \circ V\right) \le M \, \mathrm{Rad}(V). \qquad \square$$

Proof of Lemma 5.5. Let $\phi(v) = (\varphi(v_1), v_2, \dots, v_m)$ so that $\phi \circ V = \{(\varphi(v_1), v_2, \dots, v_m) :$

$v \in V\}$. Using Python-like notation where $v_{2:}$ means $(v_2, v_3, \dots, v_m) \in \mathbb{R}^{m-1}$, we have

$$\begin{aligned}
m \operatorname{Rad}(\phi \circ V) &= \mathop{\mathbb{E}}_{\sigma} \sup_{v \in V} \left[\sigma_1 \varphi(v_1) + \sigma_{2:} \cdot v_{2:}\right] \\
&= \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v \in V} \left[\varphi(v_1) + \sigma_{2:} \cdot v_{2:}\right] + \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v' \in V} \left[-\varphi(v_1') + \sigma_{2:} \cdot v_{2:}'\right] \\
&= \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v, v' \in V} \left[\varphi(v_1) - \varphi(v_1') + \sigma_{2:} \cdot (v_{2:} + v_{2:}')\right].
\end{aligned}$$

Now, for points arbitrarily close to the supremum, $\varphi(v_1) - \varphi(v_1')$ will always be nonnegative: if it were negative, simply swapping v and v' would make that term positive, and wouldn't affect the rest of the expression, making the objective bigger. Thus we can write

$$\begin{aligned}
m \operatorname{Rad}(\phi \circ V) &= \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v, v' \in V} \left|\varphi(v_1) - \varphi(v_1')\right| + \sigma_{2:} \cdot (v_{2:} + v_{2:}') \\
&\le \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v, v' \in V} \left|v_1 - v_1'\right| + \sigma_{2:} \cdot (v_{2:} + v_{2:}')
\end{aligned}$$

since φ is 1-Lipschitz. Now, notice that the objective of the maximization is identical if we swap v and v', so for any point close to the supremum with $v_1 \le v_1'$, there's an exactly equivalent one with $v_1 \ge v_1'$. Thus

$$\begin{aligned}
m \operatorname{Rad}(\phi \circ V) &\le \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v, v' \in V} v_1 - v_1' + \sigma_{2:} \cdot (v_{2:} + v_{2:}') \\
&= \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{2:}} \left(\sup_{v \in V} \left[v_1 + \sigma_{2:} \cdot v_{2:}\right] + \sup_{v' \in V} \left[-v_1' + \sigma_{2:} \cdot v_{2:}'\right]\right) \\
&= \mathop{\mathbb{E}}_{\sigma} \sup_{v \in V} v \cdot \sigma = m \operatorname{Rad}(V). \qquad \square
\end{aligned}$$

How do we use this? Well, remember that for typical supervised learning losses,

$$\begin{aligned}
(\ell \circ \mathcal{H})|_S &= \{(\ell(h, z_1), \dots, \ell(h, z_m)) : h \in \mathcal{H}\} \\
&= \{(l_{y_1}(h(x_1)), \cdots, l_{y_m}(h(x_m))) : h \in \mathcal{H}\} \\
&= (\mathbf{l}_{S_y} \circ \mathcal{H})|_{S_x},
\end{aligned}$$

where $\mathbf{l}_{S_y}$ is a vectorized version of these losses (like ϕ above) for the vector of particular labels $S_y = (y_1, \dots, y_m)$. Then we have a function of x only, so we apply it to $S_x = (x_1, \dots, x_m)$. If the functions l_{y_i} are all M-Lipschitz, then Talagrand's lemma gives us that

Note that M *here might depend on the particular* S_y*!*

$$\operatorname{Rad}((\ell \circ \mathcal{H})_S) \le M \operatorname{Rad}(\mathcal{H}|_{S_x}). \tag{5.3}$$

5.2.2 *Complexity of bounded linear functions*

When studying covering numbers, we considered logistic regression using the hypothesis class of bounded-norm linear functions,

$$\mathcal{H}_B = \{x \mapsto \langle w, x \rangle : \|w\| \le B\}.$$

To analyze that with Rademacher complexity, the key term is

$$\operatorname{Rad}((\ell_{log} \circ \mathcal{H}_B)|_S) \le \operatorname{Rad}(\mathcal{H}_B|_{S_x}),$$

using (5.3) with Lemma 4.9 that logistic loss is 1-Lipschitz. Now let's bound that latter term:

$$\begin{aligned} m\,\mathrm{Rad}(\mathcal{H}_B|_{S_x}) &= \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\|\le B} \sum_i \sigma_i \langle w, x_i \rangle \\ &= \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\|\le B} \left\langle w, \sum_i \sigma_i x_i \right\rangle \end{aligned}$$

using Cauchy-Shwartz

$$\begin{aligned} &\le \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\|\le B} \|w\| \left\| \sum_i \sigma_i x_i \right\| \\ &= B \mathop{\mathbb{E}}_{\sigma} \left\| \sum_i \sigma_i x_i \right\| \end{aligned}$$

using $(\mathbb{E}\,T)^2 \le \mathbb{E}\,T^2$ *so* $|\mathbb{E}\,T| \le \sqrt{\mathbb{E}\,T^2}$

$$\begin{aligned} &\le B \sqrt{\mathop{\mathbb{E}}_{\sigma} \left\| \sum_i \sigma_i x_i \right\|^2} \\ &= B \sqrt{\mathop{\mathbb{E}}_{\sigma} \sum_{ij} \sigma_i \sigma_j \langle x_i, x_j \rangle} \\ &= B \sqrt{\sum_i \underbrace{\mathbb{E}[\sigma_i^2]}_{1} \|x_i\|^2 + \sum_{i\ne j} \underbrace{\mathop{\mathbb{E}}_{\sigma}[\sigma_i\sigma_j]}_{0} \langle x_i, x_j \rangle}. \end{aligned}$$

Dividing both sides by m, we can rewrite this final inequality as

$$\mathrm{Rad}(\mathcal{H}_B|_{S_x}) \le \frac{B}{\sqrt{m}} \sqrt{\frac{1}{m} \sum_i \|x_i\|^2}, \tag{5.4}$$

so this bound on the complexity depends on the particular S_x that you see, similar to the issue we had with covering numbers.

a.s. is "almost surely" = "with probability one"

One solution (as we did before) is to assume that $\mathcal{D}$ is such that $\|x\| \le C$ (a.s.), something often true in practice. This would imply that $\mathrm{Rad}(\mathcal{H}_B|_{S_x}) \le BC/\sqrt{m}$ (a.s.). Note that this gives us an expected-case bound on the excess error of ERM for logistic regression of

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \frac{2BC}{\sqrt{m}}; \tag{5.5}$$

we'll see in Section 5.3 that, in this case, we can convert this into a bound saying that, with probability at least $1-\delta$,

$$L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \frac{BC}{\sqrt{m}} \left[2 + \sqrt{2 \log \frac{2}{\delta}} \right] = \mathcal{O}_p\left(\frac{BC}{\sqrt{m}} \right). \tag{5.6}$$

Compare this to the covering number-based bound we showed in (4.6):

$$L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \frac{BC}{\sqrt{m}} \left[\frac{1}{\sqrt{2}} + \frac{1}{2}\sqrt{d \log(72m)} + \sqrt{2 \log \frac{2}{\delta}} \right] = \mathcal{O}_p\left(BC \sqrt{\frac{d \log m}{m}} \right).$$

Sometimes, though, we don't want to assume this hard upper bound on $\|x\|$; for example, what if our data is Gaussian? Again using that $\mathbb{E}\,X \le \sqrt{\mathbb{E}\,X^2}$ for nonnegative

X, we can bound the expected value of (5.4) as

$$\mathop{\mathbb{E}}_{S} \operatorname{Rad}(\mathcal{H}_B|_{S_x}) \le \frac{B}{\sqrt{m}} \mathop{\mathbb{E}}_{S} \sqrt{\frac{1}{m} \sum_i \|x_i\|^2} \le \frac{B}{\sqrt{m}} \sqrt{\mathop{\mathbb{E}}_{x} \|x\|^2}. \tag{5.7}$$

This allows for much broader data distributions, as long as you can bound $\mathbb{E}\|x\|^2$. For example, for a Gaussian $x \sim \mathcal{N}(\mu, \Sigma)$ this is $\mathbb{E}\|x\|^2 = \|\mu\|^2 + \operatorname{Tr}(\Sigma)$.

This only works for the average Rademacher complexity, which is the only thing we've seen to care about yet, but in some settings you do want a high-probability bound on $\operatorname{Rad}(\mathcal{H}|_{S_x})$ *rather than an average-case one.*

We've thus shown an average-case estimation error bound for bounded-norm linear problems with Lipschitz losses with a rate of $\mathcal{O}(1/\sqrt{m})$.

5.3 CONCENTRATION

Now let's prove that high-probability bound. We'll need a new tool: *McDiarmid's inequality*, which lets us show concentration of things *other* than sample averages.

Theorem 5.6 ([McD89]). *Let* $X_1, \dots, X_m$ *be independent, and let* $f(X_1, \dots, X_m)$ *be a real-valued function satisfying the* bounded differences *condition*

$$\forall i \in [m]. \qquad \sup_{x_1,\dots,x_m,x_i'} \left| f(x_1, \dots, x_m) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_m) \right| \le c_i.$$

Then, with probability at least $1 - \delta$,

$$f(X_1, \dots, X_m) \le \mathbb{E} f(X_1, \dots, X_m) + \sqrt{\frac{1}{2} \left(\sum_{i=1}^m c_i^2 \right) \log \frac{1}{\delta}}.$$

Proof. Use $X_{i:j}$ to denote $(X_i, \dots, X_j)$. For any $k \in [m]$, freeze some arbitrary values for $x_{1:k-1} = (x_1, \dots, x_{k-1})$. We're going to consider $\mathbb{E}_{X_{k+1:m}} f(x_{1:k-1}, X_k, X_{k+1:m})$ as a random variable, which is random depending *only* on the value of X_k: the earlier arguments are frozen, and the later ones are being averaged over.

This proof has deep connections to martingale methods, but we won't talk any more about that. If you take Nick Harvey's randomized algorithms course, you can learn some more! Or read Section 2.2 of [Wai19] for a very brief intro, or read [McD89].

First, we know this variable is bounded: it can vary only in an interval of length at most c_k. By assumption, for any particular values for $x_{1:k-1}$ and $x_{k+1:m}$,

$$c_k \ge \sup_{x_k} f(x_{1:m}) - \inf_{x_k} f(x_{1:m}).$$

This is true for *any* values of $x_{k+1:m}$, so it's also true on average:

$$\begin{aligned}
c_k &\ge \mathop{\mathbb{E}}_{X_{k+1:m}} \sup_{x_k} f(x_{1:k-1}, x_k, X_{k+1:m}) - \inf_{x_k} f(x_{1:k-1}, x_k, X_{k+1:m}) \\
&\ge \mathop{\mathbb{E}}_{X_{k+1:m}} \sup_{x_k} f(x_{1:k-1}, x_k, X_{k+1:m}) + \sup_{x_k}(-f(x_{1:k-1}, x_k, X_{k+1:m})) && -\inf t = \sup(-t) \\
&= \mathop{\mathbb{E}}_{X_{k+1:m}} \sup_{x_k, x_k'} f(x_{1:k-1}, x_k, X_{k+1:m}) - f(x_{1:k-1}, x_k', X_{k+1:m}) \\
&\ge \sup_{x_k, x_k'} \mathop{\mathbb{E}}_{X_{k+1:m}} f(x_{1:k-1}, x_k, X_{k+1:m}) - f(x_{1:k-1}, x_k', X_{k+1:m}) && \text{Lemma 5.1} \\
&= \sup_{x_k} \mathop{\mathbb{E}}_{X_{k+1:m}} f(x_{1:k-1}, x_k, X_{k+1:m}) - \inf_{x_k} \mathop{\mathbb{E}}_{X_{k+1:m}} f(x_{1:k-1}, x_k, X_{k+1:m}).
\end{aligned}$$

Thus, by Hoeffding's lemma (Proposition 3.5), this variable is $\mathcal{SG}(c_k/2)$. That is, multiplying the definition of subgaussianity (Definition 3.4) by $e^{\lambda \mathbb{E} X}$ for convenience,

$$\mathop{\mathbb{E}}_{X_k} \exp\left(\lambda \mathop{\mathbb{E}}_{X_{k+1:m}} f(x_{1:k-1}, X_k, X_{k+1:m})\right) \le \exp\left(\lambda \mathop{\mathbb{E}}_{X_k} \mathop{\mathbb{E}}_{X_{k+1:m}} f(x_{1:k-1}, X_k, X_{k+1:m}) + \frac{1}{8}\lambda^2 c_k^2\right).$$

This inequality holds for any $x_{1:k-1}$, so let's take the expectation of both sides:

$$\mathop{\mathbb{E}}_{X_{1:k}} \exp\left(\lambda \mathop{\mathbb{E}}_{X_{k+1:m}} f(X_{1:m})\right) \le \mathop{\mathbb{E}}_{X_{1:k-1}} \exp\left(\lambda \mathop{\mathbb{E}}_{X_{k:m}} f(X_{1:m}) + \frac{1}{8}\lambda^2 c_k^2\right).$$

That inequality holds for each choice of k. Let's take the log of each one, and add them all up:

$$\sum_{k=1}^{m} \log \mathop{\mathbb{E}}_{X_{1:k}} \exp\left(\lambda \mathop{\mathbb{E}}_{X_{k+1:m}} f(X_{1:m})\right) \le \sum_{k=1}^{m} \left[\log \mathop{\mathbb{E}}_{X_{1:k-1}} \exp\left(\lambda \mathop{\mathbb{E}}_{X_{k:m}} f(X_{1:m})\right) + \frac{1}{8}\lambda^2 c_k^2\right].$$

Letting $a_k = \log \mathbb{E}_{X_{1:k}} \exp(\lambda \mathbb{E}_{X_{k+1:m}} f(X_{1:m}))$, we have

$$\sum_{k=1}^{m} a_k \le \sum_{k=1}^{m} a_{k-1} + \sum_{k=1}^{m} \frac{1}{8}\lambda^2 c_k^2.$$

Most of the terms cancel, leaving us a_m on the left and a_0 on the right:

$$\log \mathop{\mathbb{E}}_{X_{1:m}} \exp\left(\lambda f(X_{1:m})\right) \le \log \exp\left(\lambda \mathop{\mathbb{E}}_{X_{1:m}} f(X_{1:m})\right) + \frac{1}{8}\lambda^2 \sum_{k=1}^{m} c_k^2.$$

Taking the exponential of both sides and rearranging,

$$\mathop{\mathbb{E}}_{X_{1:m}} \exp\left(\lambda\left(f(X_{1:m}) - \mathop{\mathbb{E}}_{X_{1:m}} f(X_{1:m})\right)\right) \le \exp\left(\frac{1}{2}\lambda^2 \cdot \frac{1}{4}\sum_{k=1}^{m} c_k^2\right).$$

This is exactly the definition of $f(X_{1:m}) \in \mathcal{SG}\left(\frac{1}{2}\sqrt{\sum_{i=1}^{m} c_i^2}\right)$. The Chernoff bound for subgaussians (Proposition 3.8) then tells us that with probability at least $1 - \delta$,

$$f(X_{1:m}) \le \mathbb{E} f(X_{1:m}) + \frac{1}{2}\sqrt{\sum_{i=1}^{m} c_i^2} \cdot \sqrt{2 \log \frac{1}{\delta}}. \qquad \square$$

Considering $-f$ gives an identical form for the lower bound, and a union bound gives an absolute value version by replacing $\frac{1}{\delta}$ with $\frac{2}{\delta}$.

Notice that if $c_i = c$ for all i, then $\sqrt{\sum_{i=1}^{m} c_i^2} = c\sqrt{m}$.

(It's also worth checking for yourself that when $f(X_{1:m}) = \frac{1}{m}\sum_{i=1}^{m} X_i$, you exactly recover the bounded version of Hoeffding's inequality.)

Now that we know McDiarmid's inequality, we can *directly* apply it to get a high-probability bound:

Theorem 5.7. *Suppose that $\ell(h, z) \in [a, b]$ for all h, z. Then, with probability at least $1 - \delta$,*

$$\sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le \mathbb{E} \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] + (b - a)\sqrt{\frac{1}{2m} \log \frac{1}{\delta}}. \tag{5.8}$$

Thus, if $\hat{h}_S$ is an ERM, we have with probability at least $1 - \delta$ that

$$L_{\mathcal{D}}(\hat{h}_S) - \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \le \mathbb{E} \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] + (b - a)\sqrt{\frac{2}{m} \log \frac{2}{\delta}}. \tag{5.9}$$

Proof. Let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$. Now, we have

$$L_{\mathcal{D}}(h) - L_S(h) = L_{\mathcal{D}}(h) - L_{S^{(i)}}(h) + L_{S^{(i)}}(h) - L_S(h);$$

thus, expanding out $L_{S(i)}(h) - L_S(h)$,

$$\sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] - \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_{S^{(i)}}(h)] \le \sup_{h \in \mathcal{H}} \frac{1}{m} [\ell(h, z') - \ell(h, z)] \le \frac{b - a}{m}$$

because the loss is bounded. The same holds in the other direction:

$$\sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_{S^{(i)}}(h)] - \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] \le \sup_{h \in \mathcal{H}} \frac{1}{m} [\ell(h, z) - \ell(h, z')] \le \frac{b - a}{m}.$$

Therefore the worst-case generalization gap, $f(S) = \sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$, satisfies the bounded differences condition with $c = (b - a)/m$. Equation (5.8) follows by applying McDiarmid.

The other result follows as usual for our ERM bounds: we know that for any $h^* \in \mathcal{H}$,

$$\begin{aligned}
L_{\mathcal{D}}(\hat{h}_S) &\le L_S(\hat{h}_S) + \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] \\
&\le L_S(\hat{h}_S) + \mathbb{E} \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] + (b - a)\sqrt{\frac{1}{2m} \log \frac{2}{\delta}} && \text{(5.8), } \textit{w/ prob. } 1 - \delta/2 \\
&\le L_S(h^*) + \mathbb{E} \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] + (b - a)\sqrt{\frac{1}{2m} \log \frac{2}{\delta}} && \textit{definition of ERM} \\
&\le L_{\mathcal{D}}(h^*) + (b - a)\sqrt{\frac{1}{2m} \log \frac{2}{\delta}} + \mathbb{E} \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] + (b - a)\sqrt{\frac{1}{2m} \log \frac{2}{\delta}}, && \textit{Hoeffding, w/ prob. } 1 - \delta/2
\end{aligned}$$

and the result follows since h^* was arbitrary. □

For bounded-norm bounded-data logistic regression, using (5.5) and (4.4) in (5.9) gives (5.6).

REFERENCES

[BM02] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research* 3 (2002), pages 463–482.

[McD89] Colin McDiarmid. On the method of bounded differences. *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989, pages 148–188.

[Wai19] Martin Wainwright. *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge University Press, 2019.