CPSC 532D — 5. RADEMACHER COMPLEXITY

Danica J. Sutherland

University of British Columbia, Vancouver

Fall 2025

Last time (Chapter 4) 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, $L_{\mathcal{D}}(\hat{h}_S) - \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$.

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 sometimes 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/25w1/.

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, $S = (z_1, \dots, z_m)$ and $S' = (z'_1, \dots, z'_m)$ 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 $U = (u_1, \dots, u_m)$ and $U' = (u'_1, \dots, u'_m)$ accordingly, we have

$$L_{S'}(h) - L_S(h) = \frac{1}{m} \sum_i \sigma_i [\ell(h, u'_i) - \ell(h, u_i)].$$

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, first writing out the definition and then using this distribution over U, U′ that we made to be exactly the same thing,

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

$$\begin{aligned} \mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}} L_{S'}(h) - L_S(h) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i [\ell(h, z'_i) - \ell(h, z_i)] \right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{U,U'} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i [\ell(h, u'_i) - \ell(h, u_i)] \,\middle|\, S, S', \boldsymbol{\sigma} \right]. \end{aligned}$$

Here we're writing U and U′ as random variables; they're deterministic conditional on S, S′, and $\boldsymbol{\sigma}$, but that's fine. It then makes sense to talk about the joint distribution of (S, S′, $\boldsymbol{\sigma}$, U, U′):

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

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

Now we're taking the expectation with respect to five different random variables, but the value of S and S′ don't actually show up in the thing we're taking the expectiation of! So, we can just forget about the irrelevant variables:

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

Now we have to ask what the distribution of $(\boldsymbol{\sigma}, U, U')$ is that we're averaging with respect to. Well… if I don't know S or S′, then it's just $U, U' \sim \mathcal{D}^m$, $\boldsymbol{\sigma} \sim \text{Unif}(\pm 1)^m$, because we've forgotten about everything that couples any of the variables. Thus

$$\mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}} L_{S'}(h) - L_S(h) = \mathop{\mathbb{E}}_{U,U' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i [\ell(h, u'_i) - \ell(h, u_i)]$$

$$\leq \mathop{\mathbb{E}}_{U,U'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{h\in\mathcal{H}} \frac{1}{m} \sum_i \sigma_i \ell(h, u'_i) + \sup_{h'\in\mathcal{H}} \frac{1}{m} \sum_i (-\sigma_i) \ell(h', u_i) \right]$$

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

$$= \mathop{\mathbb{E}}_{U'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{\sigma} \sup_{h'\in\mathcal{H}} \frac{1}{m} \sum_i \sigma_i \ell(h', u'_i) + \mathop{\mathbb{E}}_{U\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{\sigma} \sup_{h\in\mathcal{H}} \frac{1}{m} \sum_i \sigma_i \ell(h, u_i)$$

since $-\sigma$ and σ have the same distribution

$$= 2 \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{\sigma} \sup_{h\in\mathcal{H}} \frac{1}{m} \sum_i \sigma_i \ell(h, z_i).$$

Renaming U, U′ to S

Let's define some notation to express this more compactly. First, let $\ell \circ \mathcal{H} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}$ be a set of functions from $\mathcal{Z}$ to $\mathbb{R}$, representing each hypothesis by the loss it might achieve on any input. Then, for an arbitrary class of functions $f : \mathcal{X} \to \mathcal{Y}$, let $\mathcal{F}|_S = \{(f(z_1), \ldots, f(z_m)) : f \in \mathcal{F}\} \subseteq \mathcal{Y}^m$: replace the functions with the vector of outputs they get on each thing in S. Combining these,

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

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

$$\mathrm{Rad}(V) = \mathop{\mathbb{E}}_{\sigma\sim\mathrm{Unif}(\pm1)^m} \sup_{v\in V} \frac{1}{m} \sum_{i=1}^{m} \sigma_i v_i = \mathop{\mathbb{E}}_{\sigma\sim\mathrm{Unif}(\pm1)^m} \sup_{v\in V} \frac{\sigma \cdot v}{m}.$$

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

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.

For intuition, compare to the following "spherical complexity":

$$\mathop{\mathbb{E}}_{s\sim\mathrm{Unif}(\{w:\|w\|=1\})} \sup_{v\in V} s \cdot v,$$

which first projects the set V along the random direction s, then asks "how far" it's possible to get along that direction. This seems somewhat sensible as a notion of size: it's a little strange that we don't check relative to the "centre" of the set, but if we think of the set as centred at the origin, this seems reasonable. (In fact its centre won't matter, as we'll see shortly.) Looking in a random "binary" direction, versus looking in a totally random direction, doesn't make a big difference, particularly in high dimensions; A2 Q3 explores that.

People don't use this notion of complexity, as far as I know, but they do use the Gaussian complexity *where $s \sim \mathcal{N}(0, I_m)$ [BM02]. Since $\|s\|$ concentrates tightly for large m, Gaussian and spherical complexity are extremely similar, and both are similar to Rademacher.*

So: why does it make sense to ask about the "size" of the set $(\ell \circ \mathcal{H})|_S$? This set determines how much "flexibility" the class $\mathcal{H}$ has, with respect to the loss ℓ, on the sample S. For example, imagine we're looking at bounded-weight linear regression, $\ell(h, (x, y)) = (h(x) - y)^2$ with $\mathcal{H} = \{x \mapsto w \cdot x : \|w\| \leq B\}$, and S contains two points. If those two points are far away from each other, many multiples of B, then predictors can do lots of things: they can get high loss on both points, low loss on one and high on the other, low loss on both points. Thus $(\ell \circ \mathcal{H})|_S$ is a large region of the positive quadrant in $\mathbb{R}^2$; this flexible class can do basically whatever it wants, and so it's natural to expect that maybe this class can overfit a lot. On the other hand, suppose the two x points in S are very close to one another, say B/10, and have the same y value. Then the predictions on the two points are constrained to be very similar, and so the model must achieve about the same loss on each point. This set is much "smaller": it's just a narrow region near the diagonal in $\mathbb{R}^2$. It's also much less possible for the model to overfit here, and so it's natural to expect the generalization gap to be smaller. This same intuition holds for any m.

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*

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

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

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

i.e. 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 4.2). For example, recall from Lemma 4.5 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*

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

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

$$\begin{aligned}(\ell \circ \mathcal{H})|_{\mathrm{S}} &= \{(\ell(h, z_1), \dots, \ell(h, z_m)) : h \in \mathcal{H}\} \\ &= \{(l_{y_1}(h(x_1)), \dots, l_{y_m}(h(x_m))) : h \in \mathcal{H}\} \\ &= \mathbf{l}_{\mathrm{S}_y} \circ \{(h(x_1), \dots, h(x_m)) : h \in \mathcal{H}\} = \mathbf{l}_{\mathrm{S}_y} \circ (\mathcal{H}|_{\mathrm{S}_x}).\end{aligned}$$

Here we apply $\mathbf{l}_{\mathrm{S}_y}$, a vectorized version of $(l_{y_1}, \dots, l_{y_m})$ for the vector of particular labels $\mathrm{S}_y = (y_1, \dots, y_m)$, to the vector of predictions on $\mathrm{S}_x = (x_1, \dots, x_m)$ for each $h \in \mathcal{H}$. If the functions l_{y_i} are each M-Lipschitz, then Talagrand's lemma gives us

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

$$\mathrm{Rad}((\ell \circ \mathcal{H})_{\mathrm{S}}) \le \mathrm{M}\,\mathrm{Rad}(\mathcal{H}|_{\mathrm{S}_x}). \tag{5.3}$$

We'll prove Lemma 5.4 based on the following special case:

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

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:

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

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

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

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

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

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

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

Proof of Lemma 5.5. Let $\phi(v) = (\varphi(v_1), v_2, \dots, v_m)$ so that $\phi \circ \mathrm{V} = \{(\varphi(v_1), v_2, \dots, v_m) : v \in \mathrm{V}\}$. Using Python-like notation where $v_{2:}$ means $(v_2, v_3, \dots, v_m) \in \mathbb{R}^{m-1}$, we have

$$\begin{aligned}m\,\mathrm{Rad}(\phi \circ \mathrm{V}) &= \mathop{\mathbb{E}}_{\sigma} \sup_{v \in \mathrm{V}} [\sigma_1 \varphi(v_1) + \sigma_{2:} \cdot v_{2:}] \\ &= \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v \in \mathrm{V}} [\varphi(v_1) + \sigma_{2:} \cdot v_{2:}] + \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v' \in \mathrm{V}} [-\varphi(v'_1) + \sigma_{2:} \cdot v'_{2:}] \\ &= \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v, v' \in \mathrm{V}} [\varphi(v_1) - \varphi(v'_1) + \sigma_{2:} \cdot (v_{2:} + v'_{2:})].\end{aligned}$$

Now, the one thing we know about ϕ is that it's 1-Lipschitz: $|\phi(v_1) - \phi(v'_1)| \le |v_1 - v'_1|$. This also implies that $\phi(v_1) - \phi(v'_1) \le |v_1 - v'_1|$: if the left-hand side is positive, it's the same as the absolute value version, while if it's negative, it's automatically less

than the nonnegative right-hand-side. Thus we have

$$m\,\mathrm{Rad}(\phi \circ \mathrm{V}) \le \frac{1}{2}\mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v,v' \in \mathrm{V}} |v_1 - v'_1| + \sigma_{2:} \cdot (v_{2:} + v'_{2:})$$

$$= \frac{1}{2}\mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v,v' \in \mathrm{V}} v_1 - v'_1 + \sigma_{2:} \cdot (v_{2:} + v'_{2:}).$$

Where did the second line come from? The objective of the first 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 $v_1 \ge v'_1$. This means that, while the set of maximizers is different between the two lines, the value of the maximum is the same. Thus, rewriting the last line to break up the two now-independent maximizations,

$$m\,\mathrm{Rad}(\phi \circ \mathrm{V}) \le \frac{1}{2}\mathop{\mathbb{E}}_{\sigma_{2:}}\left(\sup_{v \in \mathrm{V}} [v_1 + \sigma_{2:} \cdot v_{2:}] + \sup_{v' \in \mathrm{V}} [-v'_1 + \sigma_{2:} \cdot v'_{2:}]\right).$$

But this expression is exactly the same as

$$m\,\mathrm{Rad}(\phi \circ \mathrm{V}) \le \mathop{\mathbb{E}}_{\sigma_1}\mathop{\mathbb{E}}_{\sigma_{2:}} \sup_{v \in \mathrm{V}} \sigma_1 v_1 + \sigma_{2:} \cdot v_{2:} = \mathrm{Rad}(\mathrm{V}). \qquad \square$$

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,

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

To analyze that with Rademacher complexity, the key term is

$$\mathrm{Rad}((\ell_{log} \circ \mathcal{H}_\mathrm{B})|_\mathrm{S}) \le \mathrm{Rad}(\mathcal{H}_\mathrm{B}|_{\mathrm{S}_x}),$$

using (5.3) with Lemma 4.5 that logistic loss is 1-Lipschitz. Now let's bound that latter term:

$$m\,\mathrm{Rad}(\mathcal{H}_\mathrm{B}|_{\mathrm{S}_x}) = \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\| \le \mathrm{B}} \sum_i \sigma_i \langle w, x_i\rangle$$

$$= \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\| \le \mathrm{B}} \left\langle w, \sum_i \sigma_i x_i \right\rangle$$

using Cauchy-Shwartz

$$\le \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\| \le \mathrm{B}} \|w\| \left\|\sum_i \sigma_i x_i\right\|$$

$$= \mathrm{B}\mathop{\mathbb{E}}_{\sigma}\left\|\sum_i \sigma_i x_i\right\|$$

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

$$\le \mathrm{B}\sqrt{\mathop{\mathbb{E}}_{\sigma}\left\|\sum_i \sigma_i x_i\right\|^2}$$

$$= \mathrm{B}\sqrt{\mathop{\mathbb{E}}_{\sigma}\sum_{ij} \sigma_i \sigma_j \langle x_i, x_j\rangle}$$

$$= \mathrm{B}\sqrt{\sum_i \underbrace{\mathbb{E}[\sigma_i^2]}_{1}\|x_i\|^2 + \sum_{i \ne j} \underbrace{\mathop{\mathbb{E}}_{\sigma}[\sigma_i \sigma_j]}_{0}\langle x_i, x_j\rangle} = \mathrm{B}\sqrt{\sum_i \|x_i\|^2}.$$

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

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

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

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

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

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

we'll see in (5.9) that we can convert this into a high-probability bound showing $\mathrm{L}_{\mathcal{D}}(\hat{h}_{\mathrm{S}}) - \mathrm{L}_{\mathcal{D}}(h^*) = \mathcal{O}_p\left(\frac{\mathrm{BC}}{\sqrt{m}}\right)$. This is indeed notably better than the covering number-based bound we showed in (4.4) of $\mathrm{L}_{\mathcal{D}}(\hat{h}_{\mathrm{S}}) - \mathrm{L}_{\mathcal{D}}(h^*) = \mathcal{O}_p\left(\frac{\mathrm{BC}}{\sqrt{m}}\sqrt{d\log m}\right)$.

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 $\mathbb{E}\,\mathrm{X} \le \sqrt{\mathbb{E}\,\mathrm{X}^2}$, we can bound the expected value of (5.4) as

$$\mathop{\mathbb{E}}_{\mathrm{S}} \mathrm{Rad}(\mathcal{H}_{\mathrm{B}}|_{\mathrm{S}_x}) \le \frac{\mathrm{B}}{\sqrt{m}} \mathop{\mathbb{E}}_{\mathrm{S}} \sqrt{\frac{1}{m}\sum_i \|x_i\|^2} \le \frac{\mathrm{B}}{\sqrt{m}}\sqrt{\mathop{\mathbb{E}}_{x}\|x\|^2}. \tag{5.6}$$

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 $\mathrm{Rad}(\mathcal{H}|_{\mathrm{S}_x})$ rather than an average-case one.

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 $\mathbb{E}\|x\|^2 = \|\mu\|^2 + \mathrm{Tr}(\Sigma)$.

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 $\mathrm{X}_1, \dots, \mathrm{X}_m$ be independent, and let $f(\mathrm{X}_1, \dots, \mathrm{X}_m)$ be a real-valued function satisfying the* bounded differences *condition*

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

Then $f(\mathrm{X}_1, \dots, \mathrm{X}_m)$ is $\mathcal{SG}\left(\frac{1}{2}\sqrt{\sum_{i=1}^m c_i^2}\right)$, and so with probability at least $1-\delta$,

$$f(\mathrm{X}_1,\dots,\mathrm{X}_m) \le \mathbb{E}\, f(\mathrm{X}_1,\dots,\mathrm{X}_m) + \sqrt{\frac{1}{2}\left(\sum_{i=1}^m c_i^2\right)\log\frac{1}{\delta}}.$$

Before proving this, notice that 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 $\frac{2}{\delta}$.

Also, if $c_i = c$ for all i, then $\sqrt{\sum_{i=1}^m c_i^2} = c\sqrt{m}$. Typically we'll have something like

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

$c = \mathcal{O}(1/m)$ which gives us our usual $\mathcal{O}_p(1/\sqrt{m})$ concentration.

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 [Har23], you can learn some more! Or read Section 2.2 of [Wai19] for a very brief intro, or read [McD89].

Proof. Use $X_{i:j}$ to denote $(X_i, \dots, X_j)$. We're going to try to compare $f(x_{1:m})$ to its mean, to show subgaussianity. In particular, we want to upper bound

$$\mathbb{E} \exp\left(\lambda(f(X_{1:m}) - \mathbb{E} f(X_{1:m}))\right).$$

To do that, we're going to go "step by step" through each of the X_i, imagining that we've "observed" the first $i-1$ entries of X. We'd like to break out *only* the contribution of X_i, though, so let's just take an average over all the following entries: $\mathbb{E}_{X_{i+1:m}} f(x_{1:i-1}, X_i, X_{i+1:m})$. This overall expression is random, depending only on the value of X_i.

First, we'll show that this variable is bounded: it can vary only in an interval of length at most c_i. By assumption, for any particular values for $x_{1:i-1}$ and $x_{i+1:m}$,

$$c_i \geq \sup_{x_i} f(x_{1:m}) - \inf_{x_i} f(x_{1:m}).$$

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

$$\begin{aligned}
c_i &\geq \mathop{\mathbb{E}}_{X_{i+1:m}} \left[\sup_{x_i} f(x_{1:i-1}, x_i, X_{i+1:m}) - \inf_{x_i} f(x_{1:i-1}, x_i, X_{i+1:m})\right] \\
&= \mathop{\mathbb{E}}_{X_{i+1:m}} \left[\sup_{x_k} f(x_{1:i-1}, x_i, X_{i+1:m}) + \sup_{x_k}(-f(x_{1:i-1}, x_i, X_{i+1:m}))\right] && -\inf t = \sup(-t) \\
&= \mathop{\mathbb{E}}_{X_{i+1:m}} \sup_{x_i, x_i'} \left[f(x_{1:i-1}, x_i, X_{i+1:m}) - f(x_{1:i-1}, x_i', X_{i+1:m})\right] \\
&\geq \sup_{x_i, x_i'} \mathop{\mathbb{E}}_{X_{i+1:m}} \left[f(x_{1:i-1}, x_i, X_{i+1:m}) - f(x_{1:i-1}, x_i', X_{i+1:m})\right] && \text{Lemma 5.1} \\
&= \sup_{x_i} \mathop{\mathbb{E}}_{X_{i+1:m}} f(x_{1:i-1}, x_i, X_{i+1:m}) - \inf_{x_i} \mathop{\mathbb{E}}_{X_{i+1:m}} f(x_{1:i-1}, x_i, X_{i+1:m}).
\end{aligned}$$

Thus, by Hoeffding's lemma (Proposition 3.5), this variable is $\mathcal{SG}(c_i/2)$, or equivalently (Definition 3.4),

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

Let's move the mean over to the right-hand side:

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

This inequality holds for any choice of $x_{1:i-1}$, so it also holds in expectation over those choices:

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

Now, let's define $a_i = \mathbb{E}_{X_{1:i}} \exp\left(\lambda \mathbb{E}_{X_{i+1:m}} f(X_{1:m})\right)$, so that $a_m = \mathbb{E} \exp(\lambda f(X_{1:m}))$ and $a_0 = \exp(\lambda \mathbb{E} f(X_{1:m}))$. Our overall goal is exactly to bound a_m/a_0, and we have just shown that $a_i \leq a_{i-1} \cdot \exp(\frac{1}{8}\lambda^2 c_i^2)$. Thus, expanding this out recursively,

$$a_m \leq a_0 \exp\left(\sum_{i=1}^{m} \frac{1}{8}\lambda^2 c_i^2\right),$$

which is exactly that $f(X_{1:m})$ is $\mathcal{SG}\left(\frac{1}{2}\sqrt{\sum_{i=1}^m c_i^2}\right)$. The last result follows by the Chernoff bound for subgaussians (Proposition 3.8). □

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 $1 - \delta$,*

$$\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le \mathbb{E}\sup_{h\in\mathcal{H}}[L_{\mathcal{D}}(h) - L_S(h)] + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}. \tag{5.7}$$

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

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

Proof. Let $f(S) = \sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$ be the worst-case generalization gap, and write $S^{(i)} = (z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_m)$.

Now, bounded differences requires a bound on

$$\left|\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h) - L_S(h)) - \sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h) - L_{S^{(i)}}(h))\right|.$$

Notice that for any particular $h \in \mathcal{H}$, we have $L_{\mathcal{D}}(h) - L_{S^{(i)}}(h) \le \sup_{h'} L_{\mathcal{D}}(h') - L_{S^{(i)}}(h')$, and thus $-(L_{\mathcal{D}}(h) - L_{S^{(i)}}(h)) \ge -\sup_{h'}(L_{\mathcal{D}}(h') - L_{S^{(i)}}(h'))$. We can thus bound one direction of the two-sided bound that we want:

$$\begin{aligned}
\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h) - L_S(h)) - \sup_{h'\in\mathcal{H}}(L_{\mathcal{D}}(h') - L_{S^{(i)}}(h')) &\le \sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h) - L_S(h)) - (L_{\mathcal{D}}(h) - L_{S^{(i)}}(h)) \\
&= \sup_{h\in\mathcal{H}} L_{S^{(i)}}(h) - L_S(h) \\
&= \sup_{h\in\mathcal{H}} \frac{1}{m}\left(\ell(h, z_i') - \ell(h, z_i)\right) \le \frac{b-a}{m}.
\end{aligned}$$

The exact same bound holds in the other direction, e.g. by just swapping the names of S and $S^{(i)}$; thus this proves that $\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$ has bounded differences with constants $(b-a)/m$. Equation (5.7) follows by applying McDiarmid.

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

$$\begin{aligned}
L_{\mathcal{D}}(\hat{h}_S) &\le L_S(\hat{h}_S) + \sup_{h\in\mathcal{H}}[L_{\mathcal{D}}(h) - L_S(h)] \\
&\le L_S(\hat{h}_S) + \mathbb{E}\sup_{h\in\mathcal{H}}[L_{\mathcal{D}}(h) - L_S(h)] + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}} && \text{(5.7), w/ prob. } 1-\delta/2 \\
&\le L_S(h^*) + \mathbb{E}\sup_{h\in\mathcal{H}}[L_{\mathcal{D}}(h) - L_S(h)] + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}} && \textit{definition of ERM} \\
&\le L_{\mathcal{D}}(h^*) + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}} + \mathbb{E}\sup_{h\in\mathcal{H}}[L_{\mathcal{D}}(h) - L_S(h)] + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}}, && \textit{Hoeffding, w/ prob. } 1-\delta/2
\end{aligned}$$

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

For bounded-norm bounded-data logistic regression, using (5.5) and (4.3) in (5.8) gives that with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$,

$$L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \frac{BC}{\sqrt{m}}\left[2 + \sqrt{2\log\frac{2}{\delta}}\right] = \mathcal{O}_p\left(\frac{BC}{\sqrt{m}}\right). \tag{5.9}$$

It's also straightforward to generalize this to other Lipschitz losses, and in the homework we'll even use this to show a bound on $\mathcal{H}$ being deep networks.

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.

[Har23] Nick Harvey. *A second course in randomized algorithms*. March 12, 2023.

[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.