CPSC 532D — 5. RADEMACHER COMPLEXITY

Danica J. Sutherland
University of British Columbia, Vancouver
Fall 2023

Last time was our first time showing a uniform convergence bound, i.e. bounding $\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$, for an infinite $\mathcal{H}$. Today we're going to see a technique that takes a little more work to grasp intuitively but will show a slightly better result (no $\log m$), is somewhat more general, and once you understand it can be easier to use.

1 UNIFORM CONVERGENCE IN EXPECTATION

It's going to be easier, here, to 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 return to high-probability bounds on $\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$ later.

2 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 1. *Let f_y be a class of functions indexed by y, and X 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

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

For more, visit https://cs.ubc.ca/~dsuth/532D/23w1/.

This last form is itself a natural thing to think about: how much does $\mathcal{H}$ 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 sub-Gaussian 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

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

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

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

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. $\boldsymbol{\sigma} \mid U, U'$ is still just random signs; given $\boldsymbol{\sigma}$ and U, U′, S and S′ become deterministic. This gives us

This is allowed by Fubini's theorem; for a nonnegative loss, it's fine as long as $L_{\mathcal{D}}(h)$ exists. (For a negative loss, it's enough for $\mathbb{E}_z |\ell(h, z)|$ to exist.)

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

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

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

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

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

$-\boldsymbol{\sigma}$ 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 2. The *Rademacher complexity* of a set $V \subseteq \mathbb{R}^m$ is given by

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

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, 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.

THEOREM 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*

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

3 PROPERTIES OF RADEMACHER COMPLEXITY

First, note that

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

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 the hypercube,

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

This is also the complexity of the function class of all possible $\{-1, 1\}$-valued functions, as long as S has no duplicates: if we tried to do ERM in the set of "all possible classifiers," we'd only get that the expected zero-one loss is less than 2.

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

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

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

For $V + W = \{v + w : v \in V, w \in W\}$ 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.

3.1 *Talagrand's contraction lemma*

This lemma is also very helpful for bounding $\mathrm{Rad}(\mathcal{H})$ *for* $\mathcal{H}$ *that are defined compositionally, like deep networks.*

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, then bound the complexity of $\mathcal{H}$. We can do that with the following lemma.

The major way to do that is with the following results, for Lipschitz losses. (We showed that logistic loss, used in logistic regression, is 1-Lipschitz last lecture.) To remind you of the definition:

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

definition 4. A function $f : \mathcal{X} \to \mathcal{Y}$ is ρ-Lipschitz with respect to $\mathrm{dist}_{\mathcal{X}}$ and $\mathrm{dist}_{\mathcal{Y}}$ if for all $x, x' \in \mathcal{X}$, $\mathrm{dist}_{\mathcal{Y}}(f(x), f(x')) \le \rho\,\mathrm{dist}_{\mathcal{X}}(x, x')$. The smallest ρ for which this inequality holds is *the Lipschitz constant*, denoted $\|f\|_{\mathrm{Lip}}$.

The same idea establishes that for differentiable $\mathbb{R}^d \to \mathbb{R}$ *functions,* $\|f\|_{\mathrm{Lip}} = \sup_{x\in\mathcal{X}} \|\nabla f(x)\|$.

If $\mathcal{X}$ and/or $\mathcal{Y}$ are Euclidean spaces, dist is Euclidean distance unless otherwise specified. We showed last time that for differentiable $\mathbb{R} \to \mathbb{R}$ functions, $\|f\|_{\mathrm{Lip}} = \sup_{x\in\mathcal{X}} |f'(x)|$. The canonical example of a non-differentiable Lipschitz function is the absolute value.

lemma 5 (Talagrand). *Let* $\phi : \mathbb{R}^m \to \mathbb{R}^m$ *be given by* $\phi(t) = (\varphi_1(t_1), \ldots, \varphi_m(t_m))$*, where each* ϕ_i *is* ρ*-Lipschitz. Then*

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

Our proof will be based on the following special case:

lemma 6. *If* $\varphi : \mathbb{R} \to \mathbb{R}$ *is* 1*-Lipschitz,* $\mathrm{Rad}(\{(\varphi(v_1), v_2, \ldots, v_m) : v \in V\}) \le \mathrm{Rad}(V)$.

Proof of Lemma 5, assuming Lemma 6. First notice that "rotating" the vectors in V doesn't change its complexity, since σ has iid entries:

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

Now, define the function $\phi'(t) = (\frac{1}{\rho}\varphi_1(t_1), \ldots, \frac{1}{\rho}\varphi_m(t_m))$; notice that each of its components is 1-Lipschitz. So, start by applying Lemma 6 to V with $\frac{1}{\rho}\varphi_1$, then rotating, to obtain

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

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

$$\mathrm{Rad}(\phi' \circ V) \le \mathrm{Rad}(V).$$

Finally, scale by ρ, which by (1) means

$$\operatorname{Rad}(\phi \circ V) = \rho \operatorname{Rad}(\phi' \circ V) \leq \rho \operatorname{Rad}(V). \qquad \square$$

Proof of Lemma 6. Let $\phi(v) = (\varphi(v_1), v_2, \ldots, v_m)$ so that $\phi \circ V = \{(\varphi(v_1), v_2, \ldots, v_m) : v \in V\}$. We have

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

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

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

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 $v_1 \leq v_1'$, there's an exactly equivalent one with $v_1 \geq v_1'$. Thus

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

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

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

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

$$\operatorname{Rad}((\ell \circ \mathcal{H})_S) \leq \rho \operatorname{Rad}(\mathcal{H}|_{S_x}). \tag{2}$$

Note that ρ here might depend on the particular S_y!

3.2 *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

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

using (2) with our previous result that logistic loss is 1-Lipschitz. Now let's bound that latter term:

$$\begin{aligned}
m\,\mathrm{Rad}(\mathcal{H}|_{S_x}) &= \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\|\le B} \sum_i \sigma_i \langle w, x_i\rangle \\
&= \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\|\le B} \langle w, \sum_i \sigma_i x_i\rangle \\
&\le \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\|\le B} \|w\| \left\|\sum_i \sigma_i x_i\right\| && \textit{using Cauchy-Shwartz} \\
&= B \mathop{\mathbb{E}}_{\sigma} \left\|\sum_i \sigma_i x_i\right\| \\
&\le B \sqrt{\mathop{\mathbb{E}}_{\sigma} \left\|\sum_i \sigma_i x_i\right\|^2} && \textit{using } (\mathbb{E}\,T)^2 \le \mathbb{E}\,T^2 \\
&= B \sqrt{\mathop{\mathbb{E}}_{\sigma} \sum_{ij} \sigma_i \sigma_j \langle x_i, x_j\rangle} \\
&= B \sqrt{\sum_i \underbrace{\mathbb{E}[\sigma_i^2]}_{1} \|x_i\|^2 + \sum_{i\ne j} \underbrace{\mathop{\mathbb{E}}_{\sigma}[\sigma_i \sigma_j]}_{0} \langle x_i, x_j\rangle}.
\end{aligned}$$

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

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

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

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

we'll see soon that, in this case and if $BC \ge 1$, we can convert this into a high-

probability bound of the form

$$\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \frac{2BC}{\sqrt{m}}\left[1 + \sqrt{2\log\frac{2}{\delta}}\right]\right) \ge 1-\delta. \tag{3}$$

Compare this to the covering numbers-based bound we showed before:

$$\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \frac{2BC}{\sqrt{m}}\left[1 + 2\sqrt{\log\frac{2}{\delta}} + \sqrt{\frac{d}{2}\log(9m)}\right]\right) \ge 1-\delta.$$

The other way to handle the dependence on the particular S_x is to write

$$\mathop{\mathbb{E}}_{S}\mathrm{Rad}(\mathcal{H}_B|_{S_x}) \le \frac{B}{\sqrt{m}}\mathop{\mathbb{E}}_{S}\sqrt{\frac{1}{m}\sum_i \|x_i\|^2} \le \frac{B}{\sqrt{m}}\sqrt{\mathop{\mathbb{E}}_{x}\|x\|^2}. \tag{4}$$

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

This allows for broader data distributions, as long as you can bound $\mathbb{E}\|x\|^2$: e.g. you can easily handle Gaussians.

In either case, this means we've shown an average-case excess error bound for logistic regression (and mean-absolute-error linear regression, and…) with a rate of $\mathcal{O}(1/\sqrt{m})$.

4 CONCENTRATION

Now let's prove that high-probability bound. We'll need a new tool: *McDiarmid's inequality*. This is a very important concentration inequality, which holds when we have *bounded differences*.

THEOREM 7 ([McD89]). *Let $X_1,\dots,X_m$ be independent, and let $f(X_1,\dots,X_m)$ be a real-valued function satisfying*

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

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 $(X_i,\dots,X_j)$.

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].

Fix some $k \in [m]$, and 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.

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

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

by assumption. This is true for *any* values for $x_{k+1:m}$, so it's also true if we average

over them (and change $-\inf x$ to $+\sup -x$):

$$\mathop{\mathbb{E}}_{X_{k+1:m}} \sup_{x_k} f(x_{1:k-1}, x_k, X_{k+1:m}) + \sup_{x_k}(-f(x_{1:k-1}, x_k, X_{k+1:m})) \le c_k.$$

Now, using Lemma 1, this implies that

$$\sup_{x_k} \mathop{\mathbb{E}}_{X_{k+1:m}} f(x_{1:k-1}, x_k, X_{k+1:m}) + \sup_{x_k} \mathop{\mathbb{E}}_{X_{k+1:m}} (-f(x_{1:k-1}, x_k, X_{k+1:m})) \le c_k;$$

changing the sup back to an inf shows the boundedness we wanted.

Thus, by Hoeffding's lemma, this variable is $\mathcal{SG}(c_k/2)$. That is, multiplying our definition of sub-Gaussianity by $e^{\lambda\mu}$ 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:

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

Most of these terms cancel: if we combined the two sums, the result would be telescoping. So, this simplifies to

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

or, 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 sub-Gaussians 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 k, 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 8. *Suppose that $\ell(h,z) \in [a,b]$ for all h, z. Then, with probability at least $1-\delta$,*

$$\sup_{h\in\mathcal{H}} \mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}}(h) \le \mathbb{E}\sup_{h\in\mathcal{H}}[\mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}}(h)] + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}.$$

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

$$\mathrm{L}_{\mathcal{D}}(\hat{h}_{\mathrm{S}}) - \inf_{h\in\mathcal{H}} \mathrm{L}_{\mathcal{D}}(h) \le \mathbb{E}\sup_{h\in\mathcal{H}}[\mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}}(h)] + (b-a)\sqrt{\frac{2}{m}\log\frac{2}{\delta}}.$$

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

$$\mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}}(h) = \mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}^{(i)}}(h) + \mathrm{L}_{\mathrm{S}^{(i)}}(h) - \mathrm{L}_{\mathrm{S}}(h);$$

take $\sup_h$ of both sides, use $\sup_x f(x) + g(x) \le \sup_x f(x) + \sup_x g(x)$, rearrange, then use $|\sup f(x)| \le \sup|f(x)|$ to see that

$$\left|\sup_{h\in\mathcal{H}}[\mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}}(h)] - \sup_{h\in\mathcal{H}}[\mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}^{(i)}}(h)]\right|$$
$$\le \sup_{h\in\mathcal{H}} |\mathrm{L}_{\mathrm{S}^{(i)}}(h) - \mathrm{L}_{\mathrm{S}}(h)| = \sup_{h\in\mathcal{H}} \frac{1}{m}\left|\ell(h,z') - \ell(h,z_i)\right| \le \frac{b-a}{m},$$

because the loss is bounded. The first equation follows by applying McDiarmid to the function $f(\mathrm{S}) = \sup_{h\in\mathcal{H}} \mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}}(h)$.

The second follows as usual for our ERM bounds: we know that

$$\mathrm{L}_{\mathcal{D}}(\hat{h}_{\mathrm{S}}) \le \mathrm{L}_{\mathrm{S}}(\hat{h}_{\mathrm{S}}) + \mathbb{E}\sup_{h\in\mathcal{H}}[\mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}}(h)] + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}}$$
$$\le \mathrm{L}_{\mathrm{S}}(h^*) + \mathbb{E}\sup_{h\in\mathcal{H}}[\mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}}(h)] + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}}$$
$$\le \mathrm{L}_{\mathcal{D}}(h^*) + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}} + \mathbb{E}\sup_{h\in\mathcal{H}}[\mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}}(h)] + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}},$$

where the first and last inequalities each add a $\delta/2$ probability of error. □

For bounded-norm bounded-data logistic regression with $\mathrm{BC} \ge 1$, this gives (3).

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.