CPSC 532D — 4. PAC LEARNING; INFINITE $\mathcal{H}$

Danica J. Sutherland
University of British Columbia, Vancouver
Fall 2024

Recall that we previously showed Proposition 2.2:

Proposition 2.2. *Suppose $\ell(z, h)$ is almost surely bounded in $[a, b]$, $\mathcal{H}$ is finite, and $\hat{h}_S$ is any ERM in $\mathcal{H}$. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$ it holds that*

$$L_{\mathcal{D}}(\hat{h}_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \le (b - a)\sqrt{\frac{2}{m} \log \frac{|\mathcal{H}| + 1}{\delta}}.$$

Another way to state this result is that with m samples, we can achieve estimation error at most ε with probability at least $1 - (|\mathcal{H}| + 1) \exp\left(-\frac{m\varepsilon^2}{2(b-a)^2}\right)$.

Or, alternately, we can say that we can achieve estimation error at most ε with probability at least $1 - \delta$ if we have at least $\frac{2(b-a)^2}{\varepsilon^2} \log \frac{|\mathcal{H}|+1}{\delta}$ samples. This last way establishes the *sample complexity* of learning to a given estimation error ε with a given confidence $1 - \delta$.

4.1 PAC LEARNING

This last statement corresponds to one of the standard notions of learnability. Here, we're going to use a general idea of a learning algorithm as some function that takes a sample $S \in \mathcal{Z}^*$ (the set of sequences of any length from $\mathcal{Z}$) and returns a hypothesis in $\mathcal{H}$.

Definition 4.1. An algorithm $\mathcal{A} : \mathcal{Z}^* \to \mathcal{H}$ *agnostically PAC learns* $\mathcal{H}$ with a loss ℓ if there exists a function $m : (0, 1)^2 \to \mathbb{N}$ such that, for every $\varepsilon, \delta \in (0, 1)$, for every distribution $\mathcal{D}$ over $\mathcal{Z}$, for any $m \ge m(\varepsilon, \delta)$, we have that

$$\Pr_{S \sim \mathcal{D}^m, \mathcal{A}}\left(L_{\mathcal{D}}(\mathcal{A}(S)) > \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\right) < \delta,$$

where the randomness is both over the choice of S and any internal randomness in the algorithm $\mathcal{A}$. That is, $\mathcal{A}$ can *probably* get an *approximately correct* answer, where "correct" means the best possible loss in $\mathcal{H}$.

If $\mathcal{A}$ runs in time polynomial in $1/\varepsilon$, $1/\delta$, m, and some notion of the size of h^*, then we say that $\mathcal{A}$ *efficiently agnostically PAC learns* $\mathcal{H}$.

Definition 4.2. A hypothesis class $\mathcal{H}$ is *agnostically PAC learnable* if there exists an algorithm $\mathcal{A}$ which agnostically PAC learns $\mathcal{H}$.

So, ERM agnostically PAC-learns finite hypothesis classes, with the sample complexity $m(\varepsilon, \delta) = \frac{2(b-a)^2}{\varepsilon^2} \log \frac{|\mathcal{H}|+1}{\delta}$. Notice that in the definition of agnostic PAC learning, there's no limitation on the distribution – there needs to be an $m(\varepsilon, \delta)$ that works for

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

any $\mathcal{D}$. Proposition 2.2 satisfies this, but in general, it's an extremely worst-case kind of notion.

Often it's nicer to think about cases where we can make some assumptions on $\mathcal{D}$. For example, maybe the number of samples you need depends on "how hard" the particular problem is. We'll talk about this more a little later in the course. For now, it's worth mentioning one common special case:

A1 Q4 was partly about this setting.

Definition 4.3. Consider a nonnegative loss $\ell(h, z) \geq 0$. A distribution $\mathcal{D}$ is called *realizable* by $\mathcal{H}$ if there exists an $h^* \in \mathcal{H}$ such that $L_{\mathcal{D}}(h^*) = 0$.

This version is the "privileged" version that doesn't need a modifier because it's was introduced first [Val84].

Definition 4.4. An algorithm $\mathcal{A} : \mathcal{Z}^* \to \mathcal{H}$ *PAC learns* $\mathcal{H}$ with a loss ℓ if there exists a function $m : (0, 1)^2 \to \mathbb{N}$ such that, for every $\varepsilon, \delta \in (0, 1)$, for every *realizable* distribution $\mathcal{D}$ over $\mathcal{Z}$, for any $m \geq m(\varepsilon, \delta)$, we have that

$$\Pr_{S \sim \mathcal{D}^m, \mathcal{A}}(L_{\mathcal{D}}(\mathcal{A}(S)) > \varepsilon) < \delta,$$

where the randomness is both over the choice of S and any internal randomness in the algorithm $\mathcal{A}$. That is, $\mathcal{A}$ can *probably* get an *approximately correct* answer, where "correct" means zero loss.

If $\mathcal{A}$ runs in time polynomial in $1/\varepsilon$, $1/\delta$, m, and some notion of the size of h^*, then we say that A *efficiently (realizably) PAC learns* $\mathcal{H}$.

Definition 4.5. A hypothesis class $\mathcal{H}$ is *PAC learnable* if there exists an algorithm $\mathcal{A}$ which PAC learns $\mathcal{H}$.

Sometimes people say "realizable PAC learnable" or similar, to emphasize the difference versus agnostic PAC. The name "agnostic" is because the definition doesn't care whether there's a perfect h^* or not. (Notice that if $\mathcal{A}$ agnostically PAC learns $\mathcal{H}$, then it also PAC learns $\mathcal{H}$.)

The emphasis here on "how many samples for a given error" is also kind of a TCS-style framing, whereas statisticians more often ask "how much error for a given number of samples"; I tend to prefer the latter, but it's all equivalent.

If you read [SSBD14] or other work by computational learning theorists, there tends to be a lot of focus on just being learnable versus not being learnable. That problem has been solved, though, as we'll see not too much later in class; recent work focuses much more on rates than on just learnability or not, and tends to be willing to make *some* assumptions on $\mathcal{D}$ rather than either being totally general or assuming only realizability.

We've shown that anything finite is agnostically PAC learnable. That's only an upper bound, though; it *doesn't* mean that infinite things aren't learnable. Which is good, because that's what we usually want to learn!

Lemma 6.1 of [SSBD14] gives a really simple example of realizably PAC learning an infinite class, if you're curious to see that style of proof. I tried to do an agnostic version of that, but it was more complicated than I hoped, so let's do something more interesting instead.

4.2 Covering Number Bounds

This is more convenient than $\mathcal{Y} = \{0, 1\}$ here…

You usually want an intercept term, $w \cdot x + w_0$, but you can achieve that by padding x with an always-one dimension.

In *logistic regression*, our data is in a subset of $\mathbb{R}^d$, our labels are in $\mathcal{Y} = \{-1, 1\}$ and we try to predict with a confidence score in $\widehat{\mathcal{Y}} = \mathbb{R}$. Our predictors are linear functions of the form $h_w(x) = w \cdot x$, and the logistic loss is given by

$$\ell_{log}(h, (x, y)) = l_y^{log}(h(x)) = \log(1 + \exp(-h(x)y)). \tag{4.1}$$

We'll use the hypothesis class $\mathcal{H} = \{h_w = x \mapsto w \cdot x : w \in \mathbb{R}^d, \|w\| \le B\}$ for some constant B; this avoids overfitting by using really-really complex w, and is basically equivalent to doing L_2-regularized logistic regression (we'll talk about this more later). This $\mathcal{H}$ is still infinite, but it has finite volume.

Now, our analysis is going to be based on the idea that if w and v are similar predictors, i.e. $h_w(x) \approx h_v(x)$ for all x, then they'll behave similarly: $L_\mathcal{D}(h_w) \approx L_\mathcal{D}(h_v)$ and $L_S(h_w) \approx L_S(h_v)$. Thus we don't have to do a totally separate concentration bound on their empirical risks; we can exploit that they're similar.

The fundamental idea is going to be one of a "set cover," or an "ε-net." To handle an infinite $\mathcal{H}$ that's nonetheless bounded, we're going to choose some *finite* set $\mathcal{H}_0$ such that everything in $\mathcal{H}$ is close to something in $\mathcal{H}_0$, use Proposition 2.2 to say that $L_\mathcal{D}(h) - L_S(h)$ isn't too big for anything in $\mathcal{H}_0$, and then argue that since $L_\mathcal{D}(h) - L_S(h)$ is smooth, this means it can't be too big for anything in $\mathcal{H}$ at all.

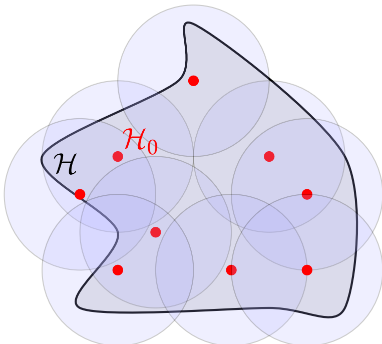

Figure 4.1: A (non-minimal) set cover.

4.2.1 *Smoothness: Lipschitz functions*

To formalize the idea that similar weight vectors give similar loss, we'll want a bound like

$$|L_\mathcal{D}(h) - L_\mathcal{D}(g)| \le M\, \rho_\mathcal{H}(h, g),$$

for some notion of a distance metric on $\mathcal{H}$. This is called a Lipschitz property.

Definition 4.6. A function $f : \mathcal{X} \to \mathcal{Y}$ is M-*Lipschitz* with respect to $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$ if for all $x, x' \in \mathcal{X}$, $\rho_\mathcal{Y}(f(x), f(x')) \le M\, \rho_\mathcal{X}(x, x')$. The smallest M for which this inequality holds is *the Lipschitz constant*, denoted $\|f\|_{\text{Lip}}$.

If $\mathcal{X}$ and/or $\mathcal{Y}$ are subsets of $\mathbb{R}^d$, ρ is Euclidean distance unless otherwise specified.

So, for example, $x \mapsto |x|$ is a 1-Lipschitz function, since $\big||x| - |y|\big| \le \big|x - y\big|$.

The notation $\|f\|_{\text{Lip}}$ is justified by the following result. If you're not sure about function spaces / norms / etc, don't worry about it (we'll come back to this later in the course); the takeaway is the two properties shown in the proof.

Lemma 4.7. *Consider a vector space of functions $\mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is a normed space, such that $f + g$ is the function $x \mapsto f(x) + g(x)$ and af is the function $x \mapsto af(x)$. $\|\cdot\|_{\text{Lip}}$ is a seminorm on this space with respect to $\|\cdot - \cdot\|_\mathcal{Y}$.*

Proof. There are two properties to show. First, subadditivity (which implies the triangle inequality):

$$\begin{aligned}\|f+g\|_{\mathrm{Lip}} &= \sup_{x\neq x'} \frac{\|f(x)+g(x)-f(x')-g(x')\|}{\rho_{\mathcal{X}}(x,x')} \\ &\leq \sup_{x\neq x'} \frac{\|f(x)-f(x')\|}{\rho_{\mathcal{X}}(x,x')} + \frac{\|g(x)-g(x')\|}{\rho_{\mathcal{X}}(x,x')} \leq \|f\|_{\mathrm{Lip}} + \|g\|_{\mathrm{Lip}}.\end{aligned}$$

Second, absolute homogeneity:

$$\|af\|_{\mathrm{Lip}} = \sup_{x\neq x'} \frac{\|af(x)-af(x')\|}{\rho_{\mathcal{X}}(x,x')} = \sup_{x\neq x'} \frac{|a|\,\|f(x)-f(x')\|}{\rho_{\mathcal{X}}(x,x')} = |a|\,\|f\|_{\mathrm{Lip}}. \qquad \square$$

It isn't a proper norm because $\|x \mapsto a\|_{\mathrm{Lip}} = 0$ for all constant functions.

So, what is $\|\mathrm{L}_{\mathcal{D}}\|_{\mathrm{Lip}}$? When $z=(x,y)$ and $\ell(h,(x,y)) = l_y(h(x))$, we have

$$\begin{aligned}|\mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathcal{D}}(g)| &= \left| \mathop{\mathbb{E}}_{z\sim\mathcal{D}} \ell(h,z) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}} \ell(g,z) \right| \\ &\leq \mathop{\mathbb{E}}_{z\sim\mathcal{D}} |\ell(h,z)-\ell(g,z)| \\ &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left| l_y(h(x)) - l_y(g(x)) \right| \\ &\leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \|l_y\|_{\mathrm{Lip}}\, \rho_{\hat{\mathcal{Y}}}(h(x),g(x)). \end{aligned} \tag{4.2}$$

So, in particular settings we want to find $\left\|l_y\right\|_{\mathrm{Lip}}$ and bound $\rho_{\hat{\mathcal{Y}}}(h(x),g(x))$ in terms of some notion of similarity between h and g.

For the first problem, since for logistic regression $l_y^{log}: \mathbb{R} \to \mathbb{R}$, this result will help:

Lemma 4.8. *Let $\mathcal{X} \subseteq \mathbb{R}$ be a connected, closed set. If a function $f : \mathcal{X} \to \mathbb{R}$ is continuous and differentiable everywhere on the interior of $\mathcal{X}$, $\|f\|_{\mathrm{Lip}} = \sup_{x\in\mathcal{X}} |f'(x)|$.*

Proof. We apply the fundamental theorem of calculus:

$$\left|f(x')-f(x)\right| = \left| \int_x^{x'} f'(x)\mathrm{d}x \right| \leq \int_x^{x'} \left|f'(x)\right| \mathrm{d}x \leq \int_x^{x'} \|f\|_{\mathrm{Lip}}\, \mathrm{d}x = \|f\|_{\mathrm{Lip}} \left|x'-x\right|. \qquad \square$$

We won't need this today, but it's worth noting that if $\mathcal{X} \subseteq \mathbb{R}^d$, the same proof idea gives us that $\|f\|_{\mathrm{Lip}} = \sup_{x\in\mathcal{X}} \|\nabla f(x)\|$.

Lemma 4.9. *For any $y \in \{-1,1\}$, $\left\|l_y^{log}\right\|_{\mathrm{Lip}} \leq 1$.*

Proof. l_y^{log} is differentiable everywhere on $\mathbb{R}$, and so using Lemma 4.8,

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}\hat{y}} l_y^{\log}(\hat{y})\right| &= \left|\frac{\mathrm{d}}{\mathrm{d}\hat{y}} \log(1+\exp(-y\hat{y}))\right| = \left| \frac{1}{1+\exp(-y\hat{y})} \exp(-y\hat{y})(-y) \right| \\ &= \left| \frac{\exp(-y\hat{y})}{1+\exp(-y\hat{y})} \times \frac{\exp(y\hat{y})}{\exp(y\hat{y})} \right| \left|-y\right| = \left| \frac{1}{1+\exp(y\hat{y})} \right| \leq 1. \qquad \square\end{aligned}$$

Plugging into (4.2), we get

$$|L_{\mathcal{D}}(h_w) - L_{\mathcal{D}}(h_v)| \le \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left\|l_y\right\|_{\text{Lip}} |h_w(x) - h_v(x)|.$$

That is, if the predictions are similar, the losses are too. We can further say that if w and v are close, then their predictions are similar:

$$|h_w(x) - h_v(x)| = |w \cdot x - v \cdot x| = |(w - v) \cdot x| \le \|w - v\| \, \|x\|$$

by Cauchy-Schwarz. Thus

$$|L_{\mathcal{D}}(h_w) - L_{\mathcal{D}}(h_v)| \le \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \|x\| \, \|l_y\|_{\text{Lip}} \right) \|w - v\|,$$

giving that $L_{\mathcal{D}}$ is $\left(\mathbb{E}_{(x,y)\sim\mathcal{D}} \|x\| \, \|l_y\|_{\text{Lip}}\right)$-Lipschitz with respect to $\rho_{\mathcal{H}}(h_w, h_v) = \|w - v\|$, and similarly L_S is $\left(\frac{1}{m} \sum_{i=1}^{m} \|x_i\| \, \|l_{y_i}\|_{\text{Lip}}\right)$-Lipschitz. (We could repeat the argument with empirical averages instead of $\mathbb{E}$, but a slicker way is to note that L_S is exactly $L_{\hat{\mathcal{D}}_S}$ for the *empirical distribution* $\hat{\mathcal{D}}_S$, the discrete distribution that puts $1/m$ probability at each member of S.) Thus we know that

$$\|L_{\mathcal{D}} - L_S\|_{\text{Lip}} \le \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \|x\| \, \|l_y\|_{\text{Lip}} + \frac{1}{m} \sum_{i=1}^{m} \|x_i\| \, \|l_{y_i}\|_{\text{Lip}}. \tag{4.3}$$

If we assume for simplicity that the distribution is bounded, $\Pr_{(x,y)\sim\mathcal{D}}(\|x\| \le C) = 1$, and that $\|l_y\|_{\text{Lip}} \le M$ for each y (as with logistic loss, where $M = 1$), then $L_{\mathcal{D}} - L_S$ is guaranteed to be $(2CM)$-Lipschitz.

4.2.2 *Putting it together with a set covering*

Now the question is: how big does $\mathcal{H}_0$ have to be? We'll use the following concept:

Definition 4.10. An η-cover of a set U is a set $T \subseteq U$ such that, for all $u \in U$, there is a $t \in T$ with $\rho(t, u) \le \eta$. The *covering number* $N(U, \eta)$ is the size of the smallest η-cover for U.

We want to cover $\mathcal{H}_B = \{h_w = (x \mapsto w \cdot x) : \|w\| \le B\}$ with the metric $\rho(h_w, h_v) = \|w - v\|$. We can immediately construct this kind of cover if we have a cover for the Euclidean ball of radius B. Section 4.2.3 bounds how big this cover needs to be:

Lemma 4.11. *Let $\eta \in (0, B]$ and $p \in [1, \infty]$. The covering number of the radius-B p-norm ball in $\mathbb{R}^d$, $U = \{x \in \mathbb{R}^d : \|x\|_p \le B\}$, satisfies*

$$\left(\frac{B}{\eta}\right)^d \le N(U, \eta) \le \left(\frac{2B}{\eta} + 1\right)^d \le \left(\frac{3B}{\eta}\right)^d.$$

(When $\eta \ge B$, trivially $N(U, \eta) = 1$.)

We now have all the tools we need for the following result about linear models with bounded Lipschitz losses.

Proposition 4.12. *Let $h_w(x) = w \cdot x$ and $\mathcal{H} = \{h_w : \|w\| \le B\}$ for some $B > 0$. Consider a loss $\ell(h, (x, y)) = l_y(h(x))$ for functions $l_y : \mathbb{R} \to \mathbb{R}$ which each have Lipschitz constant at most M and are bounded in $[a, b]$. Assume that $\|x\| \le C$ almost surely under $\mathcal{D}$. Then,* *with probability at least* $1 - \delta$,

$$\sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \leq \frac{1}{\sqrt{2m}} \left[BCM + (b-a) \sqrt{\log \frac{1}{\delta} + \frac{d}{2} \log(72m)} \right].$$

Proof. We'll first choose a η-cover $\mathcal{H}_0 = \{w_1, \dots, w_{N_\eta}\} \subset \{w \in \mathbb{R}^d : \|w\| \leq B\}$, where η is a parameter to be set later. Then, for any $h \in \mathcal{H}$, let $\mathrm{nn}_{\mathcal{H}_0}(h) \in \arg\min_{h' \in \mathcal{H}_0} \rho(h, h')$, using $\rho(h_w, h_v) = \|w - v\|$. Define the function $\Delta(h) := L_{\mathcal{D}}(h) - L_S(h)$ for brevity. Then

$$\begin{aligned}
\sup_{h \in \mathcal{H}} \Delta(h) &= \sup_{h \in \mathcal{H}} \Delta(h) - \Delta(\mathrm{nn}(h)) + \Delta(\mathrm{nn}(h)) \\
&\leq \sup_{h \in \mathcal{H}} [\Delta(h) - \Delta(\mathrm{nn}(h))] + \sup_{h' \in \mathcal{H}_0} \Delta(h') \\
&\leq 2CM\eta + \sup_{h' \in \mathcal{H}_0} \Delta(h'),
\end{aligned}$$

where the first term is because of (4.3) and $\mathcal{H}_0$ being an η-cover.

The other term is uniform convergence over a finite hypothesis class $\mathcal{H}_0$, as in Proposition 2.2. We can apply Hoeffding to each element of $\mathcal{H}_0$, giving it a failure probability of δ/N_η, and obtain that with probability at least $1 - \delta$,

$$\begin{aligned}
\sup_{h \in \mathcal{H}} \Delta(h) &\leq 2CM\eta + (b-a)\sqrt{\frac{1}{2m} \log \frac{N_\eta}{\delta}} \\
&\leq 2CM\eta + (b-a)\sqrt{\frac{1}{2m} \left[\log \frac{1}{\delta} + d \log \frac{3B}{\eta} \right]}.
\end{aligned}$$

Now, we could try to exactly optimize the value of η, but I think we won't be able to do that analytically. Instead, let's notice that if η is $o(1/\sqrt{m})$, the first term being smaller doesn't really help in rate since the other term is $1/\sqrt{m}$ anyway – but choosing a smaller η makes the $\log \frac{1}{\eta}$ worse. Also, the dependence on η there is only in a log term, so it's probably okay-ish to choose $\eta = \alpha/\sqrt{m}$ for some $\alpha > 0$, giving us

$$\sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] \leq \frac{1}{\sqrt{m}} \left[2CM\alpha + \frac{b-a}{\sqrt{2}} \sqrt{\log \frac{1}{\delta} + d \log \frac{3B\sqrt{m}}{\alpha}} \right].$$

Picking $\alpha = B/(2\sqrt{2})$ and using $\log A = \frac{1}{2} \log(A^2)$ gives the desired result. □

For our motivating problem of logistic regression, $M = 1$, but there's one catch: we can use $a = 0$ but there isn't an "inherent" upper bound for b. Given that we know

$\|x\| \le C$ and $\|w\| \le B$, though, we have that $|h(x)| = |w \cdot x| \le BC$. Thus

$$
\begin{aligned}
\ell(h,(x,y)) &= \log(1+\exp(-yh(x)) \le \log(1+\exp(BC)) =: b \\
\ell(h,(x,y)) &= \log(1+\exp(-yh(x)) \ge \log(1+\exp(-BC)) =: a \\
b - a &= \log(1+\exp(BC)) - \log(1+\exp(-BC)) \\
&= \log\left(\frac{1+\exp(BC)}{1+\exp(-BC)} \times \frac{\exp(BC)}{\exp(BC)}\right) \\
&= \log\left(\frac{1+\exp(BC)}{\exp(BC)+1} \times \exp(BC)\right) = \log\exp(BC) = BC. \qquad (4.4)
\end{aligned}
$$

Plugging into Proposition 4.12 gives us that with probability at least $1-\delta$, logistic regression with bounded-norm weights on bounded-norm data satisfies

$$
\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le \frac{BC}{\sqrt{2m}}\left[1 + \sqrt{\log\frac{1}{\delta} + \frac{d}{2}\log(72m)}\right] = \mathcal{O}_p\left(BC\sqrt{\frac{d\log m}{m}}\right). \qquad (4.5)
$$

Treating everything but m as a constant, the rate is $\mathcal{O}_p\left(\sqrt{\frac{\log m}{m}}\right)$. That $\sqrt{\log m}$ factor is actually unnecessary, but getting rid of it with covering number-type arguments requires some more advanced machinery. Instead, soon we'll see a simpler way to show a $\mathcal{O}_p(1/\sqrt{m})$ rate – in fact, a $\mathcal{O}_p(BC/\sqrt{m})$ rate, also dramatically improving the dependence on d – that will also be very generally applicable.

This machinery is called "chaining"; we probably won't cover it in class, but Wainwright [Wai19, Section 5.3.3] has a reasonable overview.

ERM bound We only wrote this proof here for $\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$, but since the loss is bounded, this implies exactly as in (1.5) an upper bound on the generalization error of any ERM $\hat{h}_S$. Using the general result from Proposition 4.12 with probability $\delta/2$, and plain Hoeffding with probability $\delta/2$ on the $L_S(h^*) - L_{\mathcal{D}}(h^*)$ term, gives us

$$
L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \frac{1}{\sqrt{2m}}\left[BCM + (b-a)\sqrt{\log\frac{2}{\delta} + \frac{d}{2}\log(72m)}\right] + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}},
$$

and using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ we can simplify to

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

Specializing to logistic regression, we can plug in $M = 1$, $b - a = BC$ so that

$$
L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \frac{BC}{\sqrt{m}}\left[\frac{1}{\sqrt{2}} + \frac{1}{2}\sqrt{d\log(72m)} + \sqrt{2\log\frac{2}{\delta}}\right] = \mathcal{O}_p\left(BC\sqrt{\frac{d\log m}{m}}\right). \qquad (4.6)
$$

A question for yourself here: does this imply that ERM agnostically PAC-learns logistic regression?

More general versions We used the following properties about the problem:

- A bounded loss, to apply Hoeffding. This could be weakened in various ways, e.g. another kind of subgaussianity, or other ways to show concentration for a finite number of points.
- A Lipschitz loss. Some form of this is definitely necessary. You could poten-

tially use a locally Lipschitz loss (where the constant varies through space), but then you have to be more careful in bounding (4.3) or similar.

- A parameterization for $\mathcal{H}$ with a covering number bound. We framed this as covering the parameter set for linear models, but you could use more general notions of covering for $\mathcal{H}$, as long as they're compatible with the metric you use for Lipschitzness in the previous part. This generality is often useful, e.g. for nonparametric $\mathcal{H}$.

4.2.3 *Aside: Bounds on covering numbers*

We'll now prove our upper bound on covering numbers. Recall their definition:

Definition 4.10. An η-cover of a set U is a set $T \subseteq U$ such that, for all $u \in U$, there is a $t \in T$ with $\rho(t, u) \leq \eta$. The *covering number* $N(U, \eta)$ is the size of the smallest η-cover for U.

We'll also use *packing numbers*: how many balls can we squeeze into a set T?

Definition 4.13. An η-*packing* of a set U is a set $T \subseteq U$ such that, for all $t, t' \in T$ with $t \neq t'$, we have $\rho(t, t') > \eta$. The *packing number* $M(U, \eta)$ is the maximal size of any η-packing.

Proposition 4.14. *A maximally-sized η-packing T of a set U is also a η-cover of U.*

Proof. Suppose there were some point $u \in U$ such that $\rho(u, t) > \eta$ for all $t \in T$. Then we could add u to the η-packing, producing a packing of size one larger; this contradicts that T was maximal. □

We're now ready to prove the result:

Lemma 4.11. *Let $\eta \in (0, B]$ and $p \in [1, \infty]$. The covering number of the radius-B p-norm ball in $\mathbb{R}^d$, $U = \{x \in \mathbb{R}^d : \|x\|_p \leq B\}$, satisfies*

$$\left(\frac{B}{\eta}\right)^d \leq N(U, \eta) \leq \left(\frac{2B}{\eta} + 1\right)^d \leq \left(\frac{3B}{\eta}\right)^d.$$

(When $\eta \geq B$, trivially $N(U, \eta) = 1$.)

Proof. By Proposition 4.14, we have that $N(U, \eta) \leq M(U, \eta)$; we'll first prove the upper bound on the packing number M. Let T be a maximal η-packing of the B-ball $U = \{w \in \mathbb{R}^d : \|w\| \leq B\}$. Thus the open $\eta/2$-balls centered at each $t \in T$, $\{w \in \mathbb{R}^d : \|w - t\|_p < \eta/2\}$, are disjoint: if they weren't, you could get from one t to another in distance less than η, contradicting that T is an η-packing. These balls are also all contained within the ball of radius $(B + \eta/2)$, since each $\|t\|_p \leq B$. Thus

$$\sum_{t \in T} \operatorname{vol}\left(\{w \in \mathbb{R}^d : \|w - t\|_p < \eta/2\}\right) \leq \operatorname{vol}\left(\{w \in \mathbb{R}^d : \|w\|_p < B + \eta/2\}\right).$$

But we know that the volume of a p-norm ball of radius R in d dimensions is $R^d V_1$, where $V_1 = \mathrm{vol}(\{w \in \mathbb{R}^d : \|w\|_p < 1\})$. Thus

$$\sum_{t \in T} \left(\frac{\eta}{2}\right)^d V_1 = M(U,\eta)\left(\frac{\eta}{2}\right)^d V_1 \leq \left(B + \frac{\eta}{2}\right)^d V_1$$

$$\text{so} \quad M(U,\eta) \leq \left(\frac{2B}{\eta} + 1\right)^d = \left(\frac{2B+\eta}{\eta}\right)^d \leq \left(\frac{3B}{\eta}\right)^d,$$

using at the end that $\eta \leq B$ to get a simpler form.

For the lower bound, it holds for a minimal cover T of any set U that

$$\mathrm{vol}(U) \leq \mathrm{vol}\left(\bigcup_{t \in T}\{w : \|w - t\|_p < \eta\}\right) \leq \sum_{t \in T} \mathrm{vol}\left(\{w : \|w - t\|_p < \eta\}\right) = N(U,\eta)V_\eta,$$

where $V_\eta = \mathrm{vol}(\{w : \|w\|_p < \eta\})$. Thus $N(U,\eta) \geq \mathrm{vol}(U)/V_\eta$. Plugging in for U being a $\|\cdot\|_p$ ball in $\mathbb{R}^d$, we obtain the desired lower bound. □

A similar upper bound holds more generally for any finite-dimensional Banach space, getting $(4B/\eta)^d$ [CS02, Proposition 5]. I don't know about a lower bound there. For infinite-dimensional Banach spaces, the lower bound is infinite [Isr15], so to use covering numbers another setup is necessary.

I don't know if the above proofs can be generalized or not.

REFERENCES

[CS02] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society* 39.1 (2002), pages 1–49.

[Isr15] Robert Israel. *Can the ball* $B(0, r_0)$ *be covered with a finite number of balls of radius* $< r_0$. Mathematics Stack Exchange. April 1, 2015.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[Val84] Leslie G. Valiant. A Theory of the Learnable. *Communications of the ACM* 27.11 (1984), pages 1134–1142.

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