CPSC 532D — 2. ERM WITH FINITE HYPOTHESIS CLASSES

Danica J. Sutherland

University of British Columbia, Vancouver

Fall 2024

In (1.5) we showed that, for any $h^* \in \mathcal{H}$,

$$L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \Big(L_{\mathcal{D}}(\hat{h}_S) - L_S(\hat{h}_S)\Big) + \Big(L_S(h^*) - L_{\mathcal{D}}(h^*)\Big).$$

We'd like to bound these two terms, which would then give us a bound on how much worse $\hat{h}_S$ is than h^*, the best thing ERM could have done. The first thing to note, though, is that anything with an S in it – so everything above except for $L_{\mathcal{D}}(h^*)$ – depends on the draw of the random training set S. It's possible that we could get some ridiculously unlikely training set where everything behaves nonsensically. So we'll need to do some kind of probabilistic bound.

Let's now try to study that formally.

2.1 ESTIMATION ERROR: ASYMPTOTICS

Let's start with the second term from (1.5):

$$L_S(h^*) - L_{\mathcal{D}}(h^*) = \frac{1}{m}\sum_{i=1}^{m} \ell(h^*, z_i) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}} \ell(h^*, z).$$

Remember that the only thing that's random here is $S = (z_1, \dots, z_m)$, since h^* is just some fixed hypothesis. So, we can frame this as $\frac{1}{m}\sum_{i=1}^{M} R_i$, where the $R_i = \ell(h^*, z_i)$ are iid random variables with mean $\mathbb{E}\,\ell(h^*, z_i) = L_{\mathcal{D}}(h)$. The law of large numbers therefore guarantees that as $m \to \infty$, $\frac{1}{m}\sum_{i=1}^{m} R_i$ converges (almost surely) to $L_{\mathcal{D}}(h*)$, and so this term in the bound converges to zero.

In fact, for many $\mathcal{H}$ and ℓ, the other term will also have the same property, implying (if h^* is a minimizer of $L_{\mathcal{D}}$) that $L_{\mathcal{D}}(\hat{h}_S) \to L_{\mathcal{D}}(h^*)$. Various formalizations of this last property are called *consistency*, and it's a nice property to have: eventually, your learning algorithm works as well as it could have. One problem with this notion, though, is that this is *all* it tells you. There's no guarantee about what happens with $m = 1,000$, or when going from $m = 1,000$ to $m = 1,000,000$, or anything at all other than "eventually it works."

A more precise analysis might use the central limit theorem. Let $\sigma^2 = \mathrm{Var}[R_i]$ and assume this is finite; informally, the CLT then says that $\frac{1}{m}\sum_{i=1}^{m} R_i$ behaves like $\mathcal{N}(0, \sigma^2/m)$. In fact, it's often true that the first term is also asymptotically normal. This is a nicer result than before: it still doesn't say anything particular for a finite m (maybe the CLT takes a long time to kick in), but it tells us a lot about the asymptotic

Formally, we'd write $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(R_i - \mathbb{E}\,R_i) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

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

behaviour, including both its limiting value but also roughly how much variation we can expect around that value.

It can be tough to find these exact limiting distributions in general, though, and they're not always true (e.g. the one I didn't state for the first term above has some kind-of strict requirements on the way that h is parameterized). A similar but somewhat looser style of bound is to say that the excess error is $\mathcal{O}_p(1/\sqrt{m})$, which is implied by the CLT result above, but can also be much easier to show. Again, this doesn't imply anything for a finite m (just like how $\mathcal{O}$ analyses don't say anything for finite input size on your algorithms), but they do say things like, for reasonably large m, observing four times as much data should roughly halve your excess error.

You can check the wiki page for a formal definition of $\mathcal{O}_p$, but it roughly means "with any constant probability, a sequence of sampled random variables is $\mathcal{O}(1/\sqrt{m})$."

The *most* preferred kind of result, though, is usually one with explicit constants: something like

$$\forall \delta > 0. \qquad \Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\hat{h}_S) - \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) \le \sqrt{\frac{2}{m}\log\frac{|\mathcal{H}|+1}{\delta}}\right) \ge 1-\delta$$

or, where B is a problem parameter,

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} L_{\mathcal{D}}(\hat{h}_S) \le \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \sqrt{\frac{8B^2}{m}}.$$

These results give you a rate, but also apply to *any* m, not just eventually. (They might not be meaningful for small m, though; if you're using 0-1 loss, it's not very helpful to say the excess error is less than four!)

2.2 UNIFORM CONVERGENCE, BOUNDED LOSS

We're first going to assume that $\ell(h,z) \in [a,b]$ for all h, z; usually $a = 0$ (but it won't hurt us to be more general), and e.g. for the 0-1 loss we have $b = 1$. For something like the square loss, it isn't "automatically" bounded, but it might be depending on $\mathcal{H}$ and $\mathcal{D}$; we'll discuss this later.

Recall that we have two things to bound in (1.5):

$$L_{\mathcal{D}}(\hat{h}_S) - L_S(\hat{h}_S) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}} \ell(\hat{h}_S, z) - \frac{1}{m}\sum_{i=1}^{m} \ell(\hat{h}_S, z_i) \tag{A}$$

and

$$L_S(h^*) - L_{\mathcal{D}}(h^*) = \frac{1}{m}\sum_{i=1}^{m} \ell(h^*, z_i) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}} \ell(h^*, z). \tag{B}$$

As we discussed, (B) is an average of iid random variables. We can bound this with the following form of *Hoeffding's inequality*, which we'll prove soon:

Proposition 2.1 (Hoeffding, simple form). *Let $(X_1, \dots, X_m)$ be independent with mean*

μ and almost surely bounded in $[a, b]$. Define $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$. Then

$$\Pr\left(\overline{X} \le \mu + (b-a)\sqrt{\frac{\log(1/\delta)}{2m}}\right) \ge 1-\delta$$

$$\Pr\left(\overline{X} \ge \mu - (b-a)\sqrt{\frac{\log(1/\delta)}{2m}}\right) \ge 1-\delta$$

$$\Pr\left(\left|\overline{X} - \mu\right| \le (b-a)\sqrt{\frac{\log(2/\delta)}{2m}}\right) \ge 1-\delta.$$

The first of these results immediately implies the other two: use the random variables $Y_i = -X_i$ for the second, and then use a union bound, Lemma 2.3, to get the third.

Applying this to the random variables $X_i = \ell(h^*, z_i)$ handles the bound for (B).

It's tempting to also try to apply this result directly to (A), which would then complete our bound and everything would be really simple. The problem is that the $\ell(\hat{h}_S, z_i)$ *aren't independent!* The choice of $\hat{h}_S$ depends on *all* of S, i.e. on all of the other z_j, so $\ell(\hat{h}_S, z_1)$ and $\ell(\hat{h}_S, z_2)$ are *not* independent.

So, how can we bound this? The most common way is called *uniform convergence*. The idea is, if we know that $L_{\mathcal{D}}(h) - L_S(h)$ is small for *all* $h \in \mathcal{H}$, then it'll be small for $\hat{h}_S$, no matter how we pick it – since it's something in $\mathcal{H}$. That is, if we know that

$$\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le \varepsilon$$

then we also have that $L_{\mathcal{D}}(\hat{h}_S) - L_S(\hat{h}_S) \le \varepsilon$. Or, stating it another way,

$$\Pr_{S\sim\mathcal{D}^m}\left(L_S(\hat{h}_S) - L_{\mathcal{D}}(\hat{h}_S) > \varepsilon\right) \le \Pr_{S\sim\mathcal{D}^m}\left(\exists h \in \mathcal{H}.\quad L_S(h) - L_{\mathcal{D}}(h) > \varepsilon\right), \tag{2.1}$$

and so bounding the right-hand side bounds the left-hand side.

How can we bound that?

2.3 FINITE $\mathcal{H}$

To start, we'll make a kind of drastic assumption: that $\mathcal{H}$ is finite, i.e. we're only considering $|\mathcal{H}|$, say 500, possible hypotheses.

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*

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

Proof. For any hypothesis h, we can allow it a "failure probability" of $\delta/(|\mathcal{H}|+1)$ in Hoeffding's inequality:

$$\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(h) - L_S(h) > (b-a)\sqrt{\frac{1}{2m}\log\frac{|\mathcal{H}|+1}{\delta}}\right) \le \frac{\delta}{|\mathcal{H}|+1}.$$

If we do this for *each* hypothesis $h \in \mathcal{H}$, we know that the probability of each particular h being bad is low. We then want to combine them into the probability that *anything* is bad; we can do this with a union bound.

This fact is really *useful.*

Lemma 2.3 (Union bound). *For any two events* A *and* B, $\Pr(A \cup B) \le \Pr(A) + \Pr(B)$.

Combining all of them together, the probability that *any* h happens to look way better than it is can be bounded as

$$\Pr_{S\sim\mathcal{D}^m}\left(\exists h \in \mathcal{H}. \quad L_{\mathcal{D}}(h) - L_S(h) > (b-a)\sqrt{\frac{1}{2m}\log\frac{|\mathcal{H}|+1}{\delta}}\right) \le |\mathcal{H}|\,\frac{\delta}{|\mathcal{H}|+1}.$$

But we'll also need the other direction for (B): h^* in particular doesn't look way worse than it actually is. Giving it the same failure probability to make things nice,

$$\Pr_{S\sim\mathcal{D}^m}\left(L_S(h^*) - L_{\mathcal{D}}(h^*) > (b-a)\sqrt{\frac{1}{2m}\log\frac{|\mathcal{H}|+1}{\delta}}\right) \le \frac{\delta}{|\mathcal{H}|+1}.$$

Now, if $(A) \le \varepsilon_A$ and $(B) \le \varepsilon_B$, then (1.5) tells us that $L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le (A) + (B) \le \varepsilon_A + \varepsilon_B$. Using another union bound,

$$\begin{aligned}
&\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) > (b-a)\sqrt{\frac{1}{2m}\log\frac{|\mathcal{H}|+1}{\delta}} + (b-a)\sqrt{\frac{1}{2m}\log\frac{|\mathcal{H}|+1}{\delta}}\right)\\
&\qquad = \Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) > (b-a)\sqrt{\frac{2}{m}\log\frac{|\mathcal{H}|+1}{\delta}}\right)\\
&\qquad\qquad \le \frac{|\mathcal{H}|}{|\mathcal{H}|+1}\delta + \frac{1}{|\mathcal{H}|+1}\delta = \delta. \qquad \square
\end{aligned}$$

Another way to state Proposition 2.2 is that with m samples, we can achieve excess error at most ε with probability at least $(|\mathcal{H}|+1)\exp\left(-\frac{m\varepsilon^2}{2(b-a)^2}\right)$.

Or, alternately, we can say that we can achieve excess 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.

2.3.1 *Is this finiteness assumption reasonable?*

Every $\mathcal{H}$ we use in practice is finite. Our models are represented on a computer with bounded memory, so we consider no more than $2^{\text{max number of bits}}$ hypotheses.

On the other hand, $|\mathcal{H}|$ might be really large. Typical vision CNNs are around a few hundred megabytes: 100 megabytes is 800,000,000 bits, and $\log(|\mathcal{H}|+1) \approx \log 2^{800{,}000{,}000} = 800{,}000{,}000 \log 2 \approx 554{,}517{,}744$ is quite big. For 0-1 loss, this would mean that for our bound to show that ERM learns a 100-MB network even to within an extremely loose $\varepsilon = 20\%$ additive error with probability at least $1-\delta = 50\%$, we'd need

$$m \ge \frac{2}{0.2^2}\left(\log(|\mathcal{H}|+1) + \log\frac{1}{0.5}\right) \approx 50\,(554 \text{ million} + 0.7) \approx 28 \text{ billion}.$$

100 MB is a relatively small model these days (ViTs are usually a few gigabytes), and 28 billion is a *lot* of samples.

But the union bound we did over $\mathcal{H}$ ignores *all* structure in $\mathcal{H}$. If we change just one parameter by 0.00001, then we're treating the error of that new hypothesis totally separately, when in reality those two errors are tightly correlated. We'll approach that soon, with various techniques that will also allow us to handle $\mathcal{H}$ with infinite size; but first, we'll go back and prove Hoeffding's inequality.