CPSC 532D — 2. ERM WITH FINITE HYPOTHESIS CLASSES

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

Quick reminder of definitions from last time:

- Data $S = (z_1, \dots, z_m) \in \mathcal{Z}^n$ iid from $\mathcal{D}$; typically $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
- Hypothesis class $\mathcal{H}$, typically of functions $h : \mathcal{X} \to \hat{\mathcal{Y}}$.
- Loss $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, typically $\ell(h, (x, y)) = l(h(x), y)$ for $l : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$.
- $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}} \ell(h, z)$, $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$.
- $\mathrm{ERM}(S) = \hat{h}_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$. For example, if $\mathcal{H}$ is all d-dimensional linear predictors and $l(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$, ERM is ordinary least squares.

We also showed that, for any $h^* \in \mathcal{H}$,

$$L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \Big(L_{\mathcal{D}}(\hat{h}_S) - L_S(\hat{h}_S)\Big) + \Big(L_S(h^*) - L_{\mathcal{D}}(h^*)\Big). \tag{1}$$

We'd like to (probabilistically) 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.

1 ERROR DECOMPOSITIONS

Here's some standard terminology to know. For any estimator $\hat{h}_S \in \mathcal{H}$ (not necessarily just the ERM), we can write

$$\underbrace{L_{\mathcal{D}}(\hat{h}_S) - L_{bayes}}_{\text{excess error}} = \underbrace{L_{\mathcal{D}}(\hat{h}_S) - \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)}_{\text{estimation error}} + \underbrace{\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_{bayes}}_{\text{approximation error}}.$$

The *excess error* is how much worse you are than the irreducible error L_{bayes}, also called the Bayes error or the error of the Bayes predictor (see A1 Q2 for more). No predictor, no matter its form, could do better than this: there's just inherent noise in the problem. Our standard notations unfortunately don't make this technically easy to write down (since ℓ's domain is $\mathcal{H}$), but if you use the $\ell(h, (x, y)) = l(h(x), y)$ form, you could write it like $L_{bayes} = \inf_{h : \mathcal{X} \to \hat{\mathcal{Y}} \text{ measurable}} l(h(x), y)$.

The *estimation error*, also called the *statistical error*, is the error that comes about from using your algorithm $\hat{h}_S$ rather than picking the best possible predictor in $\mathcal{H}$. As $m \to \infty$, this should (ideally) go to zero.

The *approximation error* doesn't (directly) depend on the number of samples you see: it's a function of your hypothesis class $\mathcal{H}$.

CPSC 340 used to use "approximation error" for the generalization gap, $L_{\mathcal{D}}(h) - L_S(h)$. This was a nonstandard use; it's been changed now in 340, and you should wipe it from your memory. :)

For example, in the polynomial regression example from last time where the truth was quadratic, using a $\mathcal{H}$ of linear functions results in some approximation error, but not much estimation error (because linear functions are easy to fit). Using a $\mathcal{H}$

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

of degree-twenty polynomials has zero approximation error (it contains the Bayes predictor) but really high estimation error (too many parameters to fit).

Intuitively, as $\mathcal{H}$ gets "bigger," approximation error decreases but estimation error increases. Let's now try to study that formally.

2 Estimation Error: Asymptotics

Let's look first at the second (simpler) term from (1):

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

If $h^* \in \arg\min_{h\in\mathcal{H}} \mathrm{L}_{\mathcal{D}}(h)$ is the optimal hypothesis, then the left-hand side is called the *estimation error*.

Remember that the only thing that's random here is $S = (z_1, \ldots, z_m)$, since h^* is just some fixed hypothesis. Thus the t_i are iid, with mean

$$\mathbb{E}\, t_i = \mathop{\mathbb{E}}_{z_i\sim\mathcal{D}} \ell(h^*, z_i) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}} \ell(h^*, z) = 0.$$

The law of large numbers therefore guarantees that as $m \to \infty$, $\frac{1}{m}\sum_{i=1}^{m} t_i$ converges (almost surely) 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 $\mathrm{L}_{\mathcal{D}}$) that $\mathrm{L}_{\mathcal{D}}(\hat{h}_S) \to \mathrm{L}_{\mathcal{D}}(h^*)$. Versions of this property are called *consistency*, and it's a nice property to have: eventually, your learning algorithm works. 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 $\sigma^2 = \mathrm{Var}[t_i]$ and assume this is finite; informally, the CLT then says that $\frac{1}{m}\sum_{i=1}^{m} t_i$ 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 behaviour, including both its limiting value but also roughly how much variation we can expect around that value.

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

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

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

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; for instance, if you're using 0-1 loss, it's not very helpful to say the excess error is less than four!)

3 UNIFORM CONVERGENCE, BOUNDED LOSS

Let's now try to analyze the excess error of ERM.

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 talk later in the course about what to do if it's not bounded.)

Recall that we have two things to bound in (1):

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

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 in a bit:

PROPOSITION 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,* $\Pr(A \cup B) \le \Pr(A) + \Pr(B)$, *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, as well as the ones we're evaluating on.

So, how can we bound this? The standard technique that we'll study for a while in this course 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 $\hat{h}_S$ in particular. That is, if we know that

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

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

$$\Pr_{S \sim \mathcal{D}^m}\left(L_S(\hat{h}_S) - L_{\mathcal{D}}(\hat{h}_S) > \varepsilon\right) \leq \Pr_{S \sim \mathcal{D}^m}\left(\forall h \in \mathcal{H}. \quad L_S(h) - L_{\mathcal{D}}(h) > \varepsilon\right), \tag{2}$$

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

How can we bound that?

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

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

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

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

This fact is really *useful.*

Apply this to *each* hypothesis $h \in \mathcal{H}$. Then we can combine them all with a *union bound*: recall that for any events A and B, $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$. Combining all of them together, this gives us that

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

Then (2) gives us a bound on (A), i.e. no h looks way better than it actually is:

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

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

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

Notice $\sqrt{\frac{1}{2m}}$ changed to $\sqrt{\frac{2}{m}}$, because we doubled it.

Now, if (A) $\leq \varepsilon_A$ and (B) $\leq \varepsilon_B$, then (A) + (B) $\leq \varepsilon_A + \varepsilon_B$, so one last union bound gives us that

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

Another way to state this result 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.

3.1.1 Is this finiteness assumption reasonable?

In some sense, every $\mathcal{H}$ we use in practice is finite. Our models are represented on a computer in a bounded amount of 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

$$m \geq \frac{2}{0.2^2}\left(\log(|\mathcal{H}| + 1) + \log\frac{1}{0.5}\right) \approx 50\,(554 \text{ million} + 0.7) \approx 27.7 \text{ billion}.$$

100 MB is a relatively small model these days (ViTs are a few gigabytes), and 28 billion is a *lot* of samples. But notice that 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 totally separately, when in reality those two errors are tightly correlated. We'll approach that next, 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.