EN.601.774 Theory of Replicable ML Spring 2025

Lecture 5

Instructor: *Jess Sorrell* Scribe: *Jess Sorrell*

A note on our ERM Experiment

Consider the following test. For a finite hypothesis class $\mathcal{H}$, I believe that my data distribution D over $\mathcal{X} \times \mathcal{Y}$ can't be arbitrarily well-modeled by any $h \in \mathcal{H}$. That is, I believe there exists some $\tau > 0$ such that

$$\min_{h \in \mathcal{H}} \ell_D(h) \geq \tau.$$

We can try to refute this hypothesis using our SQ ERM learner as follows. Set the tolerance for the SQs to be $\tau/4$. Then collect a sample $S_1 \sim_{i.i.d} D^m$ for $|S_1| \in O\left(\frac{\log(|\mathcal{H}|/\delta)}{\tau^2}\right)$, and publish $h_1 \leftarrow \mathcal{L}(S_1)$. Then we can consider two cases:

- if it's truly the case that $\min_{h \in \mathcal{H}} \ell_D(h) \geq \tau$, we'll learn a hypothesis h_1 such that $\ell_{S_1}(h_1) \geq \frac{3\tau}{4}$, except with probability at most δ.
- if it's truly the case that $\min_{h \in \mathcal{H}} \ell_D(h) < \tau/2$, we'll learn a hypothesis h_1 such that $\ell_{S_1}(h_1) \leq \frac{3\tau}{4}$, except with probability at most δ.

Another team of researchers can then attempt to replicate our results by running the same algorithm $\mathcal{L}$ on their own data $S_2 \sim_{i.i.d.} D^m$. It won't necessarily be the case that $h_1 = h_2$. However, if it's truly the case that

$$\min_{h \in \mathcal{H}} \ell_D(h) \geq \tau,$$

then, as before, $\ell_{S_2}(h_2) \geq \frac{3\tau}{4}$ except with probability at most δ. Therefore

- if it's truly the case that $\min_{h \in \mathcal{H}} \ell_D(h) \geq \tau$,

$$\Pr_{S_1, S_2}[\ell_{S_1}(h_1) \geq \tfrac{3\tau}{4} \wedge \ell_{S_2}(h_2) \geq \tfrac{3\tau}{4}] \geq 1 - 2\delta$$

- if it's truly the case that $\min_{h \in \mathcal{H}} \ell_D(h) < \frac{\tau}{2}$, $\Pr_{S_1, S_2}[\ell_{S_1}(h_1) \geq \frac{3\tau}{4} \wedge \ell_{S_2}(h_2) \geq \frac{3\tau}{4}] \leq \delta^2$

Note that even if this replication effort is successful, we can't conclude that our hypothesis is correct. There's still some chance that we drew misleading samples both times. Every successful replication effort, however, drives the probability of undetected "false discovery" down exponentially quickly (just like doubling our sample size).

Algorithmic Replicability

In our previous examples, we've been concerned with replicating the loss of the model h_1 produced by running $\mathcal{L}$ on S_1. We showed that we could do this with good probability with a large enough sample S_2, but what if we wanted to replicate some other property of h_1 or replicate *the model* h_1 itself?

Definition 0.1 (Replicability [Impagliazzo et al., 2022])**.** A randomized algorithm $\mathcal{L}$ is replicable if there exist fns $m_0, \ell_0 : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{N}$ such that for any $\rho, \varepsilon, \delta > 0$ and any distribution D, letting $m \geq m_0(\rho, \varepsilon, \delta)$ and $\ell \geq \ell_0(\rho, \varepsilon, \delta)$,

$$\Pr_{\substack{S_0, S_1 \sim D^m \\ r \sim \{0,1\}^\ell}} [\mathcal{L}(S_0; r) \neq \mathcal{L}(S_1; r)] \leq \rho$$

This definition is essentially equivalent to the notion of *pseudo-global stability* introduced in Ghazi et al. [2021].

Can we design a replicable version of our empirical risk minimization learner? All we really did was make a sequence of $|\mathcal{H}|$ non-adaptive statistical queries, so let's start by seeing if we can replicably answer a single statistical query.

Algorithm 1 $\mathsf{rSTAT}_\tau(\phi, S)$

Inputs/Parameters:

τ - tolerance parameter

δ - accuracy failure probability

ρ - replicability failure parameter

ϕ: a query $X \to [0, 1]$

S: an i.i.d. sample of size m from D

1: $\alpha = \frac{2\tau}{\rho + 1 - 2\delta}$

2: $\alpha_{\text{off}} \leftarrow_r [0, \alpha]$

3: Split $[0, 1]$ in regions: $R = \{[0, \alpha_{\text{off}}), [\alpha_{\text{off}}, \alpha_{\text{off}} + \alpha), \ldots, [\alpha_{\text{off}} + i\alpha, \alpha_{\text{off}} + (i+1)\alpha), \ldots, [\alpha_{\text{off}} + k\alpha, 1)\}$

4: $v \leftarrow \frac{1}{|S|} \sum_{x \in S} \phi(x)$

5: Let r_v denote the region in R that contains v

6: **return** the midpoint of region r_v

We now upper bound the sample complexity of rSTAT.

Theorem 0.2. *Let $\tau > 0$. Then rSTAT_τ is a replicable algorithm for answering statistical queries up to tolerance τ, with $m_0 \in O\left(\frac{\log(1/\delta)}{\tau^2(\rho - \delta)^2}\right)$.*

Proof. We begin by showing that rSTAT_τ answers statistical queries up to tolerance τ, except with probability δ.

Let $\tau' = \frac{\tau(\rho-2\delta)}{\rho+1-2\delta}$. Recall $\alpha = \frac{2\tau}{\rho+1-2\delta}$, so $\frac{2\tau'}{\alpha} = \rho - 2\delta$. Hoeffding's inequality gives us that

$$|\mathbb{E}_S[\phi] - \mathbb{E}_D[\phi]| \leq \tau' = \frac{\tau(\rho-2\delta)}{\rho+1-2\delta}$$

except with failure probability δ, so long as $|S| \geq \frac{\log(2/\delta)}{2\tau'^2}$. Outputting the midpoint of region r_v can further offset this result by at most $\frac{\alpha}{2} = \frac{\tau}{\rho+1-2\delta}$. Therefore

$$|v - \mathbb{E}_D\,\phi| \leq \frac{\tau(\rho-2\delta)}{\rho+1-2\delta} + \frac{\tau}{\rho+1-2\delta} = \tau,$$

except with probability δ, so long as the sample S satisfies $|S| \geq \frac{\log(2/\delta)}{2\tau'^2}$. Unpacking τ', we see that taking $|S| \in O\left(\frac{\log(1/\delta)}{\tau^2(\rho-\delta)^2}\right)$ suffices.

We now show that rSTAT_τ is replicable. Denote by v_1 and v_2 the values returned by the parallel runs $\mathsf{rSTAT}(S_1; r)$ and $\mathsf{rSTAT}(S_2; r)$ at line 4. We consider two cases for failure of replicability:

1. $|v_1 - v_2| > 2\tau'$ and some region's boundary falls between v_1 and v_2
2. $|v_1 - v_2| \leq 2\tau'$ and some region's boundary falls between v_1 and v_2

We bound the probability of the first case by bounding

$$\begin{aligned}\Pr_{S_1,S_2}[|v_1 - v_2| > 2\tau'] &\leq \Pr_{S_1}[|v_1 - \mathbb{E}_D[\phi]| \geq \tau'] + \Pr_{S_2}[|v_2 - \mathbb{E}_D[\phi]| \geq \tau'] \\ &\leq 2\delta\end{aligned}$$

The second case remains. Since α_{off} is chosen uniformly in $[0, \alpha]$, we have

$$\Pr_{\alpha_{\text{off}}}[r_{v_1} \neq r_{v_2}] = \frac{2\tau'}{\alpha} = \rho - 2\delta$$

Union bounding over these two events gives

$$\Pr_{S_1,S_2,r}[\mathsf{rSTAT}_\tau(\phi, S_1) \neq \mathsf{rSTAT}_\tau(\phi, S_2)] = 2\delta + \rho - 2\delta = \rho.$$

□

References

Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. User-level differentially private learning via correlated sampling. *Advances in Neural Information Processing Systems*, 34:20172–20184, 2021.

Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th annual ACM SIGACT symposium on theory of computing*, pages 818–831, 2022.