EN.601.774 Theory of Replicable ML Spring 2025

Lecture 11

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

Acknowledgements. Much of this material (and the material for the next few weeks) is lifted wholesale from the course notes of Aaron Roth and Adam Smith, available at

https://www.adaptivedataanalysis.com

Domain $\mathcal{X} = \{0,1\}^d$, $\mathcal{Y} = \{0,1\}$.

Theorem 0.1 (Bousquet-Elisseef'02)**.** *Let $\mathcal{A}$ be ε-UCO stable, for a hypothesis class $\mathcal{H}$ such that $h \in \mathcal{H}$ is bounded. That is, $h : \mathcal{X} \to [0,1]$ for all $h \in \mathcal{H}$. Then for every distribution D over $\mathcal{X} \times \{0,1\}$, we have that except with probability at most δ over $S \sim D^m$:*

$$|acc_S(h_S) - acc_D(h_S)| \leq \varepsilon + (2\varepsilon m + 1)\sqrt{\frac{\ln 2/\delta}{2m}}$$

Theorem 0.2 (McDiarmid's Inequality)**.** *Let $F : \mathcal{Z}^m \to \mathbb{R}$ be a function such that for all neighboring datasets S, S',*

$$|F(S) - F(S')| \leq \tau.$$

Then

$$\Pr_S[|F(S) - \mathop{\mathbb{E}}_S[F]| > t] \leq 2e^{\frac{-2t^2}{m\tau^2}}$$

Proof idea:

1. Use stability of $\mathcal{A}$ to show that $\mathbb{E}_S[acc_S(h_S) - acc_D(h_S)] \leq \varepsilon$
2. Use stability to show that $|(acc_S(h_S) - acc_D(h_S)) - (acc_{S'}(h_{S'}) - acc_D(h_{S'}))| \leq 2\varepsilon + \frac{M}{m}$
3. Apply McDiarmid's inquality to $F(S) = acc_S(h_S) - acc_D(h_S)$ to show that with high probability, $F(S)$ must be close to its expectation

Claim 0.3. *Let $\mathcal{A}$ be ε-UCO stable. Then for every distribution D over $\mathcal{X} \times \{0,1\}$, the expected generalization error of the classifier is at most ε, that is:*

$$|\mathop{\mathbb{E}}_S[acc_S(h_S) - acc_D(h_S)]| \leq \varepsilon$$

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_{S}[acc_S(h_S) - acc_D(h_S)] &= \mathop{\mathbb{E}}_{S}[\mathop{\mathbb{E}}_{(x,y)\sim D}[|h_S(x) - y|] - \tfrac{1}{m}\sum_{i=1}^{m}|h_S(x_i) - y_i|] \\
&= \tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{\substack{S\\(x,y)\sim D}}[|h_S(x) - y|] - \mathop{\mathbb{E}}_{\substack{S\\(x,y)\sim D}}[|h_S(x_i) - y_i|]\right) && \text{lin of exp} \\
&= \tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{\substack{S\\(x,y)\sim D}}[|h_{S_{i\to(x,y)}}(x_i) - y_i| - |h_S(x_i) - y_i|]\right) && \text{equivalent dist} \\
&\leq \tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{\substack{S\\(x,y)\sim D}}[|h_{S_{i\to(x,y)}}(x_i) - h_S(x_i)|]\right) && \text{triangle ineq} \\
&\leq \tfrac{1}{m}\sum_{i=1}^{m}\varepsilon && \varepsilon\text{-UCO stability} \\
&= \varepsilon
\end{aligned}$$

□

Claim 0.4. *Let $\mathcal{A}$ be an ε-UCO stable algorithm and $h_S = A(S)$. Let*

$$G(h_S) = acc_D(h_S) - acc_S(h_S) = \mathop{\mathbb{E}}_{(x,y)\sim D}[|h_S(x) - y|] - \tfrac{1}{m}\sum_{i=1}^{m}|h_S(x_i) - y_i|.$$

Then

$$|G(h_S) - G(h_{S'})| \leq 2\varepsilon + \tfrac{1}{m}$$

Proof.

$$\begin{aligned}
|G(h_S) - G(h_{S'})| = (&\mathop{\mathbb{E}}_{(x,y)\sim D}[|h_S(x) - y|] - \tfrac{1}{m}\sum_{i=1}^{m}|h_S(x_i) - y_i|) \\
&- (\mathop{\mathbb{E}}_{(x,y)\sim D}[|h_{S'}(x) - y|] - \tfrac{1}{m}\sum_{i=1}^{m}|h_{S'}(x_i') - y_i'|) \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim D}[|h_S(x) - h_{S'}(x)|] + \tfrac{1}{m}\sum_{i=1}^{m}|h_{S'}(x_i') - y_i'| - \tfrac{1}{m}\sum_{i=1}^{m}|h_S(x_i) - y_i| \\
&\leq \varepsilon + \frac{(m-1)\varepsilon + 1}{m} \\
&\leq 2\varepsilon + \frac{1}{m}
\end{aligned}$$

□

Theorem 0.1. The proof follows by applying McDiarmid's inequality to $F(S) = G(h_S)$. We just established that $|F(S) - F(S')| \leq 2\varepsilon + \frac{1}{m} = \tau$. It follows that

$$\begin{aligned}
\Pr_S[|acc_S(h_S) - acc_D(h_S)| > \varepsilon + (2\varepsilon m + 1)\sqrt{\frac{\ln 2/\delta}{2m}}] \\
&= 2\Pr_S\left[G(h_S) > \varepsilon + (2\varepsilon m + 1)\sqrt{\frac{\ln 2/\delta}{2m}}\right] \\
&= 2\Pr_S\left[G(h_S) > \mathop{\mathbb{E}}_S[G(h_S)] + (2\varepsilon m + 1)\sqrt{\frac{\ln 2/\delta}{2m}}\right] \\
&= 2\Pr_S\left[G(h_S) - \mathop{\mathbb{E}}_S[G(h_S)] > \tau\sqrt{\frac{m \ln 2/\delta}{2}}\right] \\
&\leq 2e^{\frac{-2m^2\tau^2 \ln 2/\delta}{2m^2\tau^2}} \\
&= 2e^{-\ln 2/\delta} \\
&= \delta
\end{aligned}$$

□

Great, so what does this tell us about our overfitting SQ learner? Does it tell us something about how we could alter the procedure for answering adaptive SQs to prevent overfitting? If we could somehow make the entire sequence of queries and answers $O(\frac{1}{m})$-UCO-stable, for instance, then by rearranging the above, we'd know that by taking $m \in O(\frac{\log 1/\delta}{\varepsilon^2})$ samples, we could obtain generalization error $O(\varepsilon)$. And hey, that's how many samples we needed to answer a single query! But how do we ensure the entire adaptive sequence of queries is stable? Empirically estimating the value of the query *is* already stable in the sense that changing a single element of the sample can change the output of the estimate by at most $\frac{1}{m}$. And yet, we can observe that the hypothesis output by the SQ learner that answers its queries with empirical estimates is highly unstable.

What if we could ensure stability under post-processing? If we could answer queries through some mechanism $\mathcal{M}$ such that, not only is $\mathcal{M}(S)$ stable, but $\mathcal{A} \circ \mathcal{M}(S)$ is stable for any $\mathcal{A}$, we'd be set!

Definition 0.5 (TV distance)**.** The total variation distance between two distributions D_1, D_2 over events $\mathcal{X}$ is defined

$$d_{TV}(D_1, D_2) = \sup_{X \subseteq \mathcal{X}} |D_1(X) - D_2(X)| = \tfrac{1}{2}\int_{x \in X} |D_1(x) - D_2(x)| dx.$$

If $\mathcal{X}$ is discrete, we have

$$d_{TV}(D_1, D_2) = \tfrac{1}{2}\sum_{x \in \mathcal{X}} |D_1(x) - D_2(x)|.$$

For random variables X, Y, we'll write $d_{TV}(X, Y)$ to denote the TV distance between the distributions of X and Y.

Definition 0.6 (TV stability)**.** A randomized algorithm $\mathcal{M}$ is ε-TV stable if for all neighboring datasets S, S',

$$d_{TV}(\mathcal{M}(S), \mathcal{M}(S')) \leq \varepsilon$$

Claim 0.7 (TV stability is preserved under post-processing.)**.** *Let X,Y be random variables over $\mathcal{Z}$. Then for every (potentially randomized) algorithm $\mathcal{A} : \mathcal{Z} \to \mathcal{O}$*

$$d_{TV}(\mathcal{A}(X), \mathcal{A}(Y)) \leq d_{TV}(X, Y)$$

Proof. Let R denote the randomness of $\mathcal{A}$ and let X, Y be random variables over $\mathcal{Z}$. We first observe that, because R is independent of X and Y

$$\begin{aligned} d_{TV}((X,R),(Y,R)) &= \tfrac{1}{2} \int_{z \in Z} \int_{r \in R} |\Pr_{X}(z) \cdot \Pr(r) - \Pr_{Y}(z) \cdot \Pr(r)| dr dz \\ &= \tfrac{1}{2} \int_{z \in Z} |\Pr_{X}(z) - \Pr_{Y}(z)| dz \\ &= d_{TV}(X, Y) \end{aligned}$$

Letting $F = \mathcal{A}^{-1}(O)$ denote the set of $(z, r) \in \mathcal{Z} \times \mathcal{R}$ such that $\mathcal{A}(z) \in O$.
Then

$$\begin{aligned} \Pr(\mathcal{A}(X) \in O) - \Pr(\mathcal{A}(Y) \in O) &= \Pr((X,R) \in F) - (\Pr(Y,R) \in F) \\ &\leq d_{TV}((X,R),(Y,R)) \\ &= d_{TV}(X, Y) \end{aligned}$$

□

Algorithm 1 Gaussian mechanism(σ^2, S)

Inputs/Parameters:
σ^2, variance for Gaussian
$S = \{x_i\}_{i=1}^m$, dataset

```
Receive a statistical query $\phi : \mathcal{X} \to [0,1]$
$\nu \leftarrow \mathcal{N}(0, \sigma^2)$
return $\frac{1}{m}\sum_{i=1}^m \phi(x_i) + \nu$
```

Claim 0.8. *The Gaussian mechanism with parameter σ^2 is $\frac{1}{2m\sigma}$-TV stable.*