EN.601.774 Theory of Replicable ML	Spring 2025
Lecture 14	
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. *Let $\mathcal{M} : \mathcal{Z}^m \to \mathcal{Q}$ be an ε-TV stable algorithm that outputs a query $q \in \mathcal{Q}$. Then for every distribution D, except with probability δ over the choice of sample:*

$$|\mathop{\mathbb{E}}_{r}[q_S(S) - q_S(D)]| \leq \varepsilon + (2\varepsilon m + 1)\sqrt{\frac{\log 2/\delta}{m}},$$

where $q_S = \mathcal{M}(S)$.

1. Use stability of $\mathcal{M}$ to prove that $|\mathbb{E}_{\substack{S\sim D^m \\ r}}[q_S(S) - q_S(D)]| \leq \varepsilon$
2. Prove that $G(S) = \mathbb{E}_r[q_S(S) - q_S(D)]$ satisfies $|G(S) - G(S')| \leq 2\varepsilon + \frac{1}{m}$
3. Apply McDiarmid's inequality to conclude that $G(S)$ must be close to $\mathbb{E}_{S\sim D^m}[G(S)]$ with high probability, and so the *expected* generalization error of q_S (over the internal randomness r of $\mathcal{M}$) must be small with high probability.

Claim 0.2. *Let $\mathcal{M} : \mathcal{Z}^m \to \mathcal{Q}$ be an ε-TV stable algorithm that outputs a query $q \in \mathcal{Q}$. Then for every distribution D,we have:*

$$|\mathop{\mathbb{E}}_{\substack{S\sim D^m \\ r}}[q_S(S) - q_S(D)]| \leq \varepsilon$$

Claim 0.3. *Let $G(S) = \mathbb{E}_r[q_S(S) - q_S(D)]$. Then $|G(S) - G(S')| \leq \varepsilon$.*

Step 3, apply McDiarmid! Now that we've done steps 2 and 3, this is really the same argument from last lecture, applying McDiarmid's inequality to $G(S) = \mathbb{E}_r[q_S(S) - q_S(D)]$.

We just established that $|G(S) - G(S')| \le 2\varepsilon + \frac{1}{m}$ (call this τ). It follows that

$$\begin{aligned}
\Pr_S[|\mathop{\mathbb{E}}_r[q_S(S) - q_S(D)]| > \varepsilon + (2\varepsilon m + 1)\sqrt{\frac{\ln 2/\delta}{2m}}] \\
&= 2\Pr_{S,r}\left[G(S) > \varepsilon + (2\varepsilon m + 1)\sqrt{\frac{\ln 2/\delta}{2m}}\right] \\
&\le 2\Pr_S\left[G(h_S) > \mathop{\mathbb{E}}_S[G(h_S)] + (2\varepsilon m + 1)\sqrt{\frac{\ln 2/\delta}{2m}}\right] \\
&= 2\Pr_S\left[G(h_S) - \mathop{\mathbb{E}}_S[G(h_S)] > \tau\sqrt{\frac{m\ln 2/\delta}{2}}\right] \\
&\le 2e^{\frac{-2m^2\tau^2\ln 2/\delta}{2m^2\tau^2}} \\
&= 2e^{-\ln 2/\delta} \\
&= \delta.
\end{aligned}$$

It follows that

$$\Pr_S\left[|\mathop{\mathbb{E}}_r[q_S(S) - q_S(D)]| > \varepsilon + (2\varepsilon m + 1)\sqrt{\frac{\log 2/\delta}{m}}\right] \le \delta.$$

Great! So now we have a (weaker than we'd like) generalization guarantee for TV-stable algorithms. But for this to help us prove generalization guarantees for a sequence of adaptive statistical queries, we now need to show that TV-stability is preserved under composition.

Previously, we showed that it's preserved under post-processing, so any algorithm that takes the output of a TV-stable algorithm as its only input will itself be TV-stable. But what about the composition of many such algorithms?

Let $\mathcal{M}$ be a mechanism that takes as input a dataset S and interacts with an analyst $\mathcal{A}$ over k rounds, receiving adaptively chosen queries from $\mathcal{A}$ and responding with answers to these queries. We can break this mechanism into k separate mechanisms $M_1, M_2, \ldots, M_k$, each of which take as input

- Dataset S
- A query from the analyst ϕ
- Global *state* (we need to add this state input to model the memory of the interaction between $\mathcal{A}$ and $\mathcal{M}$)

and output

- An answer to query ϕ

- Updated global *state*

These separate mechanisms interact with k separate algorithms $\mathcal{A}_i$, which take as input

- The answer to a query, a
- Global *state*

and output

- A new query ϕ
- Updated global *state*

We will call $\mathcal{M}$ the *adaptive sequential composition* of $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$. And write $\mathcal{A} \circ \mathcal{M}$ to denote the interactive process between $\mathcal{A}$ and $\mathcal{M}$.

$\mathcal{M}_i(S, \phi, state)$ is ε-TV stable if for all neighboring datasets S, S', for all queries ϕ, and for all values of $state_{i-1}$, the distribution over $(a, state_i)$ outputs of $\mathcal{M}_i$ satisfies:

$$d_{TV}(\mathcal{M}_i(S, \phi, state_{i-1}), \mathcal{M}_i(S', \phi, state_{i-1})) \leq \varepsilon$$

Theorem 0.4. *Let $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k)$ be the sequential adaptive composition of k mechanisms, each of which is ε-TV stable. Then for any algorithm $\mathcal{A}$ that is a post-processing of $\mathcal{M}_i$'s, the interaction $\mathcal{A} \circ \mathcal{M}$ is $k\varepsilon$-TV stable.*

Now that we've decomposed the interaction between $\mathcal{A}$ and $\mathcal{M}$ into k exchanges between $\mathcal{A}_i$ and $\mathcal{M}_i$, note that if $\mathcal{M}_i$ is ε-TV stable, then $\mathcal{A}_{i+1} \circ \mathcal{M}_i$ is also ε-TV stable by the post-processing result we showed previously. $\mathcal{A}_{i+1}$ only takes the outputs of $\mathcal{M}_i$ as input (the answer a_i and the state $state_i$), so it is simply a post-processing of a stable algorithm. So for notational simplicity, we can let $\mathcal{M}_i$ "absorb" $\mathcal{A}_{i+1}$, so now $\mathcal{M}_i$ takes as input

- Dataset S
- A query ϕ_i
- Global $state_{i-1}$ (we need to add this state input to model the memory of the interaction between $\mathcal{A}$ and $\mathcal{M}$)

and outputs

- A new query ϕ_{i+1}
- Updated global $state_i$

Let $Y = \mathcal{M}(S)$ be the random variable denoting the sequence of outputs of $\mathcal{M}$ on dataset S when it is interacting with a fixed $\mathcal{A}$. Let $\mathcal{O}^k$ be the outcome space that Y is distributed over. Y can also be written as a joint distribution $Y = (Y_1, Y_2, \ldots, Y_k)$, where $Y_i = \mathcal{M}_i(S, Y_1, Y_2, \ldots, Y_{i-1})$. We will similarly write $Z = \mathcal{M}(S')$, and $Z = (Z_1, Z_2, \ldots, Z_k)$ for $Z_i = \mathcal{M}_i(S', Z_1, Z_2, \ldots, Z_k)$.

To show that $\mathcal{M}$ is $k\varepsilon$-TV stable, it would suffice to show that $d_{TV}(Y_k, Z_k) \leq k\varepsilon$. Next time, we will show something somewhat stronger, which is

$$d_{TV}(Y, Z) \leq k\varepsilon.$$