EN.601.774 Theory of Replicable ML	Spring 2025
Lecture 13	
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:*

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

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

Wait, why aren't we proving high probability bounds? These may be high probability over the sample, but they're only in expectation over the randomness. Ideally, we would like a statement of the form

Theorem 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, except with probability δ*

$$\Pr_{S,r}[|q_S(S) - q_S(D)| > \varepsilon + f(\varepsilon, \delta, m)] \leq \delta$$

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

like we had for UCO algorithms. We will eventually prove a theorem of this form, but let's first see what happens when we try to follow the UCO argument, but for TV-stability.

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

What happens when we try to rerun that argument now?

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 $|G(S) - G(S')| \leq \varepsilon$

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.

Let $q_{S;r} = A(S;r)$ and note that trying $G(S,r) = q_{S;r}(S) - q_{S;r}(D)$ doesn't satisfy the assumptions of McDiarmid's inequality! We have no stability guarantees regarding perturbations to the randomness of, e.g., the Gaussian mechanism. Similarly, we might want to try $G_r(S) = q_{S;r}(S) - q_{S;r}(D)$, but our stability guarantees are on the distribution of outputs of $\mathcal{M}$, not a single output. The only function we have stability guarantees for here is $G(S) = \mathbb{E}_r[q_S(S) - q_S(D)]$.

Claim 0.3. *For all X, Y on $\mathcal{O}$ with $d_{TV}(X,Y) \leq \varepsilon$, and for all functions $f : \mathcal{O} \to [0,1]$,*

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \leq \varepsilon$$

Proof. Let D_X, D_Y denote the distributions of X, Y.

$$\begin{aligned}
\mathbb{E}[f(X)] - \mathbb{E}[f(Y)] &= \int_{o\in\mathcal{O}} f(o)D_X(o)do - \int_{o\in\mathcal{O}} f(o)D_Y(o)do \\
&= \int_{o\in\mathcal{O}} f(o)(D_X(o) - D_Y(o))do \\
&= \int_{o:D_X(o)>D_Y(o)} f(o)(D_X(o) - D_Y(o))do + \int_{o:D_X(o)\leq D_Y(o)} f(o)(D_X(o) - D_Y(o))do \\
&\leq \int_{o:D_X(o)>D_Y(o)} f(o)(D_X(o) - D_Y(o))do \\
&\leq \int_{o:D_X(o)>D_Y(o)} |D_X(o) - D_Y(o)|do \\
&= d_{TV}(X,Y) \\
&\leq \varepsilon
\end{aligned}$$

□

Claim 0.4. *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:*

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

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{S\sim D^m\\ r}}[q_S(S)-q_S(D)] &= \mathop{\mathbb{E}}_{S,r}[\tfrac{1}{m}\sum_{i=1}^{m} q_S(z_i) - \mathop{\mathbb{E}}_{z\sim D}[q_S(z)]] \\
&= \tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{S,r}[q_S(z_i)] - \mathop{\mathbb{E}}_{\substack{S,r\\ z\sim D}}[q_S(z)]\right) && \text{lin of exp} \\
&= \tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{\substack{S,r\\ z\sim D}}[q_S(z_i) - q_{S_{i\to z}}(z_i)]\right) && \text{equivalent dist} \\
&= \tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{\substack{S\\ z\sim D}}[\mathop{\mathbb{E}}_{r}[q_S(z_i)] - \mathop{\mathbb{E}}_{r}[q_{S_{i\to z}}(z_i)]]\right)
\end{aligned}$$

From ε-TV stability of $\mathcal{M}$, we know that $d_{TV}(q_S, q_{S_{i\to z}}) \leq \varepsilon$, but how does that help us bound

$$\mathop{\mathbb{E}}_{r}[q_S(z_i)] - \mathop{\mathbb{E}}_{r}[q_{S_{i\to z}}(z_i)]?$$

Let $f_{z_i}(q) = q(z_i)$. Then

$$\mathop{\mathbb{E}}_{r}[q_S(z_i)] - \mathop{\mathbb{E}}_{r}[q_{S_{i\to z}}(z_i)] = \mathop{\mathbb{E}}_{r}[f_{z_i}(q_S)] - \mathop{\mathbb{E}}_{r}[f_{z_i}(q_{S_{i\to z}})]$$

Note that $f_{z_i} : \mathcal{Q} \to [0,1]$, and in the equation above, it's evaluated on two random variables that have $d_{TV}(q_S, q_{S_{i\to z}}) \leq \varepsilon$. Then from our result in Step 1, we have that for all $S \in \mathbb{Z}^m, z \in \mathbb{Z}$,

$$\mathop{\mathbb{E}}_{r}[f_{z_i}(q_S)] - \mathop{\mathbb{E}}_{r}[f_{z_i}(q_{S_{i\to z}})] \leq \varepsilon$$

Applying this bound, we continue

$$\begin{aligned}
\tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{\substack{S\\ z\sim D}}[\mathop{\mathbb{E}}_{r}[q_S(z_i)] - \mathop{\mathbb{E}}_{r}[q_{S_{i\to z}}(z_i)]]\right) &\leq \tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{\substack{S\\ z\sim D}}\varepsilon\right) && \text{from } \varepsilon\text{-TV stability, Step 1} \\
&= \varepsilon
\end{aligned}$$

□

Step 2 down! On to Step 3.

Claim 0.5. *Let $G(S) = \mathbb{E}_r[q_S(S) - q_S(D)]$. Then $|G(S) - G(S')| \leq \varepsilon$.*

Proof. We will again consider functions $f : \mathcal{Q} \to [0,1],\ f_z(q) = q(z)$

$$\begin{aligned}
|G(S) - G(S')| &= |\mathop{\mathbb{E}}_r[q_S(S) - q_S(D)] - \mathop{\mathbb{E}}_r[q_{S'}(S') - q_{S'}(D)]| \\
&= |\tfrac{1}{m}\sum_{i=1} \mathop{\mathbb{E}}_r[q_S(S) - q_S(D) - q_{S'}(S') + q_{S'}(D)]| \\
&= |\tfrac{1}{m}\sum_{i=1} \mathop{\mathbb{E}}_r[q_S(z_i) - q_{S'}(z_i')] + \mathop{\mathbb{E}}_{z\sim D}\mathop{\mathbb{E}}_r[q_{S'}(z) - q_S(z)]]| \\
&= |\tfrac{1}{m}\sum_{i=1} \mathop{\mathbb{E}}_r[q_S(z_i) - q_{S'}(z_i')] + \mathop{\mathbb{E}}_{z\sim D}\mathop{\mathbb{E}}_r[f_z(q_{S'})] - \mathop{\mathbb{E}}_r[f_z(q_S)]]| \\
&\le |\tfrac{1}{m}\sum_{i=1} \mathop{\mathbb{E}}_r[q_S(z_i) - q_{S'}(z_i')] + \varepsilon| \\
&\le |\tfrac{1}{m}\sum_{i=1} \mathop{\mathbb{E}}_r[f_{z_i}(q_S) - f_{z_i'}(q_{S'})] + \varepsilon| \\
&\le |\tfrac{(m-1)\varepsilon}{m} + \tfrac{1}{m} + \varepsilon| \\
&\le 2\varepsilon + \tfrac{1}{m}
\end{aligned}$$

□