| EN.601.774 Theory of Replicable ML | Spring 2025 |
| --- | --- |

# Lecture 15

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

**Acknowledgements.** Much of this material 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 $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k)$ be the sequential adaptive composition of $k$ mechanisms, each of which is $\varepsilon$-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.*

*Proof.* 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 $\varepsilon$-TV stable, then $\mathcal{A}_{i+1} \circ \mathcal{M}_i$ is also $\varepsilon$-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 $\phi_i$
- 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 $\phi_{i+1}$
- 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 $d_{TV}(Y_k, Z_k) \leq k\varepsilon$. We will show something somewhat stronger, which is

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

Let $Y_1^i = (Y_1, Y_2, \ldots, Y_i)$ and similarly $Z_1^i = (Z_1, Z_2, \ldots, Z_i)$. We will argue inductively, showing

Base case: $d_{TV}(Y_1, Z_1) \leq \varepsilon$

Inductive step: $d_{TV}(Y_1^{i+1}, Z_1^{i+1}) \leq d_{TV}(Y_1^i, Z_1^i) + \varepsilon$.

To argue the base case, we just need the stability of $\mathcal{M}_1$. It follows immediately that

$$d_{TV}(Y_1, Z_1) = d_{TV}(\mathcal{M}_1(S), \mathcal{M}_1(S')) \leq \varepsilon.$$

We now turn to arguing the inductive step. Let $o_1^i = (o_1, o_2, \ldots, o_i)$ and let $Y_1^{i+1}(o_1^{i+1}) = \Pr[Y_1^{i+1} = o_1^{i+1}]$. We want to bound

$$\begin{aligned} d_{TV}(Y_1^{i+1}, Z_1^{i+1}) &= \tfrac{1}{2} \int_{o \in \mathcal{O}^{i+1}} |Y_1^{i+1}(o) - Z_1^{i+1}(o)| do \\ &= \tfrac{1}{2} \int_{o_1^i \in \mathcal{O}^i} \int_{o \in \mathcal{O}} |Y_1^{i+1}(o, o_1^i) - Z_1^{i+1}(o, o_1^i)| do \, do_1^i \end{aligned}$$

Observe that for all $o = (o_1, \ldots, o_{i+1}) \in \mathcal{O}^{i+1}$,

$$Y_1^{i+1}(o_{i+1}, o_1^i) = Y_{i+1}(o_{i+1} \mid o_1^i) \cdot Y_1^i(o_1^i).$$

So for all $o \in \mathcal{O}$,

$$\begin{aligned} Y_1^{i+1}(o, o_1^i) - Z_1^{i+1}(o, o_1^i) &= Y_{i+1}(o \mid o_1^i) \cdot Y_1^i(o_1^i) - Z_{i+1}(o \mid o_1^i) \cdot Z_1^i(o_1^i) \\ &= Y_{i+1}(o \mid o_1^i) \cdot Y_1^i(o_1^i) - Z_{i+1}(o \mid o_1^i) \cdot Z_1^i(o_1^i) \\ &\quad + Z_{i+1}(o \mid o_1^i) \cdot Y_1^i(o_1^i) - Z_{i+1}(o \mid o_1^i) \cdot Y_1^i(o_1^i) \\ &= Y_1^i(o_1^i)(Y_{i+1}(o \mid o_1^i) - Z_{i+1}(o \mid o_1^i)) + Z_{i+1}(o \mid o_1^i)(Y_1^i(o_1^i) - Z_1^i(o_1^i)) \end{aligned}$$

Inserting the above into our expression for $d_{TV}(Y_1^{i+1}, Z_1^{i+1})$, we obtain

$$\begin{aligned} d_{TV}(Y_1^{i+1}, Z_1^{i+1}) &= \tfrac{1}{2} \int_{o_1^i \in \mathcal{O}^i} \int_{o \in \mathcal{O}} |Y_1^{i+1}(o, o_1^i) - Z_1^{i+1}(o, o_1^i)| do \, do_1^i \\ &= \tfrac{1}{2} \int \int |Y_1^i(o_1^i)(Y_{i+1}(o \mid o_1^i) - Z_{i+1}(o \mid o_1^i)) + Z_{i+1}(o \mid o_1^i)(Y_1^i(o_1^i) - Z_1^i(o_1^i))| do \, do_1^i \\ &\leq \tfrac{1}{2} \int \int Y_1^i(o_1^i)|Y_{i+1}(o \mid o_1^i) - Z_{i+1}(o \mid o_1^i)| + \tfrac{1}{2} \int \int Z_{i+1}(o \mid o_1^i)|Y_1^i(o_1^i) - Z_1^i(o_1^i)| do \, do_1^i \\ &= \mathop{\mathbb{E}}_{o_1^i \sim Y_1^i}[d_{TV}(Y_{i+1}, Z_{i+1}) \mid Y_1^i = o_1^i] + \tfrac{1}{2} \int_{o_1^i \in \mathcal{O}^i} |Y_1^i(o_1^i) - Z_1^i(o_1^i)| do_1^i \\ &\leq \varepsilon + d_{TV}(Y_1^i, Z_1^i) \end{aligned}$$

It follows that

$$d_{TV}(\mathcal{M}(S), \mathcal{M}(S')) = d_{TV}(Y, Z) = d_{TV}(Y_1^k, Z_1^k) = k\varepsilon$$

□

# Adaptive composition for DP algorithms

**Definition 0.2** (Approximate Differential Privacy)**.** A randomized algorithm $\mathcal{M} : \mathcal{Z}^m \to \mathcal{O}$ is $(\varepsilon, \delta)$-DP if for all measurable subsets $T \subset \mathcal{O}$ and neighboring datasets $S, S'$:

$$\Pr[\mathcal{M}(S) \in T] \leq e^{\varepsilon} \Pr[\mathcal{M}(S') \in T] + \delta$$

Note that for $(\varepsilon, 0)$-DP algorithms, this is equivalent to the statement

$$\ln\left(\frac{\Pr[\mathcal{M}(S) \in O]}{\Pr[\mathcal{M}(S') \in O]}\right) \leq \varepsilon$$

**Theorem 0.3.** *For all $\varepsilon \geq 0$ and $\delta' > 0$, the adaptive composition of $k$ algorithms, each of which is $\varepsilon$-DP, is $(\epsilon\sqrt{2k \ln 1/\delta'} + k\varepsilon(e^{\varepsilon} - 1), \delta')$-DP.*

# Putting it all together

Let $\mathcal{A}$ be a statistical query algorithm that makes $k$ queries to $\mathcal{M}$, all of which are answered by the Gaussian mechanism. Then to ensure expected generalization error at most $\tau$ for $\mathcal{A} \circ \mathcal{M}$, it suffices to take $\sigma \in O(\frac{k}{\tau\sqrt{m}})$.

- We previously showed the Gaussian mechanism with parameter $\sigma$ is $\frac{1}{\sqrt{2\pi} m\sigma}$-TV stable
- We just finished showing that TV-stable algorithms compose. So the interaction $\mathcal{A} \circ \mathcal{M}$ is $\frac{k}{\sqrt{2\pi} m\sigma}$-TV stable
- We previously showed that TV-stable algorithms have small expected generalization error

Specifically, we showed that for all distributions $D$, letting $q_S \leftarrow \mathcal{A} \circ \mathcal{M}(S)$ where $\mathcal{A} \circ \mathcal{M}$ is an $\varepsilon$-TV stable algorithm, that except with probability at most $\delta$ over $S \sim D^m$

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

Plugging in $\frac{k}{\sqrt{2\pi} m\sigma}$ for $\varepsilon$, and assuming $\sigma < 1$ (which it better be if we want our statistical queries to not be totally drowned out by noise), we have

$$\begin{aligned}
|\mathop{\mathbb{E}}_{r}[q_S(S) - q_S(D)]| &\leq \frac{k}{\sqrt{2\pi} m\sigma} + (\frac{2k}{\sqrt{2\pi}\sigma} + 1)\sqrt{\tfrac{\log 2/\delta}{m}} \\
&\leq \frac{k}{\sqrt{2\pi} m\sigma} + \frac{3k}{\sqrt{2\pi}\sigma}\sqrt{\tfrac{\log 2/\delta}{m}} \\
&= \frac{k}{\sqrt{2\pi} m\sigma} + \frac{3k\sqrt{\log 2/\delta}}{\sqrt{2\pi m}\sigma} \\
&\in O\left(\frac{k\sqrt{\log 1/\delta}}{\sqrt{m}\sigma}\right)
\end{aligned}$$

So if we want generalization error no greater than $\tau$, it suffices to take

$$\sigma \in O\left(\frac{k\sqrt{\log 1/\delta}}{\sqrt{m}\tau}\right).$$

Note that if we want $\sigma \in o(1)$, we need to take

$$m \in \omega\left(\frac{k^2\sqrt{\log(1/\delta)}}{\tau^2}\right)$$

But using the composition theorem for differential privacy, the picture looks a little different.

- We can show that the Laplace mechanism with parameter $\varepsilon$ is $(\varepsilon, 0)$-DP
- We claimed that DP algorithms compose. So the interaction $\mathcal{A} \circ \mathcal{M}$ is $(\epsilon\sqrt{2k \ln 1/\delta'} + k\varepsilon(e^{\varepsilon} - 1), \delta')$-DP for all $\delta$
- We showed that $(\varepsilon, 0)$-DP algorithms are TV-stable,

  $$d_{TV} \le \tfrac{1}{2}(e^{\varepsilon} - 1) + \delta$$

  so we can show that $\mathcal{A} \circ \mathcal{M}$ is TV-stable:

$$\begin{aligned}
d_{TV}(\mathcal{A} \circ \mathcal{M}(S), \mathcal{A} \circ \mathcal{M}(S')) &\le \tfrac{1}{2}(e^{\epsilon\sqrt{2k\ln 1/\delta'} + k\varepsilon(e^{\varepsilon}-1)} - 1) + \delta' \\
&\le \tfrac{1}{2}(e^{\epsilon\sqrt{2k\ln 1/\delta'} + k\varepsilon(e^{\varepsilon}-1)} - 1) + \delta' \\
&\approx \tfrac{1}{2}(e^{\epsilon\sqrt{2k\ln 1/\delta'} + k\varepsilon^2} - 1) + \delta' && \text{if } \varepsilon < 1 \text{ from } e^{\varepsilon} \approx 1 + \varepsilon \\
&\le \tfrac{1}{2}(e^{2\epsilon\sqrt{2k\ln 1/\delta'}} - 1) + \delta' && \text{if } \varepsilon < \tfrac{1}{\sqrt{k}} \\
&\le \tfrac{1}{2}(2\epsilon\sqrt{2k \ln 1/\delta'}) + \delta' && \text{if } \varepsilon < \tfrac{1}{2\sqrt{2k\ln 1/\delta'}} \\
&= \epsilon\sqrt{2k \ln 1/\delta'} + \delta' && \text{if } \varepsilon < \tfrac{1}{2\sqrt{2k\ln 1/\delta'}} \\
&\in \tilde{O}(\varepsilon\sqrt{k}) && \text{take } \delta' = \varepsilon\sqrt{k} \text{ and ignore log factors}
\end{aligned}$$

Plugging $O(\varepsilon\sqrt{k})$ in for $\alpha$ in our expected generalization guarantees:

$$\begin{aligned}
|\mathop{\mathbb{E}}_{r}[q_S(S) - q_S(D)]| &\le \tau + (2\alpha m + 1)\sqrt{\tfrac{\log 2/\beta}{m}} \\
&\le \varepsilon\sqrt{k} + (2\varepsilon\sqrt{k}m + 1)\sqrt{\tfrac{\log 2/\beta}{m}} \\
&\le \varepsilon\sqrt{k} + 3\varepsilon\sqrt{k}m\sqrt{\tfrac{\log 2/\beta}{m}} \\
&\le \varepsilon\sqrt{k} + 3\varepsilon\sqrt{km \log 2/\beta} \\
&\in \tilde{O}(\varepsilon\sqrt{km})
\end{aligned}$$

So if I want expected generalization error smaller than $\tau$, I need $\varepsilon \in \tilde{O}(\frac{\tau}{\sqrt{km}})$.