EN.601.774 Theory of Replicable ML	Spring 2025

Lecture 12

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

Definition 0.1 (TV stability)**.** A randomized algorithm $\mathcal{M}$ is ε-TV stable if for all neighboring datasets S, S',

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

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

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

Claim 0.2. *The Gaussian mechanism with parameter σ^2 is $\frac{1}{\sqrt{2\pi}m\sigma}$-TV stable.*

Proof. For dataset S, the output is distributed as a Gaussian distribution D with $\mu = \frac{1}{m}\sum_{i=1}^m \phi(x_i)$ and variance σ^2. For dataset S', it is distributed as a Gaussian D' with $\mu' = \frac{1}{m}\sum_{i=1}^m \phi(x_i')$ and variance σ^2, where $|\mu' - \mu| \leq \frac{1}{m}$.

So for all $x \in \mathcal{X}$,

$$D(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$D'(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu')^2}{2\sigma^2}}$$

WLOG, say $\mu' = \mu + \frac{1}{m}$. Then $D(x) > D'(x)$ up to some point $\bar{\mu}$, where $D(\bar{\mu}) = D'(\bar{\mu})$.

$$\begin{aligned}(x-\mu)^2 &= (x-\mu')^2 \\ &= (x - \mu - \tfrac{1}{m})^2 \\ &= (x-\mu)^2 + \tfrac{1}{m^2} - 2(\frac{x-\mu}{m})\end{aligned}$$

So we want to find x such that

$$0 = \tfrac{1}{m} - 2(x-\mu)$$
$$x = \mu + \tfrac{1}{2m}$$

Therefore

$$\begin{aligned}
d_{TV}(D, D') &= \tfrac{1}{\sqrt{2\pi\sigma^2}} \int_{\frac{-1}{2m}}^{\frac{1}{2m}} e^{-\frac{x^2}{2\sigma^2}} dx \\
&\leq \tfrac{1}{\sqrt{2\pi\sigma^2}} \int_{\frac{-1}{2m}}^{\frac{1}{2m}} 1 dx \\
&= \frac{1}{\sqrt{2\pi} m\sigma}
\end{aligned}$$

□

Definition 0.3 (Multiplicative distance (pure DP metric))**.** The multiplicative distance between two distributions D_1, D_2 over events $\mathcal{X}$ is defined

$$d_\diamond(D_1, D_2) = \sup_{X \subseteq \mathcal{X}} \left| \ln \frac{D_1(X)}{D_2(X)} \right|$$

If $d_\diamond(D_1, D_2) \leq \varepsilon$, then for all $X \subseteq \mathcal{X}$

$$D_1(X) \leq e^\varepsilon D_2(X).$$

This notion of distance gives another stability notion called *differential privacy*

Definition 0.4 (Differential Privacy)**.** A randomized algorithm $\mathcal{M} : \mathcal{Z}^m \to \mathcal{O}$ is ε-DP if for all measurable subsets $O \subset \mathcal{O}$ and neighboring datasets S, S':

$$\Pr[\mathcal{M}(S) \in O] \leq e^\varepsilon \Pr[\mathcal{M}(S') \in O]$$

Laplace distribution $Lap(\varepsilon, \mu)$ has PDF $f(x) = \frac{1}{2\varepsilon} e^{\frac{-|x-\mu|}{\varepsilon}}$. We'll write $Lap(\varepsilon) = Lap(\varepsilon, 0)$.

Algorithm 2 Laplace mechanism(ε, S)
Inputs/Parameters:
ε, scale parameter for Laplace distribution
$S = \{x_i\}_{i=1}^m$, dataset

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

Claim 0.5. *The Laplace mechanism with parameter ε satisfies ε-DP*

Proof. The Laplace distribution $Lap(\frac{1}{m\varepsilon})$ has PDF $f(x) = \frac{m\varepsilon}{2}e^{-m\varepsilon|x-\mu|}$. For neighboring datasets, $|\mu - \mu'| \leq \frac{1}{m}$, and so for all x the ratio

$$\begin{aligned}\frac{D(x)}{D'(x)} &= \frac{e^{-m\varepsilon|x-\mu|}}{e^{-m\varepsilon|x-\mu'|}} \\ &= e^{m\varepsilon(|x-\mu'|-|x-\mu|)} \\ &\leq e^{\varepsilon}.\end{aligned}$$

□

Claim 0.6 (DP ⇒ TV stability)**.** $d_{TV}(D, D') \leq \frac{1}{2}(e^{d_\diamond(D,D')} - 1)$

Proof.

$$\begin{aligned}d_{TV}(D, D') &= \sup_{E\subseteq\mathcal{O}} |D(E) - D'(E)| \\ &\leq \sup_{E\subseteq\mathcal{O}} |e^{d_\diamond(D,D')}D'(E) - D'(E)| \\ &= \sup_{E\subseteq\mathcal{O}} |D'(E)(e^{d_\diamond(D,D')} - 1)| \\ &\leq \tfrac{1}{2}(e^{d_\diamond(D,D')} - 1) && \text{if } D'(E) \leq 1/2\end{aligned}$$

IF $D'(E) > 1/2$, then $D'(E^c) \leq 1/2$ so we have

$$\begin{aligned}d_{TV}(D, D') &= \sup_{E\subseteq\mathcal{O}} |D(E) - D'(E)| \\ &= \sup_{E\subseteq\mathcal{O}} |1 - D(E^c) - (1 - D'(E^c))| \\ &= \sup_{E\subseteq\mathcal{O}} |D(E^c) - D'(E^c))| \\ &\leq \sup_{E\subseteq\mathcal{O}} |e^{d_\diamond(D,D')}D'(E^c) - D'(E^c)| \\ &= \sup_{E\subseteq\mathcal{O}} |D'(E^c)(e^{d_\diamond(D,D')} - 1)| \\ &\leq \tfrac{1}{2}(e^{d_\diamond(D,D')} - 1)\end{aligned}$$

□