CS839: Differential Privacy and Learning February 10, 2026

7: Exponential Mechanism and Inverse Sensitivity

Instructor: Gavin Brown *Scribe:* Angus He

Disclaimer: This document is intended as an informal supplement to in-class note-taking. It has not been given the level of scrutiny expected in polished lecture notes, let alone that reserved for peer-reviewed publications.

1 Exponential Mechanism

1.1 Recap

- Data $x_1, \dots, x_n \in X^n$
- Candidates $y \in Y$
- Utility $u(y; x)$
- Sensitivity: $\forall x \sim x', y$, $|u(y; x) - u(y; x')| \leq \Delta$
- Release $p_x(y) \propto \exp\left(\frac{\varepsilon}{2\Delta} u(y; x)\right)$
- This is ε-DP

1.2 Utility Analysis

We rank all possible y's according to their utility $u(y; x)$. This ordering helps us focus on the probability of choosing a bad outcome.

Formally speaking, we order y's by decreasing utility as such:

$$\left.\begin{aligned} y_1 &= y^* = \underset{y}{\text{argmax}} u(y; x) \\ y_2 &= \\ y_3 &= \end{aligned}\right\} \text{Good, High Weight}$$

$$\left.\begin{aligned} y_4 &= \\ y_5 &= \end{aligned}\right\} \text{Bad, Low Weight}$$

Intuitively, we want to make sure that higher weight outcomes get chose with higher probability and vice versa. The Exponential Mechanism does exactly that: it makes sure that higher weight outcomes get chosen with exponentially higher probability, while lower weight outcomes get chosen with exponentially lower probability.

Next, we formally prove the probability bound of Exponential Mechanism.

Claim 1.1. *Given $|Y| < \infty$. Let $p_x \propto \exp((\varepsilon/2\Delta) \cdot u(y;x))$ denote the exponential mechanism with utility u and sensitivity Δ. Draw $\tilde{Y} \sim p_x$. $\forall t > 0$, we have*

$$\Pr\left[u(\tilde{Y};x) < u(y^*,x) - \frac{2\Delta}{\varepsilon}(t + \log|Y|)\right] \leq e^{-t}$$

Proof. We will assume $\Delta = 1$ and $u(y^*,x) = 0$ (after finishing the proof, it will be easy to see that this doesn't change the calculation). Define

$$BAD = \left\{y \in Y : u(y;x) < -\frac{2}{\varepsilon}(t + \log|Y|)\right\}$$

Then, we have

$$\Pr[\tilde{y} \in BAD] = \frac{\sum_{y \in BAD} \exp\left(\frac{\varepsilon}{2}u(y;x)\right)}{\sum_{y' \in BAD} \exp\left(\frac{\varepsilon}{2}u(y';x)\right)}$$

This equation is hard to bound: you would need to know the exact utility values for every single element in the set Y. Thus, we use another approach to bound the equation. Since $\Pr[y = y^*] \leq 1$, it follows that:

$$\begin{aligned}\Pr[\tilde{y} \in BAD] &\leq \frac{\Pr[\tilde{y} \in BAD]}{\Pr[\tilde{y} = y^*]} \\ &= \frac{\sum_{y \in BAD} \exp\left(\frac{\varepsilon}{2}u(y;x)\right)}{\exp\left(\frac{\varepsilon}{2}u(y^*;x)\right)} \\ &\leq |BAD| \exp\left(\frac{\varepsilon}{2}\left(\frac{-2}{\varepsilon}(t + \log|Y|)\right)\right) \\ &= |BAD| \cdot \frac{1}{|Y|}e^{-t} \leq e^{-t}\end{aligned}$$

□

Definition 1.2. Median Estimation: Given dataset $x_1 \leq x_2 \leq \cdots \leq x_n \in [0,R]$, where R is known a priori, and the output space is $Y = [0,R]$. The goal is to privately estimate the median of the dataset x.

Exponential Mechanism is suitable for solving median estimation because median is not "continuous". This is because changing a single datapoint might make the median jump to another value discretely. Also, the median has high global sensitivity (R in this setting). Thus, the Laplace Mechanism or Gaussian Mechanism become less suitable.

Remark 1.3. Our first idea is to set the utility function $u(y;x) = -|\text{med}(x) - y|$. The sensitivity is: $\forall x \sim x', y$, $|u(y;x) - u(y;x')| \leq R$. Then, we do

$$p_x(y) \propto \exp\left(\frac{-\varepsilon}{2R} \cdot |\text{med}(x) - y|\right) \approx \text{med}(x) + \text{Lap}\left(\frac{R}{\varepsilon}\right)$$

The main issues with this approach is that the sensitivity of median can be very large (even approaches infinity). Large Laplace noise means poorer accuracy, meaning that it will fail to generate outputs that are meaningful.

Remark 1.4. Our second idea is to set the utility function to $u(y;x) = -n\left|\frac{1}{2} - \hat{F}_x(y)\right|$, where the empirical CDF $\hat{F}_x(y)$ is defined as

$$\hat{F}_x(y) = \frac{\#\left\{i \in [n]; x_i \leq y\right\}}{n}$$

This method is better because the sensitivity is bounded. With this method, changing a single data point only changes the function by $\left|\frac{1}{n}\right|$.

Next, we prove the high-probability bound on the error of the private median estimation. Assume the dataset is ordered: $x_1 \leq x_2 \leq \cdots \leq x_n$. Notation like $x_{n(1/2+\alpha)}$ refers to the m-th example for $m = \frac{n}{2} + \alpha n$ (assuming this is an integer; we could round it). Let's define $W_{x,\alpha} := |x_{n(1/2-\alpha)} - x_{n(1/2+\alpha)}|$ to denote the width of the $\frac{1}{2} \pm \alpha$ quantiles on dataset x. This is a kind of measure of scale of the data around the median. For example, for data drawn i.i.d. from the Gaussian $\mathcal{N}(\mu, \sigma^2)$, for large n and small α we expect $W_{x,\alpha} \approx \sigma\alpha$.

Proof. Attempt 1:

$$\Pr\left[\left|\frac{1}{2} - \hat{F}_x(\tilde{y})\right| > \alpha\right] \leq \frac{\Pr\left[\left|\frac{1}{2} - \hat{F}_x(\tilde{y})\right| > \alpha\right]}{\Pr\left[\tilde{F}_x(\tilde{y}) = \frac{1}{2}\right]}$$

This will not work because $\Pr\left[\tilde{F}_x(\tilde{y}) = \frac{1}{2}\right] = 0$.

Attempt 2:

$$\begin{aligned}
\Pr\left[\left|\frac{1}{2} - \hat{F}_x(\tilde{y})\right| > \alpha\right] &\leq \frac{\Pr\left[\left|\frac{1}{2} - \hat{F}_x(\tilde{y})\right| > \alpha\right]}{\Pr\left[\left|\frac{1}{2} - \hat{F}_x(\tilde{y})\right| \leq \alpha\right]} \\
&\leq \frac{\int_{BAD} \exp\left(\frac{-\varepsilon}{2} u(y;x)\right) \mathrm{dy}}{\int_{\overline{BAD}} \exp\left(\frac{-\varepsilon}{2} u(y;x)\right) \mathrm{dy}} \\
&\leq \frac{\mathrm{Vol}(BAD) \cdot \exp\left(\frac{-\varepsilon n}{2}\alpha\right)}{\left|x_{n(1/2-\alpha)} - x_{n(1/2+\alpha)}\right| \cdot \exp\left(\frac{-\varepsilon n}{2}\alpha\right)} \\
&= \frac{\mathrm{Vol}(BAD)}{W_{x,\alpha}}
\end{aligned}$$

This is better, as the upper bound is not ∞. However, problem occurs because the term $\exp\left(\frac{-\varepsilon n}{2}\alpha\right)$ cancels out, so the probability doesn't go down as n increases. Worse, unless the points around the median are extremely spread out, we will have $\left|x_{n(1/2-\alpha)} - x_{n(1/2+\alpha)}\right| \ll R$ and this upper bound on the probability will be vacuous.

Attempt 3: This one will work, and represents a common approach when the output space is infinite.[1]

$$
\begin{aligned}
Pr\left[\left|\frac{1}{2} - \hat{F}_x(\tilde{y})\right| > \alpha\right] &\leq \frac{\Pr\left[\left|\frac{1}{2} - \hat{F}_x(\tilde{y})\right| > \alpha\right]}{\Pr\left[\left|\frac{1}{2} - \hat{F}_x(\tilde{y})\right| \leq \frac{\alpha}{2}\right]} \\
&\leq \frac{R \cdot \exp\left(\frac{-\varepsilon n}{2}\alpha\right)}{W_{x,\frac{\alpha}{2}} \cdot \exp\left(\frac{-\varepsilon n}{4}\alpha\right)} \\
&= \exp\left(\log\left(\frac{R}{W_{x,\frac{\alpha}{2}}}\right) - \frac{\varepsilon\alpha n}{4}\right) \\
&\leq \beta
\end{aligned}
$$

Thus we need $n \gtrsim \frac{\log(R/\beta W_{x,\alpha/2})}{\varepsilon\alpha}$ samples to output an approximate median. If the width around the median is not extremely small, this is quite sample-efficient. □

1.3 Inverse Sensitivity Mechanism

The core idea of Inverse Sensitivity Mechanism is that instead of measuring the sensitivity, we measure how many datapoints would need to change to make the output becomes y. We run the exponential mechanism with a special utility function.

Definition 1.5. Let function $f : X^n \to Y$. Define

$$\operatorname{len}_f(y; x) = \min_{x'} \left\{d(x, x') : f(x') = y\right\}$$

Where d is the Hamming distance.

If $f(x) = y$, $\operatorname{len}_f(y; x) = 0$. If y is very different from the output, many changes to the dataset are required.

The mechanism samples

$$p_x(y) \propto \exp\left(-\frac{\varepsilon}{2}\operatorname{len}_f(y; x)\right)$$

Claim 1.6. *This is ε-DP.*

Proof. For any $x \sim x'$ and y, we have $|\operatorname{len}_f(y; x) - \operatorname{len}_f(y; x')| \leq 1$, because we define adjacency of x and x' using the Hamming distance. □

[1] If you know more about the problem and can prove upper and/or lower bounds on the volume of level sets at various utilities, you can sometimes do better and shave log factors. See e.g., Bassily et al. [2014], Theorem 3.2 and Brown et al. [2021], Lemma 3.10.

Intuitively, the inverse sensitivity method works well on median because medium is "unstable", but the number of points needed to move to median is "stable".

Claim 1.7 (Informal)**.** *The Inverse Sensitivity mechanism is highly accurate.*

We will see (on Homework 2 and later in class) that the Inverse Sensitivity Mechanism is, in many settings, the most accurate differentially private algorithm one can hope for. An interesting line of (ongoing) work is related to understanding in what situations we can hope to do better.

Another important set of questions relates to computation: when can we implement it efficiently? Note that there are multiple immediate problems:

- Is it possible to even *compute* $\mathrm{len}_f(y; x)$ for a given x, y pair? If the space of datasets is infinite, it might not be clear how do so.
- Can we efficiently compute $\mathrm{len}_f(y; x)$?
- Can we efficiently sample from the resulting distribution?

Bibliographic Notes

The exponential mechanism was introduced by McSherry and Talwar [2007] and is a workhorse of differentially private algorithms. Every textbook and class notes will cover it.

The inverse sensitivity mechanism is showed up in a few applications as a kind of "folklore," but was brought to prominence and studied systematically by Asi and Duchi [2020]. It plays a key role in recent highlights of the literature on DP statistics, including the robustness-to-privacy transformation [Hopkins et al., 2022, Asi et al., 2023] and privacy wrappers [Linder et al., 2025].

References

Hilal Asi and John C Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in neural information processing systems*, 33:14106–14117, 2020.

Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. In *International Conference on Machine Learning*, pages 1121–1146. PMLR, 2023.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems*, 34:7950–7964, 2021.

Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. *arXiv preprint arXiv:2212.05015*, 2022.

Ephraim Linder, Sofya Raskhodnikova, Adam Smith, and Thomas Steinke. Privately evaluating untrusted black-box functions. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, page 2350–2361, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715105. doi: 10.1145/3717823.3718247. URL https://doi.org/10.1145/3717823.3718247.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.