Lecture Notes on Differential Privacy

Lectured by Jordan Awan
Scribed and edited by Yu-Wei (Gary) Chen and Jihun (Jimmy) Hwang

Last updated: July 23, 2025

Course Description

Differential privacy (DP) has emerged as the leading framework for formal privacy protection and is widely adopted by tech companies such as Apple, Google, and Microsoft, as well as by the US Census Bureau. DP introduces randomness into statistical procedures to obscure the contribution of any individual in a dataset, making it difficult to identify specific information. Key questions in the field of privacy include: (1) How should privacy be defined, and what properties should a privacy definition have? (2) How can algorithms be designed for different tasks to ensure privacy guarantees are met? (3) Once a randomized, privacy-preserving statistic is produced, how can we incorporate this randomness to perform valid statistical inference?

In this course, we will explore differential privacy and address each of these questions. We will begin by examining the essential properties an algorithm must possess to avoid being "blatantly non-private." We will then present differential privacy as a formal framework for privacy preservation. We will derive several key properties of the DP framework and develop general-purpose DP mechanisms. Additionally, we will study extensions and variations of differential privacy, including local DP, approximate DP, concentrated DP, Renyi DP, and f-DP. Lastly, we will consider methods for performing valid statistical inference on privatized data.

Contents

1 Introduction and Prerequisite

1.1 Motivation

In 2006, Netflix launched the Netflix Prize, an open competition to develop the best movie recommendation algorithm based on a user's past movie ratings and the dates those movies were watched. To protect privacy, all personally identifiable information (PII) was removed from the database. However, on IMDb, user ratings are publicly accessible and can serve as a unique identifier for individuals. To fully anonymize this data, one would need to remove all ratings and dates, which would effectively result in the complete destruction of the dataset.

Differential privacy (DP) is a guarantee made by the data holder or curator to the data subject, ensuring that their privacy is preserved regardless of the availability of other studies, datasets, or information sources. DP allows confidential data to be made available for analysis without requiring data usage agreements, data protection plans, or restricted access.

1.1.1 The Picture of Privacy

To formalize Differential Privacy, we consider the following cases:

- (Fundamental Law of Information Recovery) Overly accurate answers to too many questions (statistics) will destroy privacy in a speculated way.
  - The goal of DP is to postpone the inevitable loss of privacy as long as possible.
- (Paradox of Privacy) There are two extremes: perfect level of privacy and perfect accuracy. We want to learn nothing about an individual yet learn useful information about a population.

Example 1.1 (Insurance Premiums)**.** Suppose that a medical database shows that smoking causes cancer. Learning this, an insurance company may raise premiums for smokers.

Question 1.1. Has the smoker (that is not in the dataset) been harmed by the analysis?

Ans: Maybe yes, because his premium has become more expensive. On the other hand, it could help him reversely: realizing smoking is unhealthy, he may quit smoking .

Question 1.2. Has the smoker's privacy been violated?

Ans: Certainly more is known about the smoker than before, but his information was not leaked. But this is not going to be considered as a privacy violation in DP.

DP says that the privacy was not violated because the impact on the smokers is **the same whether or not he was in the study**. It is the **conclusion** from the study that affected the smoker, **not his presence or absence in the data set**. DP ensures that the same conclusions are reached whether or not any particular individual opts in or out of the dataset.

Example 1.2 (Target Online Shopping Recommendation)**.** A pregnant teenager started buying items such as prenatal vitamins. The platform now recommends similar or relevant items based on her purchase history. Viewing the items being recommended to the teenage girl, her parents could find out that she is pregnant.

1.1.2 Privacy-Preserving Data Analysis

There are some issues with privacy-preserving data analysis.

- Data cannot be fully anonymized **while** remaining useful.
  - The richer the data, the more useful it is. In general, more attributes, individuals, etc. lead to a more useful dataset.
  - In fact, one can identify an individual by just a few key features; for example: zip code, sex, date of birth, etc.
- Re-identification of 'anonymized' records is not the only risk.
  - If a record is tied to an individual, they could have compromising information. Instead, we could publish summary statistics of a dataset that do not merely pertain to an individual.
  - Simply knowing an individual is or is not in a data set could be harmful; for example, a teenage girl is categorized as pregnant or not.
- Questions on large sets are not safe.
  - Questions on large sets can be combined in **differential attack**; for example, suppose Mr. X is known to be in the database. By taking the difference of the following two questions, we could extract Mr. X's sickle-cell trait.
    1. What is the number of people with sickle-cell trait?
    2. What is the number of people with sickle-cell trait and are not named Mr. X?
- Query/Question/Statistic auditing is problematic.
  - Early ruling out questions that could compromise privacy is difficult, especially when it is the combination of pieces of information that causes privacy violation later on.
  - Refusing to answer a question may also cause private information leakage; for example, One refuses to say if he supports the new president.
- Summary statistics are not safe.
  - Differential attack can be carried out as we discussed above.

1.2 O-notations

The big-O and little-o notations help us reason about whether the amount of noise an estimator/statistic has is on the desirable order/rate. Let $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}^+$.

Definition 1.1 (Big-O)**.** $f(x) = O(g(x))$ if there exists $M > 0$ and $x_0 \in \mathbb{R}$ such that $|f(x)| \leq Mg(x)$ for all $x \geq x_0$.[1]

[1] In computer science, we are usually only interested in $f, g\colon \mathbb{N} \to \mathbb{R}^+$.

We think of $f(x) = O(g(x))$ meaning that **asymptotically** the rate of f is bounded by the rate of g, i.e. $|f| \leq g$ **asymptotically up to constants**.

Definition 1.2 (little-o)**.** $f(x) = o(g(x))$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$.

We think of $f(x) = o(g(x))$ as "roughly $f < g$ by much."

Definition 1.3 (Big-Ω)**.** $f(x) = \Omega(g(x))$ *iff* $g(x) = O(f(x))$. So, f is lower bounded asymptotically by g, i.e. $f \geq g$ **asymptotically**.

Definition 1.4 (Big-Θ)**.** $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $f(x) = \Omega(g(x))$, i.e. f is **bounded below and above** by g **asymptotically**.

Definition 1.5 (Little-ω)**.** $f(x) = \omega(g(x))$ *iff* $g(x) = o(f(x))$, i.e. $f > g$ by much.

Of these notations, O and o notations are the most important.

Remark 1.1 (Tips)**.** When working with o, O, Θ, Ω, and ω,

- If $f(x)$ is the sum of several terms and one of the terms has the largest growth, then others can be omitted.
- If $f(x)$ is the product of several factors, any constants (not depending on x) can be omitted.

Example 1.3.

1. $f(x) = 6x^4 - 2x^3 + 5$ then $f(x) = O(\underbrace{x^4}_{\text{simplest expression possible}})$.
2. $\log\log x = o(\log(x))$ because by L'Hospital's rule,

$$\lim_{x\to\infty} \frac{\log\log x}{\log x} = \lim_{x\to\infty} \frac{\frac{1}{x\log x}}{\frac{1}{x}} = \lim_{x\to\infty} \frac{1}{\log x} = 0$$

1.3 O in Probability Notation

When working with random variables, it is also helpful to have an "o" notation. Let $(x_n)_{n=1}^{\infty}$ be a sequence of r.v.s and $(a_n)_{n=1}^{\infty}$ be a sequence of positive real values (constants).

Definition 1.6 (Small o_p)**.** We write $X_n = o_p(a_n)$ to mean

$$\frac{X_n}{a_n} \xrightarrow{p} 0 \qquad \text{(convergence in probability)}$$

Equivalently, we can write $\frac{X_n}{a_n} = o_p(1)$; more precisely,

$$\begin{aligned} y_n = o_p(1) &\iff y_n \xrightarrow{p} 0 \\ &\iff \lim_{n\to\infty} \Pr[|y_n| \geq \varepsilon] = 0 \quad \text{for all } \varepsilon > 0 \end{aligned}$$

Definition 1.7 (Big O_p)**.** We write $X_n = O_p(a_n)$ as $n \to \infty$ to mean that $\dfrac{X_n}{a_n}$ is **stochastically bounded**, that is,

$$\forall \varepsilon > 0,\ \exists M > 0 \text{ and } N > 0 \text{ s.t. } \Pr\Big[\big|X_n/a_n\big| > M\Big] < \varepsilon \quad \text{for all } n > N$$

Example 1.4 (Theorem 14.4.1, Discrete Multivariate Analysis, Bishop et al. [2007])**.** If $(X_n)_{n=1}^{\infty}$ is a sequence of r.v.s each with finite variance, then

$$(X_n - \mathbb{E}[X_n]) = O_p\Big(\sqrt{\mathrm{Var}(X_n)}\Big)$$

Moreover, if (a_n) is a sequence such that $a_n^{-2}\,\mathrm{Var}(X_n) \to 0$, then by Chebyshev inequality,

$$\Pr\Big[a_n^{-1}\big|X_n - \mathbb{E}[X_n]\big| \geq \varepsilon\Big] \leq \frac{a_n^{-2}\mathrm{Var}(X_n)}{\varepsilon^2} \to 0$$

Therefore, $a_n^{-1}(X_n - \mathbb{E}[X_n]) \xrightarrow{p} 0$, i.e.

$$X_n - \mathbb{E}[X_n] = o_p(a_n)$$

Definition 1.8. We will also write

$$\begin{aligned}
X_n = \Omega_p(a_n) &\iff \frac{a_n}{X_n} \text{ is } \textbf{stochastically bounded} \\
X_n = \omega_p(a_n) &\iff \frac{a_n}{X_n} \xrightarrow{p} 0 \\
X_n = \Theta_p(a_n) &\iff X_n = O_p(a_n) \text{ and } X_n = \Omega_p(a_n)
\end{aligned}$$

Theorem 1.1 (Central Limit Theorem)**.** Let (X_n) be a sequence of i.i.d. r.v.s with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$, then

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2\big).$$

In O_p notation,

$$\sqrt{n}(\overline{X}_n - \mu) = O_p(1)$$

since any distribution is **stochastically bounded**. This implies $\overline{X}_n - \mu = \underbrace{O_p\left(\frac{1}{\sqrt{n}}\right)}_{\text{statistical/sampling/estimation error}}$.

Remark 1.2. In fact, $\overline{X}_n - \mu = \Theta_p\left(\dfrac{1}{\sqrt{n}}\right)$.

Corollary 1.1 (Consistent Estimators)**.** Many estimators $\widehat{\theta}(X_n)$ for Θ (e.g. MLE) satisfy

$$\widehat{\theta}(X_n) - \theta = O_p\left(\frac{1}{\sqrt{n}}\right),$$

which is called an $\sqrt{n}$-**consistent** estimator.

Remark 1.3. In DP, we want the added noise to be at most $o_p(\frac{1}{\sqrt{n}})$ such that it does not interfere with the sampling error $O_p(\frac{1}{\sqrt{n}})$. This is the least acceptable rate of noise.

2 Blatant Non-Privacy

The main reference for this chapter is Dinur and Nissim [2003].

This was in 2003-2004 when people were coming up with notions of privacy. These notions were later used to develop DP. We start with the notion of "blatant non-privacy." The goal was to agree on something that is definitely a privacy violation.

Here are some motivating questions:

1. How inaccurate must responses be to not completely destroy privacy?
    - If we are able to recreate all records in a dataset, this is definitely not private.
2. How does that answer to 1. depend on the number of **queries**?

Example 2.1 (A simple abstraction)**.** Let us consider as follows:

- Each person has a single bit of information, either 0 or 1. Hence, the database $d = (d_1, \cdots, d_n)$ with $d_i \in \{0, 1\}$.
- An attacker picks a subset $S \subseteq \{1, 2, \cdots, n\}$ and asks **(queries)** how many 1's are in the rows S of d. Denote $A(S)$ be the correct answer.
- For an arbitrary privacy mechanism, call $r(S)$ the randomized response, which approximates $A(S)$ and the error of one query is $E(S, r(S)) = |A(S) - r(S)|$.

To achieve privacy, we would want the error to be non-trivial.

Definition 2.1 (Blatantly Non-Private)**.** A mechanism is **blatantly non-private** if for every possible database d, the adversary can construct a candidate database C that agrees with d on all but $o(n)$ **entries**, i.e. $\|C - d\|_0 = o(n)$[a]. That is, the proportion of correctly recovered entries goes to 1.

[a] Can be L^1-norm as well

If R_n is the number of correct entries, then $\frac{R_n}{n} = \frac{n - o(n)}{n} \to 1$. This is considered blatantly non-private. Therefore, avoiding blatantly non-private is a low bar for a privacy mechanism.

Remark 2.1. The adversary need not know which entries are correct or not. If they did know which entries were correct and which were incorrect, they could recover the full dataset by flipping the bits of the incorrect entries.

Theorem 2.1. Let r be a mechanism on a database d with distortion bounded by E, i.e. $|r(S) - A(S)| \leq E$ for all S and any runs of r, then there exists an adversary that can reconstruct d to within $4E$ positions (entries).

Proof of Theorem 2.1. The attack is as follows:

1. (Estimate the number of 1's on all subsets and can get $r(S)$ for all $S \subseteq [n] := \{1, ..., n\}$)
2. (Rule out all "distant" or incompatible datasets)

   For each candidate $C \in \{0,1\}^n$, if there exists $S \subseteq [n]$ such that $\left|\sum_{i \in S} C_i - r(S)\right| > E$, then rule out C. Output the first C that is not ruled out.

Note that d can never be ruled out, so the attack will output some C. We now argue that C and d differ by at most $4E$ entries.

Call $I_0 = \{i \mid d_i = 0\}$ and $I_1 = \{i \mid d_i = 1\}$. By construction of C, consider the error

$$\left| r(I_0) - \sum_{i \in I_0} C_i \right| \leq E,$$

and by assumption

$$\left| r(I_0) - \sum_{i \in I_0} d_i \right| \leq E.$$

By triangle inequality, C and d differ in at most $2E$ entries among I_0 coordinates. Similarly, C and d differ in at most $2E$ entries on I_1. Therefore, C and d differ at most $4E$ entries. ■

Proposition 2.1. To avoid blatant non-privacy, error (with all 2^n queries) E must be $\boxed{\Omega_p(n)}$.

Question 2.1. What if we limit the attack to $O(n)$ queries?

If we view the database as a random sample, the each query $A(S)$ is a binomial r.v. $A(s) \overset{d}{\sim} \text{Binom}(n, p)$ for some p, then

$$A(s) = np + \underbrace{\Theta_p(\sqrt{n})}_{\text{sampling error}}$$

To have a reasonably accurate output $r(S)$, we would like (best-case) the mechanism error to be smaller, i.e. $\left|A(S) - r(S)\right| = o_p(\sqrt{n})$. The next result shows when the error is $o(\sqrt{n})$ and the attacker asks $O(n)$ queries, we still **cannot avoid** blatant non-privacy [Dinur and Nissim, 2003].

Theorem 2.2. Let $\eta > 0$ and at least $\frac{1}{2} + \eta$ questions have error less than $\alpha(n) > 0$. Then there is an attack using $O(n)$ questions to reconstruct the database in all but $\left(\frac{2\alpha(n)}{\eta}\right)^2$ entries.

Basically, it is saying that there is an attack using $O(n)$ questions with error at most $(2\alpha(n)/\eta)^2$.

Example 2.2. If $\alpha(n) = o(\sqrt{n})$ in Theorem 2.2, then the number of wrong positions is

$$\left(\frac{2\alpha(n)}{\eta}\right)^2 = \frac{4(o(\sqrt{n}))^2}{\eta^2} = o(n)$$

and we have blatant non-privacy.

This suggests that we need $\Omega(\sqrt{n})$ noise.

Remark 2.2 (Takeaway)**.** Here are some conclusions:

- Significant amount of noise must be introduced to protect privacy.
- The more queries are answered, the more noise we need.

Some DP results to shed light on how much error should be added:

1. For ε-DP and q queries, noise is $\Theta_p(q)$, which is much more that the error in Theorem 2.2.
2. For μ-GDP and q queries, noise is $\Theta_p(\sqrt{q})$.
3. If $q = o(\sqrt{n})$ in ε-DP or $q = o(n)$ in μ-GDP, we can get away with $o_p(\sqrt{n})$ noise.

3 Background and Fundamentals of DP

The main reference for this lecture is Dwork and Roth [2014].

We model the database as $X \in \mathcal{X}^n$ where each "row" (entry of the list) corresponds to an individual and $\mathcal{X}$ represents the set of possible contributions from one person. For now, assume n is public.

Definition 3.1. A privacy mechanism $\mathcal{M}$ is **a set of distributions/probability measures** M_X on a measurable space $(\mathcal{Y}, f)$ indexed by $\mathcal{X}^n$, i.e.

$$\mathcal{M} = \{M_X \mid X \in \mathcal{X}^n\}.$$

For brevity, we write $\mathcal{M} \colon \mathcal{X}^n \to \mathcal{Y}$ or $M \colon \mathcal{X}^n \to \mathcal{Y}$. We may also write $M(X)$ to denote a random variable $\overset{d}{\sim} M_X$.

$$\left.\begin{aligned} X &\twoheadrightarrow \boxed{\mathcal{M}} \twoheadrightarrow M(X) \overset{d}{\sim} M_X \\ X' &\twoheadrightarrow \boxed{\mathcal{M}} \twoheadrightarrow M(X') \overset{d}{\sim} M_{X'} \end{aligned}\right\} \text{approximately indistinguishable}$$

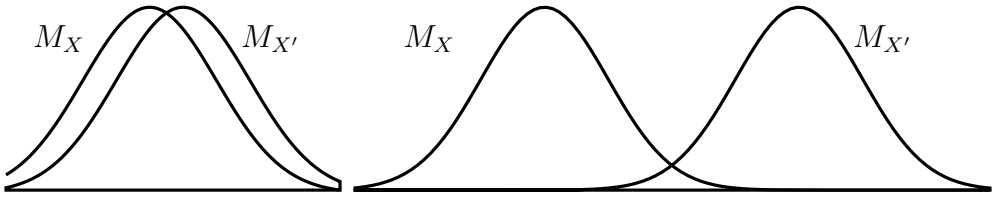


Figure 1: Left: (approximately) indistinguishable; right: distinguishable

3.1 The Idea Behind Differential Privacy (DP)

If $X, X' \in \mathcal{X}^n$ are databases differing in one person's contribution, then the distributions M_X and $M_{X'}$ should be **close**. Back tracking, the adversary will have a hard time discerning where the input came from. Intuitively, this means that observing an outcome from either M_X or $M_{X'}$, it is difficult to determine whether the input was X or X'. In other words, M_X and $M_{X'}$ are approximately indistinguishable[2].

Definition 3.2 (Hamming Distance)**.** $H \colon \mathcal{X}^n \times \mathcal{X}^n \to \mathbb{Z}^{\geq 0}$ is given by

$$H(X, X') = \#\{i \mid X_i \neq X_i'\}$$

and counts the number of rows (entries) where X_i and X_i' disagree.

[2] Imagine two normal distributions with slightly different averages.

So, X and X' **differ in one person's contribution** exactly when $H(X, X') \leq 1$. We say that X and X' are **neighboring databases** or **adjacent** (or "differing in one entry").

Example 3.1. The output of a privacy mechanism can be on any measurable space, and can be designed for any purpose. For example, $\mathcal{M}$ could be designed to give a private version of some statistics or query.

(1) $\mathcal{M}$ gives an approximate/noisy count for some property.

(2) Noisy estimate of mean/variance/sufficient statistic.

(3) Approximate regression coefficient or ML parameters.

$\mathcal{M}$ could also be designed to output a private synthetic dataset, which has similar properties as the original dataset. While we will focus on real or vector valued statistics, $\mathcal{M}$ could also be designed to produce outputs in more complex spaces such as densities or functional regression estimates.

Definition 3.3 (Differential Privacy)**.** Let $\varepsilon \geq 0$ and $\delta \in [0, 1]$. A privacy mechanism $\{M_X \mid X \in \mathcal{X}^n\}$ on $(\mathcal{Y}, \rho)$ is (ε, δ)-Differentially Private ((ε, δ)-DP) if for all measurable sets $S \in \rho$, and all $H(X, X') \leq 1$,

$$M_X(S) \leq e^{\varepsilon} M_{X'}(S) + \delta$$

Or alternatively, $\boxed{\Pr[M_X \in S] \leq e^{\varepsilon} \Pr[M_{X'} \in S] + \delta}$.

Remark 3.1 (Important Special Cases)**.**

- If $\delta = 0$, $\mathcal{M}$ is called ε-DP or **pure DP**. In contrast, (ε, δ)-DP is also called approximate DP.
- If $\varepsilon = 0$, then δ simply becomes a total variation distance. And if $\delta = 0$ as well then we get a **perfect privacy**.
- If $\varepsilon = \infty$, then there is **no privacy**.

Note that the probability is only over the randomness in $\mathcal{M}$. DP does not assume a probability distribution for X.

3.2 Randomized Response (RR)

Suppose that there is a survey asking a yes-or-no question "*Have you ever* _____ ?" and survey takers are asked to follow the following instructions:

$$\text{Your Answer: } \begin{cases} \text{Truth} & \text{w.p. } 1 - p \\ \text{Yes} & \text{w.p. } p/2 \\ \text{No} & \text{w.p. } p/2 \end{cases}$$

Theorem 3.1. Randomized Response (RR) satisfies $\boxed{\varepsilon\text{-DP}}$ where $\varepsilon =$ (TBD in proof).

Proof of Theorem 3.1. Fix any individual, whose conditional probabilities are

$$\Pr[\text{YES} \mid \text{YES}] = 1 - p + \frac{p}{2} = 1 - \frac{p}{2} = \Pr[\text{NO} \mid \text{NO}]$$
$$\Pr[\text{YES} \mid \text{NO}] = \frac{p}{2} = \Pr[\text{NO} \mid \text{YES}]$$

($\Pr[A \mid B]$ here denotes the probability of an individual says A given/when their truth is B.)

Then, after a massive cancellation for other individuals' ratio, the ratios to be checked are

$$\frac{\Pr[\text{YES} \mid \text{YES}]}{\Pr[\text{YES} \mid \text{NO}]} = \frac{1 - p/2}{p/2} = \frac{2}{p} - 1 = \frac{2-p}{p}$$

and

$$\frac{\Pr[\text{NO} \mid \text{NO}]}{\Pr[\text{NO} \mid \text{YES}]} = \frac{1 - p/2}{p/2} = \frac{2-p}{p} = \frac{\Pr[\text{YES} \mid \text{YES}]}{\Pr[\text{YES} \mid \text{NO}]}$$

for all individuals. Hence, as we set $\varepsilon = \log\left(\frac{2-p}{p}\right)$, RR satisfies ε-DP. ■

Remark 3.2. Consider different values of p,

- If $p = 1$, then it is the same as flipping a coin. $\varepsilon = \log\left(\frac{2-1}{1}\right) = \log(1) = 0$ so we achieve 0-DP, which is perfect privacy.
- If $p = 0$, then we always respond with truth, so we have ∞-DP, i.e. no privacy.
- In general, to achieve ε-DP, we set p to be $e^{\varepsilon} = \frac{2-p}{p} \implies p = \frac{2}{1+e^{\varepsilon}}$ and this gives ε-DP.

Example 3.2. Let θ be the true population proportion of participants with the property. The outcome of an individual answering "YES", denoted by Y, from RR, has expectation:

$$\begin{aligned}\mathbb{E}[Y] &= \underbrace{\Pr[\text{YES} \mid \text{YES}]}_{\text{mechanism}} \underbrace{\Pr[\text{Truth} = \text{YES}]}_{\text{model}} + \Pr[\text{YES} \mid \text{NO}] \Pr[\text{Truth} = \text{NO}] \\ &= \left(1 - \frac{p}{2}\right)\theta + \frac{p}{2}(1-\theta) = \frac{p}{2} + (1-p)\theta,\end{aligned}$$

which is biased but linear in θ. Encode $1 = \text{YES}$ and $0 = \text{NO}$, and let $T \sim \text{Bern}(\theta)$ be the truth, then

$$(Y \mid T) \sim \text{Bern}\left(\left(1 - \frac{p}{2}\right)T + \frac{p}{2}(1-T)\right).$$

It is direct to construct $\frac{Y-p/2}{1-p}$ as an unbiased estimate of θ with variance

$$\begin{aligned}
\mathrm{Var}\left(\frac{Y-p/2}{1-p}\right) &= \frac{1}{(1-p)^2}\mathrm{Var}(Y) = \frac{1}{(1-p)^2}\left[\mathbb{E}\Big[\mathrm{Var}\big(Y \mid T\big)\Big] + \mathrm{Var}\Big(\mathbb{E}\big[Y \mid T\big]\Big)\right] \\
&= \frac{1}{(1-p)^2}\left[\mathbb{E}_{T\sim\mathrm{Bern}(\theta)}\left[\frac{p}{2}\left(1-\frac{p}{2}\right)\right] + \mathrm{Var}\left(\frac{p}{2}+(1-p)T\right)\right] \\
&= \frac{1}{(1-p)^2}\left[\frac{p}{2}\left(1-\frac{p}{2}\right) + (1-p)^2\theta(1-\theta)\right] \\
&= \underbrace{\frac{\frac{p}{2}\left(1-\frac{p}{2}\right)}{(1-p)^2}}_{\text{extra error from privacy}} + \underbrace{\theta(1-\theta)}_{\text{variance without privacy}}
\end{aligned}$$

by the law of total variance.

Suppose our dataset has n individuals and RR gives $(Y_1,\ldots,Y_n)$. Then, $\widehat{\theta} = \dfrac{\overline{Y}-p/2}{1-p}$ is an unbiased estimator for θ where $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. Hence,

$$\begin{aligned}
\mathrm{Var}\Big(\widehat{\theta}\Big) &= \frac{1}{n}\mathrm{Var}\left(\frac{y_i-p/2}{1-p}\right) \\
&= \frac{1}{n}\left(\frac{p/2\ (1-p/2)}{(1-p)^2} + \theta(1-\theta)\right) \\
&= \frac{1}{n}\left(\frac{e^\varepsilon}{(e^\varepsilon-1)^2} + \theta(1-\theta)\right) \\
&\approx \frac{1}{n}\left(\frac{1}{\varepsilon^2} + \theta(1-\theta)\right) \text{ as } \varepsilon\to 0
\end{aligned}$$

On the positive side, we have a $\sqrt{n}$-consistent estimator for θ, but the asymptotic variance is inflated. This is because the error introduced is $\Theta_p\left(\dfrac{1}{\sqrt{n}}\right)$—the same rate as the statistical estimation error.

In fact, RR satisfies a stronger notion of privacy than ε-DP, called local-DP, where even the data collector does not learn the responses. Note that all local ε-DP are central ε-DP. As we will discuss later, it is possible to design a better (central) DP mechanism to estimate $\widehat{\theta}$ with $O\big(\frac{1}{n}\big)$ error. A popular option is the Laplace mechanism.

Here is a new motivating question:

Question 3.1. How do we show a mechanism satisfies ε-DP?

We need to show that

$$\Pr[M_X \in S] \le e^\varepsilon \Pr[M_{X'} \in S]$$

for all adjacent X and X', and all measurable sets S. The key lemma is as follows.

Lemma 1. Suppose that M_X is a continuous (discrete) distribution for all X, with PDF f_X, then if $f_X(t) \leq e^{\varepsilon} f_{X'}(t)$ for all t then $\mathcal{M}$ satisfies ε-DP.

Proof of Lemma 1. Let S be a measurable set, then

$$\Pr[M_X \in S] = \int_S f_X(t)dt \leq e^{\varepsilon} \int_S f_{X'}(t)dt \leq e^{\varepsilon} \Pr[M_{X'} \in S]$$

For discrete cases, replace integrals with sums. ■

Remark 3.3 (Inverse of Lemma 1)**.** If the inequality holds for a.e. t instead of for all t, then the inverse of Lemma 1 holds as well.

3.3 Laplace Distribution and Mechanism

Most statistics/queries of interests are real-vector valued, i.e. $f\colon \mathcal{X}^n \to \mathbb{R}^k$. We can add (independent) noise to f to make it DP, but the noise must be scaled to the sensitivity of f; **how much** can f change when going from X to X' adjacent? And how do we **quantify** that?

Definition 3.4 (ℓ_1-sensitivity)**.** The ℓ_1-sensitivity of a function $f\colon \mathcal{X}^n \to \mathbb{R}^k$ is

$$\Delta f = \sup_{H(X,X')\leq 1} \left\| f(X') - f(X) \right\|_1 = \sup_{H(X,X')\leq 1} \sum_{i=1}^{k} |f_i(X) - f_i(X')|$$

Definition 3.5 (Laplace Distribution)**.** The **Laplace distribution** $\mathrm{Lap}(m, s)$ is a continuous real valued r.v. with density $\frac{1}{2s} \exp\left(-\frac{|x-m|}{s}\right)$, where m is the location and s is the scale.

Remark 3.4.

- $\mathrm{Lap}(m, s)$ is also known as the double exponential distribution with mean m and variance $2s^2$.
- If $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\lambda)$ where λ is called the rate parameter, then $(X_1 - X_2) \sim \mathrm{Lap}\left(0, \frac{1}{\lambda}\right)$.
- If $Y \sim \mathrm{Lap}(0, s)$ then $|Y| \sim \mathrm{Exp}\big(1/s\big)$.

Theorem 3.2 (Laplace Mechanism)**.** Let $\varepsilon > 0$. Given $f\colon \mathcal{X}^n \to \mathbb{R}^k$ with ℓ_1-sensitivity Δf, the Laplace mechanism is

$$M_X = f(X) + (L_1, \cdots, L_k)^{\top}$$

where $L_i \overset{\text{i.i.d.}}{\sim} \mathrm{Lap}\left(0, \frac{\Delta f}{\varepsilon}\right)$, then M satisfies $(\varepsilon, 0)$-DP.

Proof of Theorem 3.2. Let $H(X, X') \leq 1$. Call p_X the PDF of M_X and $p_{X'}$ the PDF of $M_{X'}$. Let $t \in \mathbb{R}^k$, then

$$\begin{aligned}
\frac{p_X(t)}{p_{X'}(t)} &= \prod_{i=1}^{k} \frac{\exp\left(\frac{-\varepsilon}{\Delta f}|f_i(X) - t_i|\right)}{\exp\left(\frac{-\varepsilon}{\Delta f}|f_i(X') - t_i|\right)} = \prod_{i=1}^{k} \exp\left(\frac{\varepsilon}{\Delta f}\left(|f_i(X') - t_i| - |f_i(X) - t_i|\right)\right) \\
&\leq \prod_{i=1}^{k} \exp\left(\frac{\varepsilon}{\Delta f}|f_i(X) - f_i(X')|\right) \qquad (\because \text{Triangle inequality}) \\
&= \exp\left(\frac{\varepsilon}{\Delta f}\left\|f(X) - f(X')\right\|_1\right) \\
&\leq \exp\left(\frac{\varepsilon}{\Delta f}\Delta f\right) = \exp(\varepsilon) \qquad (\because \text{Supremum in Definition 3.4})
\end{aligned}$$

■

Remark 3.5. Note that Laplace mechanism adds noise proportional to the sensitivity and inversely proportional to ε.

- More sensitive requires more noise.
- More privacy (smaller ε) requires more noise.

Example 3.3 (Counting Query)**.** How many elements in the database have property P?

- This type of query is simple but commonly used.
- The sensitivity of any counting query is 1. Hence, ε-DP can be achieved by adding $\mathrm{Lap}(1/\varepsilon)$.
- If we have m counting queries, upper bound on ℓ_1-sensitivity of the vector is m. So to get ε-DP, we add i.i.d. $\mathrm{Lap}(m/\varepsilon)$.

Example 3.4 (Histogram Query)**.** Sometimes, the counting queries are structurally disjoint. Consider the following question: Are you currently a resident of (Alaska, Alabama, Arizona, $\cdots$, Wyoming)?

Each person appears in only one count. Changing their value can affect at most **two queries**, so ℓ_1-sensitivity is 2, no matter how many queries there are! So we add $\mathrm{Lap}(2/\varepsilon)$ to get ε-DP.

Remark 3.6. The 2020 Census consists of several histograms and counting queries, and they add noise to achieve DP.

Example 3.5 (Cf. Example 3.2)**.** Everyone has a $1/0$ true answer to a question, modeled as $\mathrm{Bern}(p)$. We want to estimate the true p. With RR, our estimator had variance

$$\frac{p(1-p)}{n} + \frac{e^{\varepsilon}}{n(e^{\varepsilon}-1)^2} \approx \frac{p(1-p)}{n} + \frac{1}{n\varepsilon^2},$$

which has rate $\overline{X} + O_p\left(\frac{1}{\sqrt{n\varepsilon}}\right)$. Instead, we could use the Laplace mechanism:

$$\sum_{i=1}^{n} I(X_i = 1) + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right)$$

Dividing by n, the estimator $\overline{X} + \mathrm{Lap}\left(\frac{1}{n\varepsilon}\right)$ has variance

$$\frac{p(1-p)}{n} + \frac{2}{n^2\varepsilon^2},$$

which has rate $\overline{X} + O_p\left(\frac{1}{n\varepsilon}\right)$.

Remark 3.7 (Security without Obscurity)**.** The DP mechanism can be publicly known without affecting the privacy guarantee. This is important for statistical inference.

Example 3.6. Census swapping pre-DP vs. DP algorithm

3.4 Post-Processing

An analyst without additional knowledge of the database cannot make the output of a DP mechanism less private by applying some function.

Remark 3.8. The post-processing property comes from Data Processing Inequality in Information Theory, which says that for KL Divergence, Renyi Divergence, Total Variation...,

$$D(X\|Y) \geq D(f(X)\|f(Y)).$$

Proposition 3.1 (Post-Processing)**.** Let $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{Y}$ be a (ε, δ)-DP mechanism. Let $f\colon \mathcal{Y} \to \mathcal{Z}$ be a randomized or deterministic mapping, then $f \circ \mathcal{M}\colon \mathcal{X}^n \to \mathcal{Z}$ is (ε, δ)-DP.

Proof. We prove the result when f is deterministic. Fix $X, X' \in \mathcal{X}^n$ s.t. $H(X, X') \leq 1$, and fix a measurable set $S \subseteq \mathcal{Z}$. Let $T := \left\{y \in \mathcal{Y} \;\middle|\; f(y) \in S\right\} = f^{-1}(S)$. Then,

$$\begin{aligned}\Pr[f(M_X) \in S] &= \Pr[M_X \in T] \\ &\leq e^{\varepsilon} \Pr[M_{X'} \in T] + \delta \\ &= e^{\varepsilon} \Pr[f(M(y)) \in S] + \delta.\end{aligned}$$

The bound is tight when f^{-1} exists. ■

Post-processing is also useful in mechanism design. We can privatize summary statistics or an entire synthetic database and then process them to estimate specific quantities, perform hypothesis testing and construct confidence intervals, or conduct Bayesian inference.

3.5 Composition

When data contributes to two or more differentially private outputs, the privacy guarantee degrades in a controlled manner.

Theorem 3.3 (Composition)**.** Let $\mathcal{M}_i \colon \mathcal{X}^n \to \mathcal{Y}_i$ be $(\varepsilon_i, \delta_i)$-DP for $i = 1, \cdots, k$. Then if $\underline{\mathcal{M}} \colon \mathcal{X}^n \to \prod_{i=1}^k \mathcal{Y}_i$ is defined by

$$\underline{M}(X) = \big(M_1(X), M_2(X), \cdots, M_k(X)\big),$$

$\underline{\mathcal{M}}$ satisfies $\left(\sum_{i=1}^k \varepsilon_i, \sum_{i=1}^k \delta_i\right)$-DP.

Proof. We prove the case where $\delta = 0$ and $k = 2$. Let $X, X' \in \mathcal{X}^n$ and $H(X, X') \le 1$. Also let

$$\begin{aligned} &f_X^1 \text{ be the PDF/PMF of } M_1(X) \\ &f_X^2 \text{ be the PDF/PMF of } M_2(X) \\ &f_X^{12} \text{ be the/PMF PDF of } \underline{M}(X) \\ &\dots \text{ and similar for } X' \end{aligned}$$

Then it suffices to show that $\dfrac{f_X^{12}(t_1, t_2)}{f_{X'}^{12}(t_1, t_2)} \le e^{\varepsilon_1 + \varepsilon_2}$ for all $t_1 \in \mathcal{Y}_1, t_2 \in \mathcal{Y}_2$.

$$\frac{f_X^{12}(t_1, t_2)}{f_{X'}^{12}(t_1, t_2)} = \frac{f_X^1(t_1) f_X^2(t_2)}{f_{X'}^1(t_1) f_{X'}^2(t_2)} \le e^{\varepsilon_1} e^{\varepsilon_2} = e^{\varepsilon_1 + \varepsilon_2}$$

It is straightforward to extend to $k > 2$, but $\delta \neq 0$ requires more complicated argument. ■

While the result for δ guarantee is loose, there will be tighter composition bounds for (ε, δ)-DP-case in our later discussion.

3.6 Group Privacy

What if the two databases differ by a family of size k?

Proposition 3.2 (Group Privacy for Pure DP)**.** Any $(\varepsilon, 0)$-DP mechanism $\mathcal{M}$ is $(k\varepsilon, 0)$-DP for groups of size k. More precisely, for all $H(X, X') \le k$ and set of outputs $S \subseteq \mathrm{Range}(\mathcal{M})$,

$$\Pr[M_X \in S] \le e^{k\varepsilon} \Pr[M_{X'} \in S]$$

Proof of Proposition 3.2. Let $X_0, X_1, \cdots, X_k$ be a sequence s.t. $X_0 = X$ and $X_k = X'$, and $H(X_{i-1}, X_i) \le 1$ for all $i = 1, \cdots, k$. Then,

$$\Pr[M(X_0) \in S] \le e^{\varepsilon} \Pr[M(X_1) \in S] \le e^{2\varepsilon} \Pr[M(X_2) \in S] \le \cdots \le e^{k\varepsilon} \Pr[M(X_k) \in S]$$

Hence, M is $(k\varepsilon, 0)$-DP. ■

While $(\varepsilon, 0)$-DP has a strict lower bound, it inhibits flexibility in obtaining, for example, an unbiased estimator for the private mean, which can instead be achieved by (ε, δ)-DP. Note that there is a more complex group privacy result for (ε, δ)-DP.

We listed other DP properties as follows:

- **Protection Against Arbitrary Risks**

  Let S be any set of outputs of concern. Then $\Pr[M(x) \in S]$ can increase by at most a factor of e^{ε} when an individual joins or leaves the database.

- **Automatic Protection Against Linkage Attacks**

  DP provides protection regardless of the attacker's knowledge, whether it comes from past, present, or future data sources (databases).

- **Quantification of Privacy Loss**

  Instead of a binary safe/not safe classification, we have a continuous measure of the privacy guarantee. This allows us to ask:
  1. For a fixed privacy bound, which techniques provide better utility or accuracy?
  2. For a fixed level of accuracy, which techniques offer stronger privacy?
     [High privacy (with low utility) vs. High utility (with low privacy)]

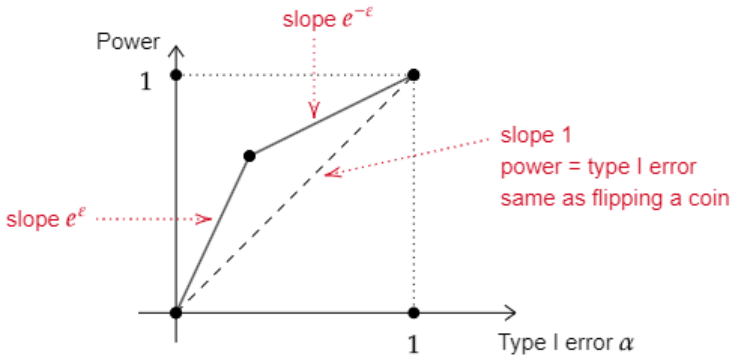
3.7 Connection to Hypothesis Testing

Suppose an adversary knows $n-1$ rows of a database (everything in the database except the last row), and believe that the last row is one of two options. Given a DP output, the adversary will still struggle to determine the last row.

Theorem 3.4. Let $X, X' \in \mathcal{X}^n$ s.t. $H(X, X') \leq 1$. Let $M\colon \mathcal{X}^n \to \mathcal{Y}$ be a privacy mechanism. Consider the hypothesis test

$$H_0\colon X \text{ versus } H_1\colon X'$$

based on the output of M, at type I error α. Then M satisfies (ε, δ)-DP if and only if the power is bounded above by $\min\{e^{\varepsilon}\alpha + \delta, e^{-\varepsilon}(\alpha - 1 + \delta) + 1\}$.



Proof of Theorem 3.4. We prove the $\implies$ direction with a deterministic rejection rule. Let R be any rejection set, then

$$\underbrace{\Pr[M(X) \in R]}_{\substack{\text{Prob. we reject} \\ \text{when } X \text{ is true}}} \leq \alpha \text{ and } \Pr[M(X) \in R^c] \geq 1 - \alpha$$

Let the power $\beta = \Pr[M(X') \in R]$, then (ε, δ)-DP implies

$$\beta = \Pr[M(X') \in R] \leq e^{\varepsilon} \Pr[M(X) \in R] + \delta \leq e^{\varepsilon}\alpha + \delta$$

which gives us the first upper bound. We also have

$$1 - \alpha = \Pr[M(X) \in R^c] \leq e^{\varepsilon} \Pr[M(X') \in R^c] + \delta = e^{\varepsilon}(1 - \beta) + \delta,$$

which gives $\beta \leq e^{-\varepsilon}(\alpha - 1 + \delta) + 1$. ■

3.8 Differentially Private Linear Regression

To perform linear regression on privatized data, for example, we need to leverage the above properties.

Given dataset $D = (X, y)$ with $d_i = (X_i, y_i) \in \mathbb{R}^{p+1}$ such that $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$. We model $y_i = X_i\beta + \varepsilon_i$ where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ and $\beta \in \mathbb{R}^p$, and our usual simulator for β is

$$\widehat{\beta} = (X^\top X)^{-1} X^\top y,$$

which consists of the following quantities:

$$s = \left(\sum_{i=1}^n X_{ij} \ , \ \sum_{i=1}^n X_{ij}^2 \ , \ \sum_{i=1}^n X_{ij}X_{ik} \ , \ \sum_{i=1}^n X_{ij}y_i \right)$$

for all $j < k$. In literature, we often assume $X_{ij} \in [0, 1]$ and $y_i \in [0, 1]$, then each of these quantities/sums has sensitivity 1 and there are $3p + \binom{p}{2}$ of these sums. The creativity here: rather than privatizing $\widehat{\beta}$, which is very sensitive to outlier (the worst case), we could privatize the elements of $\widehat{\beta}$ that are not as sensitive.

So, adding $\text{Lap}\left(0, \dfrac{3p + \binom{p}{2}}{\varepsilon}\right)$ noise to each of the quantities in s preserves ε-DP. Combine them into $\widetilde{X^\top X}$ and $\widetilde{X^\top y}$, then by post processing, $\widetilde{\beta} = (\widetilde{X^\top X})^{-1}(\widetilde{X^\top y})$ is a DP estimate of β.

Remark 3.9 (Clamping)**.** If data are not naturally bounded, unlike counts, then clamping is often used. Choose bounds $L \leq U$ such that you expect $L \leq t(X_i) \leq U$ for most X_i in practice. The clamped value is

$$t(X_i)\big]_L^U = \begin{cases} U & \text{if } t(X_i) > U \\ t(X_i) & \text{if } L \leq t(X_i) \leq U \\ L & \text{if } t(X_i) < L, \end{cases}$$

then $T'(x) := \displaystyle\sum_{i=1}^n t(X_i)\big]_L^U$ has sensitivity $\Delta = U - L$.

4 Alternative DP Regimes

The main reference for this chapter is Li et al. [2017].

4.1 Bounded versus Unbounded DP

This is an exploration of what **Differential** Privacy means? Recall that (ε, δ)-DP requires

$$\Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S] + \delta$$

for all sets S and all **adjacent** X and X'.

The concept "adjacent" is meant to capture "**differing** in one entry/individual." So far, we have set $X \in \mathcal{X}^n$ where n is fixed and known, and used Hamming distance $H(X, X') \leq 1$ to determine adjacency. However, we may be worried that n itself may be a sensitive quantity. For example, suppose X is the database of citizens with the sickle cell trait. Knowing n, the sample size could be used for attack!

1. With Hamming distance, we change one entry (**bounded**).
2. Instead, we could add or remove an individual (**unbounded**).

Then, let $X \in \mathcal{X}^* = \emptyset \cup \mathcal{X} \cup \mathcal{X}^2 \cup \cdots$, where $\mathcal{X}^*$ is the set of all possible databases, and define

$$d(X, X') = 1,$$

if X' can be obtained from X by adding or deleting an entry. With such notation, we have

1. **Bounded** (add/delete) DP if $\mathcal{X}^n$ is used with distance measure H
2. **Unbounded** (change) DP if $\mathcal{X}^*$ is used with distance measure d

Proposition 4.1. Let $\mathcal{M} \colon \mathcal{X}^n \to \mathcal{Y}$ be a mechanism satisfying ε-DP (add/delete), then $\mathcal{M}$ satisfies 2ε-DP (change).

Proof. Let $X, X' \in \mathcal{X}^n$ differing in one entry. Let $\widetilde{X}$ be the database with that entry deleted $\mathcal{X}^{n-1}$. Then,

$$\underbrace{X,}_{\text{add } 1} \widetilde{X} \underbrace{,X'}_{\text{delete } 1}$$

is a sequence of neighboring databases in unbounded sense. Then, we can apply either group privacy or DP inequality twice to conclude. ■

Remark 4.1 (What is the difference?)**.** The key difference between bounded DP and unbounded DP is whether the sample size n is public or protected. Even if n is not a concern in and of itself, when using bounded DP, we need to be careful when applying DP mechanisms on subsets of the dataset.

Proposition 4.2. Let M satisfies f-DP (add/delete) then it satisfies $f(1-f)$-DP (change).

We can use group privacy result to prove this.

Proposition 4.3 (Parallel Composition for Unbounded DP)**.** Let $\mathcal{M}_1, \cdots, \mathcal{M}_k$ be k mechanisms satisfying ε_1-DP, $\cdots$, ε_k-DP (unbounded) respectively, where $\mathcal{M}_i : \mathcal{X}^* \to \mathcal{Y}_i$. Let f be a deterministic partitioning function, and let $X_1, \cdots, X_k$ be the resulting partitions from applying f to database X. Then $(M_1(X_1), M_2(X_2), \cdots, M_k(X_k))$ satisfies $(\max_i \varepsilon_i)$-DP (unbounded).

proof sketch. Let X, X' be two adjacent databases. The extra individuals lies in one of the k partitions. Whenever it is, say i, the privacy parameter is ε_i (other mechanisms are not affected). So the worst case if $\max_{i=1,\ldots,k} \varepsilon_i$. ■

Remark 4.2 (Proposition 4.3)**.** For tradeoff functions, the result would be:

$$f^*(\alpha) = \text{Hull}\left(\min_i f_i\right)(\alpha)\text{-DP}.$$

We take the convex hull for f_i's pointwise to ensure that f^* is a convex function.

Remark 4.3 (What can go wrong with bounded DP?)**.** Let $\mathcal{M}_n \colon \mathcal{X}^n \to \mathbb{R}$ be the (bounded) DP mechanism, which just returns n, satisfying 0-DP (well-defined for $n = 0, 1, 2, \ldots$). Let f be any partition function (e.g. men/women, infected/not infected, race, tabulation of all US citizens ...). Then, running $\mathcal{M}_n$ on each partition tells exactly how many records are in each partition. Obviously, this does not satisfy 0-DP as suggested by parallel composition (Proposition 4.3).

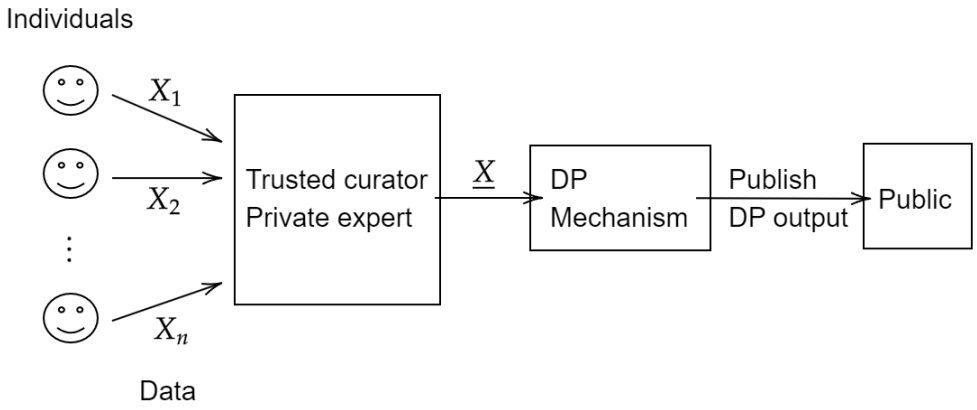
Takeaway: **Bounded-DP does not satisfy parallel composition!**

We can still use partitions/subsampling in bounded-DP, but need to analyze the whole system.

Example 4.1 (DP on graphs)**.** (Node)-DP is too dramatic and often not meaningful, whereas (Edge)-DP is more relaxed.

4.2 Local DP

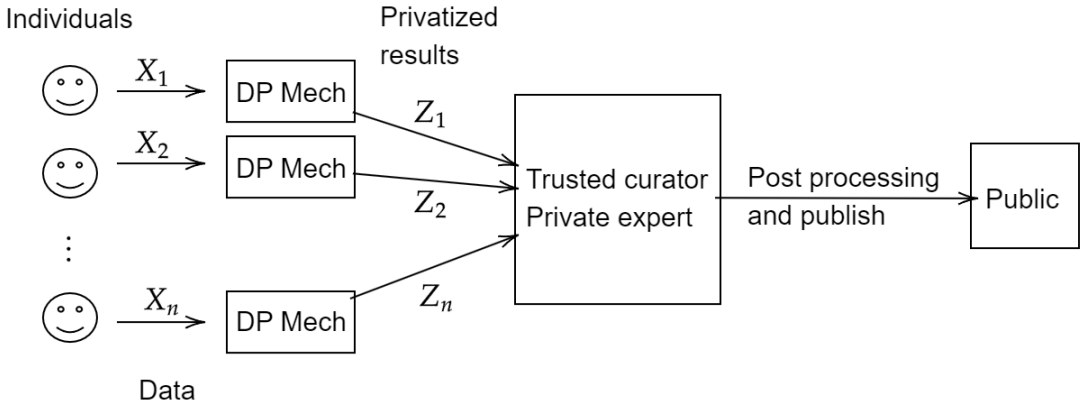
So far in this class, we have been working in the "Central Model" of DP that looks as follows.



The key part of this mechanism design is that we assume the privacy expert has access to the original database $\underline{X} = (X_1, \cdots, X_n)$ and he applies a DP mechanism to $\underline{X}$. But what if we do not trust the data curator? Or we are worried having the original data in one place (could be hacked)?

4.2.1 Local Model

A DP mechanism is applied "locally" (such as on the individual's device) before sending the results to the curator.



A complexity arises in local DP mechanisms: after receiving the DP response Z_1 from person 1, we may want to ask a different questions to person 2. That is, different questions may be asked based on previous answers and multiple rounds of communication can be required. We can express the mechanism applied to the i-th person as

$$Z_i \sim M(\ \cdot\ \mid \underbrace{X_i = x}_{\text{person } i}, Z_j = z_j, \forall j \neq i).$$

Definition 4.1 (LDP)**.** For a privacy parameter $\varepsilon > 0$, we say that M is ε-locally differentially private (ε-LDP) if

$$M(S \mid X_i = x, Z_j = z_j, j \neq i) \leq e^{\varepsilon} M(S \mid X_i = x', Z_j = z_j, j \neq i)$$

for all measurable sets S, all values of z_j, $j \neq i$ and all $x, x' \in \mathcal{X}$. When z_i is generated based only on X_i, then this simplifies to

$$\sup_S \sup_{x,x' \in \mathcal{X}} \frac{M(S \mid X_i = x)}{M(S \mid X_i = x')} \leq e^{\varepsilon}$$

Example 4.2 (Local DP Mechanisms)**.** (1) Randomized response is LDP when

$$Z_i = \begin{cases} X_i & \text{w.p. } \frac{e^{\varepsilon}}{1+e^{\varepsilon}} \\ 1 - X_i & \text{w.p. } \frac{1}{1+e^{\varepsilon}}, \end{cases}$$

where $X_i \in [0, 1]$.

(2) Laplace mechanism is LDP when, say $X_i \in [a, b]$,

$$Z_i = X_i + \frac{b-a}{\varepsilon} L_i,$$

where $L_i \sim \text{Lap}(0, 1)$.

Remark 4.4. Note that every ε-LDP mechanism satisfies ε-DP. All Central DP (bounded) mechanisms are local DP when applied to dataset of size 1.

4.2.2 Comparison of Minimax Rates

If LDP is stronger and safer than the central model, why not always use LDP? This is because of the error rate!

For $k > 1$, consider the families $\mathcal{P}_k$ of distributions P such that $\mathbb{E}_P\left[|X|^k\right] \leq 1$ and we want to estimate the mean $\theta(P) = \mathbb{E}_P[X] \in [-1, 1]$. We assume $k \geq 2$ throughout.

Remark 4.5. Higher moments control the tail of the distribution, which turns out affecting the rate in DP.

Definition 4.2 (Minimax: Without privacy constraints)**.** The minimax is defined as

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}_k} \mathbb{E}\left[\widehat{\theta}(\underline{X}) - \theta(P)\right]^2 \gtrsim \frac{1}{n}$$

and the matching upper bound is achieved by the sample mean.

Theorem 4.1 (Duchi et al. [2013, 2018]). Under local model ε-LDP ,

$$\inf_{\substack{\mathcal{M}\text{: }\varepsilon\text{-LDP} \\ \widehat{\theta}\text{ post processing}}} \sup_{P\in\mathcal{P}_k} \mathop{\mathbb{E}}_{\substack{X\sim P \\ Z\sim M(X)}} \left[\widehat{\theta}(Z) - \theta(P)\right]^2 \gtrsim \left(\frac{1}{n\varepsilon^2}\right)^{\frac{k-1}{k}}$$

Theorem 4.2 (Barber and Duchi [2014]). Under central model ε-DP,

$$\inf_{\mathcal{M}\text{: }\varepsilon\text{-DP}} \sup_{P\in\mathcal{P}_k} \mathop{\mathbb{E}}_{\substack{\text{randomness of} \\ X \text{ and } M}} \left[M(X) - \theta(P)\right]^2 \gtrsim \frac{1}{n} + \left(\frac{1}{n^2\varepsilon^2}\right)^{\frac{k-1}{k}}$$

Comparing two models, we have

- If $k = 2$:
  - LDP gives rate $\frac{1}{\varepsilon\sqrt{n}}$: significantly worse than $\frac{1}{n}$ from the non-private case.
  - ε-DP gives rate $\frac{1}{n} + \frac{1}{\varepsilon n} = \left(\frac{\varepsilon+1}{\varepsilon}\right)\frac{1}{n}$: same rate as non-private but lower effective sample size (ESS).
- As $k \to \infty$, $\mathbb{E}\left[|X|^k\right] \le 1$ becomes equivalent to $|X| \le 1$, and
  - LDP gives rate $\frac{1}{n\varepsilon^2}$: same rate as non-private but lower effective sample size (ESS).
  - ε-DP gives rate $\frac{1}{n} + \frac{1}{\varepsilon^2 n^2}$: same rate as non-private as $\frac{1}{\varepsilon^2 n^2}$ is asymptotically negligible!

As long as ε does not shrink too fast, in central model, the privacy error is either of the same order or smaller than the non-private statistical estimation error. Under LDP, unless $k = \infty$, the privacy error dominates the statistical estimation error.

5 Report Noisy Max

5.1 Algorithm of Report Noisy Max (RNM)

Example 5.1. We may want to consider the following question:

Q. What is the most common medical condition?

Suppose we want to know which condition is (approximately) the most common in medical database. For each diagnosis, we ask if they have the condition. Note that each person may have multiple conditions, so sensitivity is $m = \#$ of possible conditions. To release all noisy counting queries it requires $O_p(m/\varepsilon)$ noise.

Suppose there are m queries, each with sensitivity 1. The algorithm goes as follows:

1. Add independent $\mathrm{Lap}\left(0, \frac{2}{\varepsilon}\right)$ noise to each.
2. Return the index of the largest (noisy) value.

In fact, the algorithm reports the argument of the noisy max rather than the noisy max itself!

Proposition 5.1. Report Noisy Max satisfies $(\varepsilon, 0)$-DP.

Proof of Proposition 5.1. Let $H(X, X') \leq 1$ and call $f(X)$ and $f(X')$ the vector of queries under X and X'. Note that for all $j = 1, \cdots, m$

$$f_j(X') - 1 \leq f_j(X) \leq f_j(X') + 1 of \tag{5.1}$$

since the sensitivity is 1 for each query.

Fix $i \in \{1, \cdots, m\}$ and consider the PMFs $\Pr[i \mid X]$ and $\Pr[i \mid X']$. Fix $r_{-i} \sim \mathrm{Lap}^{m-1}\left(0, \frac{2}{\varepsilon}\right)$, which is the noise for all except the i-th entry. By definition, we wish to show $\Pr[i \mid X] \leq \Pr[i \mid X']$, which is equivalent of showing

$$\mathbb{E}_{r_{-i}} \Pr[i \mid X, r_{-i}] \leq \mathbb{E}_{r_{-i}} \Pr[i \mid X', r_{-i}].$$

Then, it suffices to show

$$\Pr[i \mid X, r_{-i}] \leq \Pr[i \mid X', r_{-i}].$$

We start the proof from defining

$$r^* := \min_{r_i} : \, f_i(X) + r_i > f_j(X) + r_j \quad \forall j \neq i, \tag{5.2}$$

the smallest noise needed to make ith the noisy max.

Once r_{-i} is fixed, we see that i is the output under X if and only if $r_i \geq r^*$, i.e.

$$\Pr[i \mid X, r_{-i}] = \Pr(r_i \geq r^*).$$

For all $j \neq i$, $f_i(X) + r^* > f_j(X) + r_j$ by Eq. (5.2). Then,

$$\begin{aligned} 1 + f_i(X') + r^* &\geq f_i(X) + r^* && \text{(Eq. (5.1))} \\ &> f_j(X) + r_j && \text{(Eq. (5.2))} \\ &\geq (f_j(X') - 1) + r_j && \text{(Eq. (5.1))}, \end{aligned}$$

which implies $f_i(X') + (r^* + 2) \geq f_j(X') + r_j$. So, if $r_i \geq r^* + 2$, then under X' , the output is i (in the worst case). Hence, by applying the Laplace mechanism $r_i \sim \mathrm{Lap}\big(0, \frac{2}{\varepsilon}\big)$, we get

$$\begin{aligned} \Pr[i \mid X', r_{-i}] \geq \Pr[r_i \geq r^* + 2] &= \Pr[r_i - 2 \geq r^*] \\ &\geq e^{-\varepsilon} \Pr[r_i \geq r^*] \\ &= e^{-\varepsilon} \Pr[i \mid X, r_{-1}], \end{aligned}$$

which implies $\Pr[i \mid X, r_{-i}] \leq e^{\varepsilon} \Pr[i \mid X', r_{-i}]$. To carefully complete the proof by following the claim, we take expectation/marginalize over r_{-i} on both sides to obtain

$$\mathop{\mathbb{E}}_{r_{-i}}\big[\Pr[i \mid X, r_{-i}]\big] \leq e^{\varepsilon} \mathop{\mathbb{E}}_{r_{-i}}\big[\Pr[i \mid X', r_{-i}]\big] \iff \Pr[i \mid X] \leq e^{\varepsilon} \Pr[i \mid X'].$$

Then, it follows that RNM satisfies $(\varepsilon, 0)$-DP. ■

5.2 RNM Analysis

Measure the utility of RNM by the expected utility gap (excess risk) i.e. the expected difference between the $\max_i f_i(X)$ and $f_{M(X)}(X)$ (holding X fixed)

Example 5.2. Suppose we have k observations, we have $f_1(X) = 100$ and $f_i(X) = 90$ for other i's. When $k = 1000$, it is more likely to have some f_i surpassing f_1 after noise is added.

We can analyze RNM with any additive noise (continuous and symmetric) $r_i \overset{iid}{\sim} F$.

Lemma 2. Let $f_1, \ldots, f_k : \mathcal{X}^n \to \mathbb{R}$ be k queries with sensitivity 1, then RNM with noise distribution F (continuous and symmetric) has expected utility gap bounded by

$$\max_i f_i(X) - \mathop{\mathbb{E}}_{\substack{M(X) \\ (r_1, \ldots, r_k)}} f_{M(X)}(X) \leq \mathbb{E}|r_1| + \mathbb{E} \max_{2 \leq i \leq k} |r_i|$$

Proof. Fix X. Assume without loss of generality that $f_i(X) \leq 0$ for all i and $\max_i f_i(X) = 0$ (relabeled if necessary). Assume further that $i = 1$ is the maximizer.

Given RVs $r_1, \ldots, r_k$. Note that $M(X) \overset{d}{=} M(X; (r_i)_{i=1}^k)$ and

$$M(X; (r_i)_{i=1}^k) = \begin{cases} 1, \text{if } \max_{2 \leq j \leq k}[f_j(X) + r_j] \leq r_1 \\ 0, \text{if } i \geq 2 \text{ and } f_i(X) + r_i \geq \max_{j \neq i}[f_j(X) + r_j], \end{cases}$$

where ties happen with probability 0. We want a lower bound

$$\begin{aligned}
\mathbb{E} f_{M(X)}(X) &= \underbrace{\mathbb{E} f_{M(X)}(X) I(M(X)=1)}_{=0} + \mathbb{E} f_{M(X)}(X) I(M(X)\neq 1) \\
&= \mathbb{E} f_{M(X)}(X) I\left(\max_{2\le j\le k}[f_j(X)+r_j] > r_1\right) \\
&= \mathbb{E} f_{M(X)}(X) I(f_{M(X)}(X) > r_1 - \max_{2\le j\le k} r_j) \qquad (5.3) \\
&\ge \mathbb{E}\left(r_1 - \max_{2\le j\le k} r_j\right) I(f_{M(X)}(X) > r_1 - \max_{2\le j\le k} r_j) \qquad (5.4) \\
&\ge \mathbb{E}\left(r_1 - \max_{2\le j\le k} r_j\right) I(0 > r_1 - \max_{2\le j\le k} r_j) \qquad (5.5) \\
&\ge -\mathbb{E}|r_1 - \max_{2\le j\le k} r_j| \\
&\ge -[\mathbb{E}|r_1| + \mathbb{E}\max_{2\le j\le k}|r_j|]
\end{aligned}$$

Eq. (5.3): Under $M(X)\neq 1$, we have $r_1 < \max_{2\le j\le k}[f_j(X)+r_j] = f_{M(X)}(X) + r_{M(X)}$. We can now make the indicator set larger by including elements from $\{M(X)=1\}$—in which $f_{M(X)}(X)=0$)—to obtain an upper bound $f_{M(X)}(X) + \max_{2\le j\le k} r_j$ of $f_{M(X)}(X) + r_{M(X)}$. This does not change the equality because $f_{M(X)}(X) = 0$.

Eq. (5.4): We substitute $f_{M(X)}(X)$ with it lower bound $r_1 - \max_{2\le j\le k} r_j$, which makes the expectation (integral) smaller. This is similar in the proof of Chebyshev's inequality.

Eq. (5.5): Under $M(X)\neq 1$, $f_{M(X)}(X)\le 0$, which makes the indicator set larger. ■

Lemma 3. Let $X_1,\dots,X_k$ be RVs with moment generating function M for $t>0$, then

$$\mathbb{E}\max_i X_i \le \frac{1}{t}\log k + \frac{1}{t}\log(M_X(t)).$$

Proof. Let $t>0$,

$$\begin{aligned}
\mathbb{E}\max_i X_i &\le \frac{1}{t}\log\left(\mathbb{E}\max_i \exp(tX_i)\right) \qquad \text{(by Jensen's inequality)} \\
&\le \frac{1}{t}\log\big(k\mathbb{E}\exp(tX_i)\big),
\end{aligned}$$

where we use the fact that $\exp(tX)$ is monotone in X and max is (crudely) upper bounded by sum of non-negative values. ■

Proposition 5.2. RNM with $\mathrm{Lap}\big(0,\frac{2}{\varepsilon}\big)$ noise has expected utility gap

$$\max_i f_i(X) - \mathbb{E}_{\substack{M(X)\\(r_1,\dots,r_k)}} f_{M(X)}(X) \le \frac{2}{\varepsilon}(2\log(k-1)+1+2\log 2),$$

which is $O\Big(\frac{\log k}{\varepsilon}\Big)$.

Proof. Recall if $L \sim \text{Lap}\big(0, \frac{2}{\varepsilon}\big)$ and $|L| \sim \text{Exp}(\mu = \frac{2}{\varepsilon})$ from lemma above, MGF for Exp(μ) is $M(t) = (1-\mu t)^{-1}$ for $t < \frac{1}{\mu}$.

Choosing $t = \frac{1}{2\mu}$, it gives

$$\mathbb{E} \max_i f_i(X) \leq 2\mu \log k + 2\mu \log 2,$$

where $X_i \overset{iid}{\sim} \text{Exp}(\mu)$. Thus,

$$\begin{aligned}
\max_i f_i(X) - \mathbb{E}_{M(X)} f_{M(X)} &\leq \mathbb{E}|L_i| + \mathbb{E} \max_{2 \leq j \leq k} |L_j| \\
&\leq \frac{2}{\varepsilon} + \frac{4}{\varepsilon} \log(k-1) + \frac{4}{\varepsilon} \log 2 \\
&= \frac{2}{\varepsilon}(1 + 2\log(k-1) + 2\log 2) \\
&= O\left(\frac{\log k}{\varepsilon}\right).
\end{aligned}$$

■

5.3 RNM Applications

Example 5.3 (Linear Regression Selection Lei et al. [2018])**.** We may have multiple regression models, but we want to find the simplest one which explains the data well. Equivalently, we want to determine which entries of β are non-zero. Assume

$$\begin{aligned}
&|y_i| \leq r \text{ for } i = 1, \cdots, n \\
&|X_{ij}| \leq 1 \text{ for } i = 1, \cdots, n \text{ and } j = 1, \cdots, d \\
&\|\beta\|_1 \leq R
\end{aligned}$$

The optimal fit for the model M is

$$\widehat{\beta}_M = \underset{\substack{\|\beta\|_1 \leq R \\ \beta \in \Theta_M}}{\text{argmin}} \underbrace{-2\ell(\beta; D)}_{= \sum_{i=1}^n (y_i - X_i^\top \beta)^2} + \varphi_n |M|$$

where ℓ is log likelihood. Note that

- $\varphi_n = 2$ gives AIC;
- $\varphi_n = \log(n)$ gives BIC;
- $\varphi_n = 0$ gives standard least squares regression.

For a model M, its score is

$$\boxed{\min_{\substack{\|\beta\|_1 \leq R \\ \beta \in \Theta_M}} -2\ell(\beta; D) + \varphi_n |M|}$$

which has sensitivity $(r+R)^2$. This is because $(y_i - X_i^\top \beta)^2$ has the minimum value 0 and needs to calculate maximum i,

$$\max\Big(y_i - X_i^\top \beta\Big)^2 \le \max\Big(|y_i| + |X_i^\top \beta|\Big)^2 \le \left(r + \underbrace{\max |X_i^\top \beta|}_{\le \|X_i\|_\infty \|\beta\|_1}\right)^2 \le (r+R)^2$$

by Holder's inequality. So, $-2\ell(\beta; D)$ has sensitivity $(r+R)^2$.

Example 5.4 (Private Model Selection Via Report Noisy Max)**.** For each candidate model M,

$$\begin{aligned}
\ell_R(M; D) &= \max_{\substack{\beta \in \Theta_M \\ \|\beta\|_1 \le R}} -\frac{1}{2}\sum_{i=1}^n (y_i - X_i^\top \beta)^2 \\
L_R(M; D) &= -2\ell_R(M; D) + \varphi_n |M| \\
\widetilde{L}_R(M; D) &= L_R(M; D) + \frac{2(r+R)^2}{\varepsilon} L_M,
\end{aligned}$$

where $L_M \sim \mathrm{Lap}(0, 1)$. Return $\mathrm{argmin}_M \widetilde{L}_R(M; D)$ satisfying ε-DP and it has error $O\Big(\frac{\log 2^d (r+R)^2}{\varepsilon}\Big) = O\Big(\frac{d(r+R)^2}{\varepsilon}\Big)$.

6 Objective Perturbation

6.1 Objective Perturbation for Empirical Risk Minimization

We are interested in calculating

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} l(\theta; d_i),$$

where $d_i = (X_i, y_i)$ (represents the data for individual i) and l is a loss of the form $l(\theta; d_i) = g(X_i^\top \theta, y_i)$. But, often the sensitivity of $\widehat{\theta}$ is too high to use an additive mechanism. If the loss f

1. is convex on Θ,
2. has continuous Hessian, i.e. $\nabla^2 \ell(\theta; x)$ is continuous in x and θ,
3. has finite sensitivity of gradient, i.e. $\sup_{D,D'} \sup_{\theta} \|\nabla l(\theta, D) - l(\theta, D')\|_2 = \Delta < \infty$,
4. λ is an upper bound on eigenvalues of $\nabla^2 l(\theta, x)$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$,

then the Extended Objective Perturbation [Awan and Slavković, 2021] produces $\widetilde{\theta}$ as follows

1. choose $0 < q < 1$,
2. set $\gamma = \frac{\lambda}{\exp(\varepsilon(1-q))-1}$,
3. sample V from density $\propto \exp(\frac{-\varepsilon q}{\Delta}\|V\|)$,
4. $\widetilde{\theta} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} l(\theta, d_i) + \frac{1}{n} r(\theta) + \underbrace{\frac{\gamma}{2n}\theta^\top\theta}_{\text{regularization}} + \underbrace{\frac{1}{n}V^\top\theta}_{\text{random linear term}}$.

We can also view it as

$$\widetilde{\theta} = \arg_\theta \text{zero} \frac{1}{n}\sum_{i=1}^{n} \nabla l(\theta; d_i) + \frac{1}{n}\nabla r(\theta) + \frac{\gamma}{n}\theta + \frac{V}{n},$$

which is conceptually similar to RNM as it only reports the argmin (argmax) after noise addition. Note that when V is Gaussian, it satisfies (ε, δ)-DP.

Remark 6.1. Objective perturbation has developed as follows:

- Chaudhuri and Monteleoni [2008] develops privacy-preserving logistic regression algorithms, highlighting a novel objective perturbation method and its link to regularization.
- Chaudhuri et al. [2011] introduces objective perturbation for differentially private ERM, outperforming output perturbation [Dwork et al., 2006] in theory and experiments.

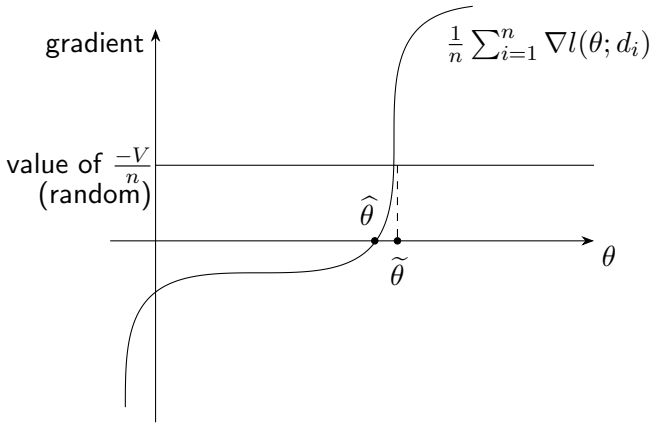


Figure 2: Extended Objective Perturbation

- Kifer et al. [2012] extends objective perturbation for private ERM and introduces new sparse regression algorithms effective in high-dimensional settings.

Remark 6.2. As illustrated in Figure 2, by choosing a random y-intercept, the mechanism will (hopefully) result in a small perturbation in $\widetilde{\theta}$ compared to $\widehat{\theta}$.

Theorem 6.1. The Extended Objective Perturbation satisfies ε-DP.

Proof. We assume $r(\theta)$ is twice differentiable and $\Theta \in \mathbb{R}^m$—techniques in Kifer et al. [2012] allow us to extend the result to arbitrary convex $r(\theta)$ and convex Θ. It suffices to show that for all $a \in \mathbb{R}^m$ and $H(D, D') \leq 1$,

$$\frac{pdf(\widetilde{\theta} = a \mid D)}{pdf(\widetilde{\theta} = a \mid D)'} \leq \exp(\epsilon)$$

Let $a \in \mathbb{R}^m$ and D, D' be given. If $\widetilde{\theta} = a$, then

$$a = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} l(\theta; d_i) + r(\theta) + \frac{\gamma}{2}\theta^\top\theta + V^\top\theta.$$

Taking the gradient and setting to zero, we solve for V as

$$V(a; D) = -\left(\sum_{i=1}^{n} \nabla l(a; d_i) + \nabla r(a) + \gamma a\right).$$

Applying change of variable, we get

$$\frac{pdf(\widetilde{\theta} = a \mid D)}{pdf(\widetilde{\theta} = a \mid D)'} = \frac{f(V(a; D))}{f(V(a; D)')} \frac{|\det(\nabla V(a; D'))|}{|\det(\nabla V(a; D))|},$$

where f is the density of V. Now we want to bound each factor separately. First we have

$$\frac{f(V(a;D))}{f(V(a;D)')} \leq \exp(\frac{\varepsilon q}{\Delta}\Delta) = \exp(\varepsilon q),$$

since the K-Norm mechanisms satisfy ε-DP (**HW1**) and scale is proportional to sensitivity of gradient.

For the second factor, we assume without the loss of generality that $d_i = d_i{}'$ for $i = 1, \ldots, n-1$. Call $A = -\nabla V(a;D)$, $B = -\nabla V(a;D')$, and $C = \sum_{i=1}^{n} \nabla^2 l(\theta; d_i) + \nabla^2 r(a) + \gamma I_m$. Note that $A = C + \nabla^2 l(a; d_n)$ and $B = C + \nabla^2 l(a; d_n{}')$. Then,

$$\begin{aligned}
\frac{|\det(\nabla V(a;D'))|}{|\det(\nabla V(a;D))|} &= \frac{\det(B)}{\det(A)} \\
&= \frac{\det(C + \nabla^2 l(a; d_n{}'))}{\det(C + \nabla^2 l(a; d_n))} \\
&= \frac{\det C}{\det C} \frac{\det(I_m + C^{-1}\nabla^2 l(a; d_n{}'))}{\det(I_m + C^{-1}\nabla^2 l(a; d_n))} \\
&\leq \frac{1 + \lambda/\gamma}{\underbrace{1}_{\text{lower bound on det because eigenvalues} \geq 1}} \\
&= 1 + \frac{\lambda}{\lambda}[\exp(\varepsilon(1-q)) - 1] \\
&= \exp(\varepsilon(1-q)).
\end{aligned}$$

Since $l(a; d_n{}') = g(a^\top X_n; y_n{}')$, the Hessian can be expressed as

$$\nabla^2 l(a; d_n{}') = g''(a^\top X_n; y_n{}') \underbrace{X_n X_n^\top}_{\text{rank} \leq 1},$$

which renders $C^{-1}\nabla^2 l(a; d_n{}')$ rank less than or equal to 1. Note that determinant is product of eigenvalues. So, the eigenvalues of $I_m + C^{-1}\nabla^2 l(a; d_n{}')$ are all 1 except the largest, which is $\leq 1 + \frac{\lambda}{\gamma}$. This can be seen from γ being a lower bound on eigenvalues of C.

Multiplying the two bounds, we get

$$\frac{pdf(\widetilde{\theta} = a \mid D)}{pdf(\widetilde{\theta} = a \mid D)'} \leq \exp(\varepsilon q)\exp(\varepsilon(1-q)) = \exp(\varepsilon).$$

■

6.2 Analysis of Objective Perturbation

For later convenience, let

$$\widehat{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n} l(\theta; d_i) \qquad \text{with } \widehat{\theta} = \operatorname*{argmin}_{\theta} \widehat{L}(\theta; D)$$

$$L^{\#}(\theta; D) = \widehat{L}(\theta; D) + \frac{\gamma}{2n}\theta^\top\theta \qquad \text{with } \theta^{\#} = \operatorname*{argmin}_{\theta} L^{\#}(\theta; D)$$

$$L^{\text{priv}}(\theta; D) = L^{\#}(\theta; D) + \frac{1}{n}V^\top\theta \qquad \text{with } \theta^{\text{priv}} = \operatorname*{argmin}_{\theta} L^{\text{priv}}(\theta; D)$$

Remark 6.3. This decomposition allows us to analyze the error rate of $\|\theta^{\text{priv}} - \widehat{\theta}\|$ in two steps: variance (from DP randomness) and bias (from regularization).

Our first step is to understand $(\theta^{\text{priv}} - \theta^{\#})$, also known as the variance term. Consider

$$\nabla L^{\#}(\theta^{\text{priv}}; D) = \underbrace{\nabla L^{\#}(\theta^{\#}; D)}_{=0} + \nabla^2 L^{\#}(\theta_1; D)(\theta^{\text{priv}} - \theta^{\#}),$$

where θ_1 is between θ^{priv} and $\theta^{\#}$. Thus,

$$(\theta^{\text{priv}} - \theta^{\#}) = (\nabla^2 L^{\#}(\theta_1; D))^{-1} \underbrace{\nabla L^{\#}(\theta^{\text{priv}}; D)}_{=\nabla L^{\text{priv}}(\theta^{\text{priv}}; D) - \frac{V}{n} = -\frac{V}{n}} = (\nabla^2 L^{\#}(\theta_1; D))^{-1}(\frac{-V}{n})$$

Assume $\nabla^2 L^{\#}(\theta_1; D) \to \Sigma^{\#}$, we have

$$n(\theta^{\text{priv}} - \theta^{\#}) \xrightarrow{d} (\Sigma^{\#})^{-1}V$$

since $V \overset{d}{=} -V$. Then,

$$\|\theta^{\text{priv}} - \theta^{\#}\| = O_p\left(\|(\Sigma^{\#})^{-1}\|\frac{\Delta}{\epsilon q n}\|V\|\right) = O_p\left(\frac{\|(\Sigma^{\#})^{-1}\|\Delta m}{\epsilon q n}\right),$$

where the second inequality is given by Awan and Slavković [2021].

However, $\theta^{\#}$ is a biased estimator for θ. If the bias is larger than the privacy error, Objective Perturbation may not work well. So, our second step is to understand $(\theta^{\#} - \widehat{\theta})$, also known as the bias term. Consider

$$\nabla L^{\#}(\widehat{\theta}; D) = \underbrace{\nabla L^{\#}(\theta^{\#}; D)}_{=0} + \nabla^2 L^{\#}(\theta_2; D)(\widehat{\theta} - \theta^{\#}),$$

where θ_2 is between $\widehat{\theta}$ and $\theta^{\#}$. Thus,

$$(\widehat{\theta} - \theta^{\#}) = (\nabla^2 L^{\#}(\theta_2; D))^{-1} \underbrace{\nabla L^{\#}(\widehat{\theta}; D)}_{\nabla \widehat{L}(\widehat{\theta}; D) + \frac{\gamma}{n}\widehat{\theta} = \frac{\gamma}{n}\widehat{\theta}} = (\nabla^2 L^{\#}(\theta_2; D))^{-1}\left(\frac{\gamma}{n}\widehat{\theta}\right)$$

Assume $\nabla^2 L^{\#}(\theta_2; D) \to \Sigma^{\#}$, we have

$$n(\widehat{\theta} - \theta^{\#}) \to (\Sigma^{\#})^{-1}\gamma\widehat{\theta}.$$

Then,

$$\|\theta^{\#} - \widehat{\theta}\|_2 = O\left(\frac{\|\Sigma^{\#}\|^{-1}\gamma\|\widehat{\theta}\|_2}{n}\right) = O\left(\frac{\|\Sigma^{\#}\|^{-1}\gamma\sqrt{m}}{n}\right).$$

By the triangle inequality and combining the variance and bias terms, we obtain

$$\|\theta^{\mathsf{priv}} - \widehat{\theta}\| = O_p\left(\frac{\|\Sigma^{\#}\|^{-1}\gamma\sqrt{m}}{n} + \frac{\|(\Sigma^{\#})^{-1}\|\Delta m}{\epsilon q n}\right) = O_p\left(\frac{\|(\Sigma^{\#})^{-1}\|\Delta m}{\epsilon q n}\right),$$

where the second inequality holds by assuming ϵ, q, γ are bounded.

7 Exponential Mechanism

The main reference for this chapter is McSherry and Talwar [2007].

So far, Laplace mechanisms, RNM, Objective Perturbation all assume "nearby" values given nearby "utility"....are useful when values that are close in ℓ_1-distance have similar utility.

However, this is not always the case!

Example 7.1 (Pumpkin Merchandise)**.** Suppose that we have lots of pumpkins to sell and have four potential buyers: A, B, C, and D. A, B, and C are all willing to pay up to $\$1$/pumpkin, but D is willing to pay up to $\$4.01$/pumpkin. What is the optimal price?

- Price at $\$4.01$: the revenue is $\$4.01$.
- Price at $\$1$: the revenue is $\$4$.
- Price at $\$1.01$: the revenue is $\$1.01$.
- Price at $\$4.02$: the revenue is $\$0$.

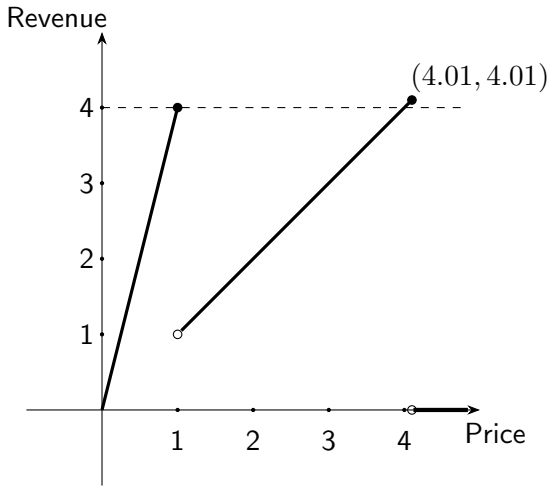


Figure 3: Pumpkin Merchandise

Note the discontinuities of this function (price vs. revenue) and that adding noise directly to the prize could completely destroy utility! For example, adding noise at $(\text{price}, \text{revenue}) = (4.01, 4.01)$ could make your revenue 0 at half of the times.

The exponential mechanism allows us to answer queries with arbitrary utility functions

$$u \colon \mathcal{X}^n \times \mathcal{Y} \to \mathbb{R},$$

takes a database and a output and returns utility. We prefer higher utility.

Definition 7.1. The sensitivity of u is

$$\Delta u = \sup_{y \in \mathcal{Y}} \sup_{H(X,X') \leq 1} \big| u(X, y) - u(X', y) \big| < \infty$$

Theorem 7.1. The exponential mechanism (ExpMech) outputs the value $y \in \mathcal{Y}$ with probability/density proportional to

$$\exp\left(\frac{\varepsilon}{2\Delta u} u(X, y)\right)$$

with respect to a nontrivial base measure on $\mathcal{Y}$, and it satisfies $(\varepsilon, 0)$-DP.

Proof of Theorem 7.1. Call μ the base measure, then for $H(X, X') \leq 1$ and $y \in \mathcal{Y}$, the ratio of the densities is

$$\frac{\frac{\exp\left(\frac{\varepsilon}{2\Delta u} u(X,y)\right)}{\int \exp\left(\frac{\varepsilon}{2\Delta u} u(X,z)\right) d\mu(z)}}{\frac{\exp\left(\frac{\varepsilon}{2\Delta u} u(X',y)\right)}{\int \exp\left(\frac{\varepsilon}{2\Delta u} u(X',z)\right) d\mu(z)}} = \underbrace{\frac{\int \exp\left(\frac{\varepsilon}{2\Delta u} u(X', z)\right) d\mu(z)}{\int \exp\left(\frac{\varepsilon}{2\Delta u} u(X, z)\right) d\mu(z)}}_{=:A} \underbrace{\frac{\exp\left(\frac{\varepsilon}{2\Delta u} u(X, y)\right)}{\exp\left(\frac{\varepsilon}{2\Delta u} u(X', y)\right)}}_{=:B} = AB$$

B can be simplified as

$$B = \frac{\exp\left(\frac{\varepsilon}{2\Delta u} u(X, y)\right)}{\exp\left(\frac{\varepsilon}{2\Delta u} u(X', y)\right)} = \exp\left(\frac{\varepsilon}{2\Delta u}\big(u(X, y) - u(X', y)\big)\right) \leq \exp\left(\frac{\varepsilon}{2\Delta u}\Delta u\right) = \exp\left(\frac{\varepsilon}{2}\right)$$

For A, we do the following

$$\begin{aligned} A &= \frac{\int \exp\left(\frac{\varepsilon}{2\Delta u} u(X', z)\right) d\mu(z)}{\int \exp\left(\frac{\varepsilon}{2\Delta u} u(X, z)\right) d\mu(z)} \\ &\leq \frac{\int \exp\left(\frac{\varepsilon}{2\Delta u}\big(u(X, z) + \Delta u\big)\right) d\mu(z)}{\int \exp\left(\frac{\varepsilon}{2\Delta u} u(X, z)\right) d\mu(z)} = e^{\frac{\varepsilon}{2}} \frac{\int \exp\left(\frac{\varepsilon}{2\Delta u} u(X, z)\right) d\mu(z)}{\int \exp\left(\frac{\varepsilon}{2\Delta u} u(X, z)\right) d\mu(z)} = e^{\frac{\varepsilon}{2}} \end{aligned}$$

Combining the two, we get $AB \leq e^{\frac{\varepsilon}{2}} e^{\frac{\varepsilon}{2}} = e^{\varepsilon}$ and this concludes the proof. ■

7.1 Utility of the Exponential Mechanism

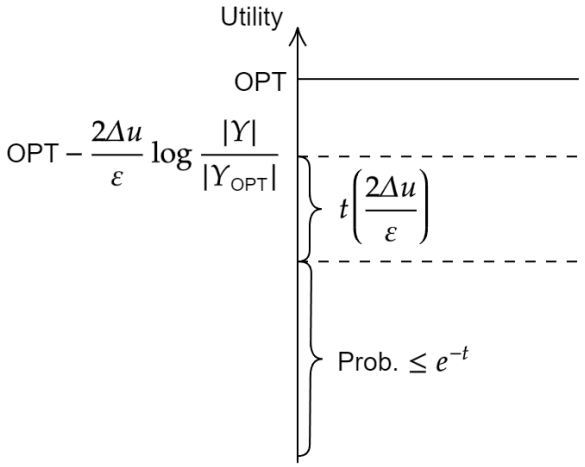
For a given database X and a utility measure $u\colon \mathcal{X}^n \times y \to \mathbb{R}$, let $\mathrm{OPT}_u(X) := \max_{y \in \mathcal{Y}} u(X, y)$ denote the maximum utility score over the possible outputs $y \in \mathcal{Y}$, with respect to the database X.

We will bound the probability that the ExpMech returns a "good" element of $\mathcal{Y}$, where "good" is measured by the difference between $u(X, M(X))$ and $\mathrm{OPT}_u(X)$, where $M(X)$ is ExpMech with utility u at ε-DP[3]. We will see in the next theorem that it is unlikely that this difference is greater than $O\left(\frac{\Delta u}{\varepsilon} \log |\mathcal{Y}|\right)$ assuming that $\mathcal{Y}$ is discrete with cardinality $|\mathcal{Y}|$. Practically, since $\mathcal{Y} = \mathbb{R}^d = (\#\text{float})^d$, $O\left(\frac{\Delta u}{\varepsilon} \log |\mathcal{Y}|\right) = O\left(\frac{\Delta u}{\varepsilon} d\right)$, which the standard error rate of an ExpMech.

[3]Since $M(X)$ is a r.v., $u(X, M(X))$ is also a r.v.

Theorem 7.2. Assume $|\mathcal{Y}| < \infty$. Let X be a fixed database and let $\mathcal{Y}_{\mathrm{OPT}_u(X)} = \{y \in \mathcal{Y} \mid u(X,y) = \mathrm{OPT}_u(X)\}$ denote the set of outputs which attain the utility score $\mathrm{OPT}_u(X)$. Then,

$$\Pr\left[u(X, M(X;u)) \le \mathrm{OPT}_u(X) - \frac{2\Delta u}{\varepsilon}\left(\log\frac{|\mathcal{Y}|}{|\mathcal{Y}_{\mathrm{OPT}}|} + t\right)\right] \le e^{-t}$$



Proof of Theorem 7.2. We have

$$\begin{aligned}
\Pr\big[u(X, M(X;u)) \le c\big] &= \frac{\sum_y I(u(X,y) \le c)\exp\left(\frac{\varepsilon}{2\Delta u}u(X,y)\right)}{\sum_y \exp\left(\frac{\varepsilon}{2\Delta u}u(X,y)\right)} \\
&\le \frac{\sum_y \exp\left(\frac{\varepsilon}{2\Delta u}c\right)}{\sum_y \exp\left(\frac{\varepsilon}{2\Delta u}u(X,y)\right)} \qquad (7.3) \\
&= \frac{|\mathcal{Y}|\exp\left(\frac{\varepsilon}{2\Delta u}c\right)}{\sum_y \exp\left(\frac{\varepsilon}{2\Delta u}u(X,y)\right)} \\
&\le \frac{|\mathcal{Y}|\exp\left(\frac{\varepsilon}{2\Delta u}c\right)}{\sum_{y\in\mathcal{Y}_{\mathrm{OPT}}} \exp\left(\frac{\varepsilon}{2\Delta u}\mathrm{OPT}_u(X)\right)} \qquad (7.4) \\
&= \frac{|\mathcal{Y}|\exp\left(\frac{\varepsilon}{2\Delta u}c\right)}{|\mathcal{Y}_{\mathrm{OPT}}|\exp\left(\frac{\varepsilon}{2\Delta u}\mathrm{OPT}_u(X)\right)} \\
&= \frac{|\mathcal{Y}|}{|\mathcal{Y}_{\mathrm{OPT}}|}\exp\left(\frac{\varepsilon}{2\Delta u}\left(c - \mathrm{OPT}_u(X)\right)\right) \qquad (7.5)
\end{aligned}$$

Eq. (7.3): (A common and useful trick) Find the upped bound by considering $I(u(X,y) \le c) \le 1$ and $u(X,y) \le c$ as $I(u(X,y) \le c) = 1$.

Eq. (7.4): Restrict the set to be summed over to its subset $\mathcal{Y}_{\mathrm{OPT}}$ and use the fact that $u(X,y) = \mathrm{OPT}_u(X)$ for the subset.

Now, we set $c = \mathrm{OPT}_u(X) - \frac{2\Delta u}{\varepsilon}\left(\log\frac{|\mathcal{Y}|}{|\mathcal{Y}_{\mathrm{OPT}}|} + t\right)$ to get the desired result. ■

Corollary 7.1. Fixing X,

$$\Pr\left[u(X, M(X)) \leq \mathrm{OPT}_u(X) - \frac{2\Delta u}{\varepsilon}\big(\log|\mathcal{Y}| + t\big)\right] \leq e^{-t}$$

This is because we always have $|\mathcal{Y}_{\mathrm{OPT}}| \geq 1$.

Example 7.2. Consider the question of determining which two medical conditions, A and B, are more common. Suppose that

$$A(X) = \#\text{ of } A \text{ in database } X$$
$$B(X) = \#\text{ of } B \text{ in database } X$$

and assume $A(X) > B(X)$. The utility measure is

$$u(X, A) = A(X) \text{ and } u(X, B) = B(X)$$

and note that $\Delta u = 1$. By our analysis of the exponential mechanism (denoted as M here),

$$\begin{aligned}\Pr[M(X; u) = B] &= \Pr[u(X, M(X; u)) \leq B(X)] \\ &\leq \frac{2}{1}\exp\left(\frac{\varepsilon}{2}(B(X) - A(X))\right),\end{aligned}$$

where the inequality is given by Eq. (7.5) with $|\mathcal{Y}| = 2$, $|\mathcal{Y}_{\mathsf{OPT}}| = 1$, $\Delta_u = 1$, and $\mathrm{OPT}_u(X) = A(X)$. We can see that as the gap between $A(X)$ and $B(X)$ widens, the probability that B output becomes exponentially smaller.

7.2 Connection between ExpMech and RNM

In the case that $|\mathcal{Y}|$ is finite, one could apply either ExpMech or RNM as our utility results showed that they have comparable performance. In fact, ExpMech is a special case of RNM with the use of a different noise distribution.

Q: How to sample from a pmf on finite $\mathcal{Y} = \{1, 2, \ldots, N\}$?

Suppose we want to sample Y from a discrete distribution $\{1, 2, \ldots, N\}$ where the un-normalized log-probability of outcome k is

$$z_k = \log(P(Y = k)) + c,$$

where c is the normalizing constant unknown and note that $z_k \in \mathbb{R}$.

- The obvious solution is to construct pmf/cdf and use standard samplers

$$\pi(k) = P(Y = k) = \frac{\exp(z_k)}{\sum_{i=1}^{N}\exp(z_k)}.$$

  A minor issue with this solution is the expense to computer sum. A potentially more major issue would be that its numerical instability on $\exp(z_k)$ could be very large or very small.

- Alternatively, we can use the Gumbel-max trick:

$$y = \underset{k=1,\dots,N}{\operatorname{argmax}}(z_k + G_k),$$

where $G_k \overset{iid}{\sim}$ Gumbel$(0,1)$ with pdf $f(x) = \exp(-x - e^{-x})$ and cdf $F(x) = \exp(-e^{-x})$. The benefits are that it avoids summation and works in log-space avoiding numerical issues.

Proposition 7.1. The Gumbel-max trick gives

$$Y \sim \pi.$$

Proof. Notice that Y does not change if we replace z_k with $\log(\pi_k)$. Define $U_k = \log(\pi_k) + G_k$, then

$$\begin{aligned}
P(Y = k) &= P(U_k \geq U_i, \forall i \neq k) \\
&= \int_{-\infty}^{\infty} P(U_k \geq U_i, \forall i \neq k \mid U_k = u_k) p(u_k) du_k \\
&= \int_{-\infty}^{\infty} \prod_{i \neq k} P(U_k \geq U_i \mid U_k = u_k) p(u_k) du_k \\
&= \int_{-\infty}^{\infty} \prod_{i \neq k} P(G_i \leq U_k - \log(\pi_i) \mid U_k = u_k) p(u_k) du_k \\
&= \int_{-\infty}^{\infty} \prod_{i \neq k} \exp(-e^{\log(\pi_i) - u_k} - u_k) f(u_k - \log(\pi_k)) du_k \\
&= \int_{-\infty}^{\infty} \prod_{i \neq k} \exp(-e^{\log(\pi_i) - u_k} - u_k) \exp(-[u_k - \log(\pi_k)] - e^{-[u_k - \log(\pi_k)]}) du_k \\
&= \int_{-\infty}^{\infty} \exp\left(\sum_{i \neq k} \pi_i e^{-u_k} \right) \pi_k \exp(-u_k - \pi_k e^{-u_k}) du_k \\
&= \pi_k \int_{-\infty}^{\infty} \exp(-u_k - \underbrace{\sum_{i=1}^{N} \pi_i}_{=1} e^{-u_k}) du_k \\
&= \pi_k
\end{aligned}$$

The last quality holds since the integrand is the pdf of Gumbel(0,1). ■

Example 7.3. Exercise: Repeat the analysis of RNM using Gumbel noise.

Remark 7.1. The main benefit of ExpMech over RNM is that it is well defined when $\mathcal{Y}$ is not finite.

7.3 Utility Measure of Median and Quantiles

7.3.1 The Utility of Median

Example 7.4. Suppose there are $2n+1$ data points in $[a,b]$, have n of them equal to a and $n+1$ of them equal to b, then the last point can change the median substantially, i.e. changing one value from b to a shows that the sensitivity is $(b-a)$, which is the range!

Unlike sample mean, sample median can be very sensitive! So, instead of adding noise directly to the median, we can express it with a utility function.

Given $X_1,\dots,X_n$, what is a good $u(X,m)$ such that $\underset{m}{\operatorname{argmax}}\, u(X,m) = \operatorname{Median}(X)$?

- One option is the standard loss function for median

$$\sum_{i=1}^{n} |X_i - m|,$$

  which does not work well.

- Alternatively, we can use

$$u(X,m) = \left|\frac{1}{2} - \widehat{F}_n(m)\right|$$

  where $\widehat{F}_n(m) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} I(X_i \le m)$ the empirical CDF. We will see that $u(X,m)$ works better than the standard loss function.

In general, the α-quantile is $Q_\alpha(\underline{X}) = \operatorname{argmin}_t \big|n\alpha - \#\{i \mid X_i \le t\}\big|$. Write $\operatorname{Median}(\underline{X}) = \underset{t}{\operatorname{argmin}}\left|\dfrac{n}{2} - \#\{i \mid X_i \le t\}\right|$. To check its sensitivity, we consider X and X' that only differ in the i-th entry, then

$$\begin{aligned}
&\left|\big|n\alpha - \#\{i \mid X_i \le t\}\big| - \Big|n\alpha - \#\big\{i \mid X_i' \le t\big\}\Big|\right| \\
\le &\Big|n\alpha - \#\{i \mid X_i \le t\} - n\alpha - \#\big\{i \mid X_i' \le t\big\}\Big| \\
\le &1
\end{aligned}$$

as counts have sensitivity 1. This gives $\Delta u = 1$. So, ExpMech draws t according to the density proportional to

$$\exp\left(-\frac{\varepsilon}{2}\big|n\alpha - \#\{i \mid X_i \le t\}\big|\right),$$

but this is not a well-defined density as this expression is not integrable. So, we need a non-trivial base measure such as uniform on an interval or a "prior" distribution on t.

7.3.2 The Utility of Quantile

Lemma 4. Suppose $(X_n)_{n=1}^{\infty}$ is a random vector and $\mathbb{E}\big[|X_n|\big] < \infty$, then $X_n = O_p(\mathbb{E}\big[|X_n|\big])$.

Proof of Lemma 4. By Markov inequality, for some constant γ,

$$\Pr[|X_n| \geq \varepsilon] \leq \frac{\mathbb{E}\big[|X_n|\big]}{\varepsilon} \implies \Pr\left[\frac{|X_n|}{\mathbb{E}\big[|X_n|\big]} \geq \frac{1}{\gamma}\right] \leq \gamma$$

■

Lemma 5. Let $U_1, \cdots, U_n \overset{\text{i.i.d.}}{\sim} U(0,1)$, then $U_{(s)} - U_{(r)} = O_p\left(\frac{s-r}{n}\right)$.

Proof of Lemma 5. Since $U_{(s)}, U_{(r)} \overset{\text{i.i.d.}}{\sim} U(0,1)$, we have $U_{(s)} - U_{(r)} \sim \text{Beta}(s-r, n-s+r+1)$ and hence

$$\mathbb{E}\big[U_{(s)} - U_{(r)}\big] = \frac{s-r}{(s-r)+(n-s+r+1)} = \frac{s-r}{n+1} = O\left(\frac{s-r}{n}\right)$$

So by Lemma 4, we conclude $U_{(s)} - U_{(r)} = O_p\left(\frac{s-r}{n}\right)$. ■

Theorem 7.6. Let $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$, where F is a continuous CDF with PDF f. Choose a quantile level $\alpha \in (0,1)$ s.t. $f > 0$ and f continuous in a neighborhood around $F^{-1}(\alpha)$. And let

$$T \sim \exp\left(-\frac{\varepsilon n}{2}\left|\alpha - \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq T)\right|\right) I(a \leq T \leq b)$$

be the output of the exponential mechanism where it is assumed that $F^{-1}(\alpha) \in [a,b]$. Let the width $\Lambda = b - a$ and assume that T is sampled with accuracy of d decimal values and that $\Lambda \geq 3/10^d$, then

$$T - X_{(n\alpha)}\big|_{X_1,\cdots,X_n} = O_p\left(\frac{d + \log\Lambda}{\varepsilon n f(F^{-1}(\alpha))}\right)$$

where $X_{(n\alpha)}$ is the $n\alpha$-th[a] order statistics and is the α-quantile of $X_1, \cdots, X_n$.

[a] May need to round, but the result does not change.

Proof of Theorem 7.6. The number of candidates for the exponential mechanism is $10^d\Lambda \geq 3$. For simplicity, we assume $\alpha n \in \mathbb{Z}$, then for any given X,

$$\Pr\left[-n\left|\alpha - \widehat{F}(T)\right| \leq 0 - \frac{2}{\varepsilon}\Big(\log(10^d\Lambda) + t\Big) \;\Big|\; X\right] \leq e^{-t}$$

where $\widehat{F}(t) := \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t)$. Since $\log(10^d\Lambda) = d\log(10) + \log\Lambda \ge 1$, the inequality above becomes

$$e^{-t} \ge \Pr\left[\frac{\varepsilon n\left|\alpha - \widehat{F}(T)\right|}{d\log(10)+\log\Lambda} \ge 2\left(\frac{d\log(10)+\log\Lambda+t}{d\log(10)+\log\Lambda}\right) \;\middle|\; X\right]$$
$$\ge \Pr\left[\frac{\varepsilon n\left|\alpha - \widehat{F}(T)\right|}{d\log(10)+\log\Lambda} \ge 2(1+t) \;\middle|\; X\right].$$

Let $\gamma := e^{-t}$ with $t = \log(1/\gamma)$, then

$$\Pr\left[\frac{\varepsilon n\left|\alpha - \widehat{F}(t)\right|}{d\log(10)+\log\Lambda} \ge 2\left(1+\log\frac{1}{\gamma}\right) \;\middle|\; X\right] \le \gamma.$$

Taking expectation over X on both sides to get marginal probability bound, we derive that

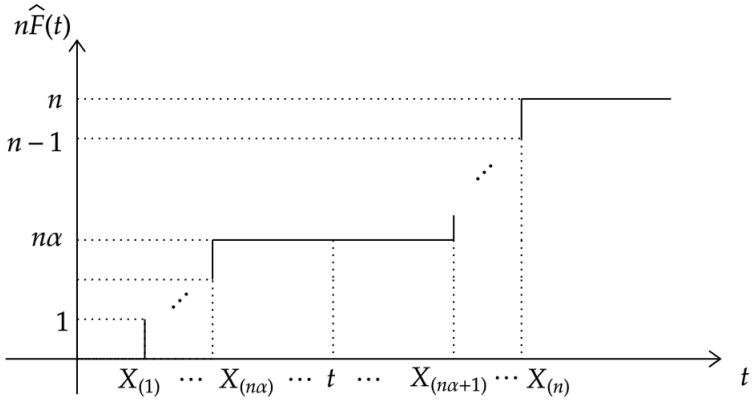
$$\frac{\varepsilon n\left|\alpha - \widehat{F}(T)\right|}{d\log(10)+\log\Lambda} = O_p(1),$$

which implies $\alpha - \widehat{F}(T)\Big|_X = O_p\left(\frac{d+\log\Lambda}{\varepsilon n}\right)$.

Remark 7.2. Since $n\alpha$ is the index of the order statistics for the α-quantile, we can also write

$$n\alpha - n\widehat{F}(t) = O_p\left(\frac{d+\log\Lambda}{\varepsilon}\right)$$

The graph of $n\widehat{F}(t) = \sum_{i=1}^{n} I(X_i \le t)$ would look a lot like a "staircase" that increases by 1 at each $X_{(i)}$'s.



We will use this figure to develop some intuition. Denote $\mathcal{O} := O_p\left(\frac{d+\log\Lambda}{\varepsilon}\right)$ for simplicity, then

$$T = X_{(n\alpha+\mathcal{O})} + \text{small error}$$

but we want to understand the difference of order statistics, since

$$T - X_{(n\alpha)} \approx X_{(n\alpha+\mathcal{O})} - X_{(n\alpha)}.$$

Now back to the proof. Becasue of the i.i.d. assumption, we can write $X_{(s)} = F^{-1}(U_{(s)})$, where $U_{(s)}$ is the order statistic from $U_1, \ldots, U_n \overset{i.i.d.}{\to} U(0,1)$. Since the quantile function is a monotone increasing function,

$$\begin{aligned} X_{(s)} - X_{(r)} &= F^{-1}(U_{(s)}) - F^{-1}(U_{(r)}) \\ &= F^{-1}\left(U_{(s)} + O_p\left(\frac{s-r}{n}\right)\right) - \left(F^{-1}(U_{(r)})\right) \\ &= \underbrace{\frac{1}{f(F^{-1}(U_{(r)}))}}_{\text{Derivative of } F^{-1} \text{ at } U_{(r)}} O_p\left(\frac{s-r}{n}\right) + \underbrace{o_p\left(\frac{s-r}{n}\right)}_{\text{negligible}} \end{aligned}$$

by Taylor Approximation. We want to put the fraction into the O_p rate, so we need to argue that it converges to some constant. In our case, we will use $r = n\alpha$ and $s - r = O_p\left(\frac{d+\log\Lambda}{\varepsilon}\right)$. Thus, by Slutsky's theorem and $U_{(r)} \overset{d}{\to} \alpha$, we get

$$\frac{1}{f(F^{-1}(U_{(r)}))} \overset{d}{\to} \frac{1}{f(F^{-1}(\alpha))} = O_p\left(\frac{1}{f(F^{-1}(\alpha))}\right).$$

Therefore,

$$X_{(s)} - X_{(r)} = O_p\left(\frac{s-r}{nf(F^{-1}(\alpha))}\right).$$

Thus,

$$T - X_{(n\alpha)} = \left(T - X_{(n\widehat{F}(T))}\right) + \left(X_{(n\widehat{F}(T))} - X_{(n\alpha)}\right). \tag{7.7}$$

For the second term on the RHS is

$$X_{(n\widehat{F}(T))} - X_{(n\alpha)} = O_p\left(\frac{d+\log\Lambda}{\varepsilon n f(F^{-1}(\alpha))}\right). \tag{7.8}$$

As for the first term on the RHS, since $X_{(n\alpha)} \le T < X_{(n\widehat{F}(T)+1)}$, and both $X_{(n\widehat{F}(T)+1)}$ and $X_{(n\widehat{F}(T))}$ converge to $F^{-1}(\alpha)$ (because $\frac{n\widehat{F}(T)+\text{constant}}{n} \to \alpha$ as $n \to \infty$), the error

$$T - X_{(n\widehat{F}(T))} \le X_{(n\widehat{F}(T))+1)} - X_{(n\widehat{F}(T))} = O_p\left(\frac{1}{nf(F^{-1}(\alpha))}\right) \tag{7.9}$$

by Lemma 5.

We can see that the term in Eq. (7.9) is dominated by that of Eq. (7.8). So combining them into Eq. (7.7), we get

$$T - X_{(n\alpha)}\big|_X = O_p\left(\frac{d + \log \Lambda}{n\varepsilon f(X_{(n\alpha)})}\right).$$

Note that the error is lower when $X_{(n\alpha)} \approx F^{-1}(\alpha)$ is near a mode of f (for unimodal, $\alpha = 1/2$). When $F^{-1}(\alpha)$ is at a low value of f (tail or where there is little support), the error increases. ■

7.4 Exponential Mechanism for Empirical Risk Minimization

7.4.1 Motivation

Q: How should we go about designing utility/risk/loss function?

We are given $X_1, \cdots, X_n$ i.i.d. from an unknown distribution F. We have a real-valued loss function $L(\theta, X)$ and ideally, we want to find

$$\theta^* = \operatorname*{argmin}_\theta \mathop{\mathbb{E}}_{X\sim F}\big[L(\theta, X)\big]$$

Since all we have are $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$, we instead solve

$$\widehat{\theta} = \operatorname*{argmin}_\theta \underbrace{\frac{1}{n}\sum_{i=1}^n L(\theta, X_i)}_{\text{empirical risk}}.$$

It is important that $\frac{1}{n}\sum L(\theta, X_i)$ is a sum over the X_i's.

Example 7.5. Applications in statistics and machine learning that fit in this framework:

- Maximum Likelihood Estimation (MLE) / Maximum A-Posterior (MAP)
- Support Vector Machine
- Neural Network
- Linear, logistic, quantile regression (GLMs)
- M estimators

7.4.2 Empirical Risk Minimization (ERM)

The sensitivity of an empirical risk function is

$$\begin{aligned}\Delta(n) &:= \sup_{H(X,X')\le 1} \sup_{\theta} \left| \frac{1}{n}\sum_{i=1}^{n} L(\theta, X_i) - \frac{1}{n}\sum_{i=1}^{n} L(\theta, X_i') \right| \\ &= \sup_{X_i, X_i'} \sup_{\theta} \frac{1}{n} \left| L(\theta, X_i) - L(\theta, X_i') \right|\end{aligned}$$

We may omit the prefactor of $1/n$ so that the sensitivity $\Delta := n\Delta(n)$ is a constant not involving n and consider its exponential mechanism

$$\widetilde{\theta} \sim \alpha \exp\left(-\frac{\varepsilon}{2\Delta} \sum_{i=1}^{n} L(\theta, X_i) \right)$$

which satisfies ε-DP. But its utility needs to be analyzed.

Remark 7.3 (α-Strongly Convex)**.** The condition (1) in Theorem 7.10 can be seen as a convex function being globally lower bounded by some quadratic function. For example, absolute value functions are not strongly convex.

Theorem 7.10 (Awan et al. [2019])**.** Let $X_1, X_2, \ldots$ be a sequence of data points. Call $L_n(\theta) = \sum_{i=1}^{n} L(\theta, X_i)$, which satisfy:

(1) (α-Strongly Convex) $\frac{1}{n} L_n(\theta)$ are twice differentiable and convex, and there exists $\alpha > 0$ such that the eigenvalues of $\frac{1}{n} L_n''(\theta)$ are greater than α for all n and θ.

(2) Let $\widehat{\theta}_n = \operatorname{argmin}_{\theta} L_n(\theta)$ and assume $\widehat{\theta}_n \to \theta^*$ and $\frac{1}{n} L_n''(\widehat{\theta}) \to \Sigma^{-1}$, where Σ is positive definite.

(3) L_n has sensitivity Δ constant in n.

(4) Base measure g is bounded, positive, and continuous in a neighborhood of θ^*.

Then $\theta \sim \alpha \exp\left(-\frac{\varepsilon}{2\Delta} L_n(\theta) \right) g(\theta)$ satisfies

$$\sqrt{n}(\theta - \widehat{\theta}) \xrightarrow{d} \mathcal{N}\left(0, \frac{2\Delta}{\varepsilon}\Sigma\right)$$

if $g = 1$. That is, the noise due to privacy is $O_p\left(\frac{1}{\sqrt{n}}\right) = O_p\left(\sqrt{\frac{\Delta}{n\varepsilon}}\right)$, the same as the statistical estimation error (Pitfalls!)

Proof of Theorem 7.10. The density of the exponential mechanism is

$$f_n(\theta) = c_n^{-1} g(\theta) \exp\left(-\frac{\varepsilon}{2\Delta} L_n(\theta)\right)$$

where c_n is the normalizing constant. Now call $z = \sqrt{n}(\theta - \widehat{\theta})$. Its density is

$$f_n(z) = c_n^{-1} n^{-1/2} g\left(\widehat{\theta} + \frac{z}{\sqrt{n}}\right) \exp\left(-\frac{\varepsilon}{2\Delta} L_n\left(\widehat{\theta} + \frac{z}{\sqrt{n}}\right)\right).$$

We will show that $f_n(z)$ converge to a multivariate normal.

By the assumptions and Taylor expansion

$$L_n\left(\widehat{\theta} + \frac{z}{\sqrt{n}}\right) = L_n(\widehat{\theta}) + \underbrace{z^\top \frac{L_n'(\widehat{\theta})}{\sqrt{n}}}_{=0 \text{ by assumption (2)}} + z^\top L_n''(\widehat{\theta}) \frac{z}{2n} + \underbrace{o(1)}_{\substack{L_n'' \text{ is greater} \\ \text{than } \alpha}}$$

The last term $o(1)$ there is basically another equivalent way of representing $O(1/\sqrt{n})$. The first term does not depend on z, so it can be absorbed into the constant. So, only the third term appears in the density, and by assumption (2), $\frac{1}{n} L_n''(\widehat{\theta}) \to \Sigma^{-1}$.

Now for other terms in $f_n(z)$, note first that

$$\left| g\left(\widehat{\theta} + \frac{z}{\sqrt{n}}\right) - g(\theta^*) \right| \to 0$$

Next, we can check the integrating constant

$$\begin{aligned}\left(\frac{1}{\text{integrating constant}}\right) &= c_n n^{-\frac{1}{2}} \exp\left(\frac{\varepsilon}{2\Delta} L_n(\widehat{\theta})\right) \\ &= \int g\left(\widehat{\theta} + \frac{z}{\sqrt{n}}\right) \exp\left(-\frac{\varepsilon}{2\Delta}\left(L_n\left(\widehat{\theta} + \frac{z}{\sqrt{n}}\right) - L_n(\widehat{\theta})\right)\right) dz\end{aligned}$$

By α-strong convexity,

$$-\left(L_n(\widehat{\theta} + \frac{z}{\sqrt{n}}) - L_n(\widehat{\theta})\right) \leq -\frac{n\alpha}{2} \left\| \frac{z}{\sqrt{n}} \right\|^2 = -\frac{\alpha \|z\|^2}{2}.$$

Since $\exp\left(-\frac{\alpha \|z\|^2}{2}\right)$ is integrable, g is bounded in a neighborhood of θ^*. So, by the dominated convergence theorem, the integrating constant converges, and so

$$f_n(z) \to f(z) \propto g(z) \exp\left(-\frac{\varepsilon}{2\Delta} \frac{z^\top \Sigma^{-1} z}{2}\right),$$

which, if $g = 1$, is the density of a multivariate normal. By Scheffe's theorem, we conclude, if $g = 1$,

$$\sqrt{n}(\widetilde{\theta} - \widehat{\theta}) \xrightarrow{d} \mathcal{N}\left(0, \frac{2\Delta}{\varepsilon}\Sigma\right).$$

■

Remark 7.4 (Important Conclusion of Theorem 7.10)**.** We see that when exponential mechanism is applied to a strongly convex empirical risk, the resulting mechanism introduce $O_p(1/\sqrt{n})$ noise, rather than the ideal $o_p(1/\sqrt{n})$.

7.4.3 One-Dimensional Illustration for ExpMech Asymptotics in ERM

For θ close to $\widehat{\theta}(X)$,

$$L(\theta; X) \approx b_n(\theta - \widehat{\theta}(X))^2 + a_n,$$

where $b_n \to b$ and $a_n \to a$. By the utility result for ExpMech,

$$L(\widetilde{\theta}; X) - L(\widehat{\theta}; X) = O_p(\frac{\Delta}{\epsilon n}),$$

where Δ is the sensitivity of $l(\theta; X_i)$ and $L(\theta; X) = \frac{1}{n}\sum_{i=1}^{n} l(\theta; X_i)$.

Thus,

$$L(\widetilde{\theta}; X) - L(\widehat{\theta}; X) = b_n(\theta - \widehat{\theta}(X))^2 = O_p(\frac{\Delta}{\epsilon n})$$

which implies $|\widetilde{\theta} - \widehat{\theta}(X)| = O_p(\sqrt{\frac{\Delta}{\epsilon n}})$. This is the rate definded in the previous theroem.

Remark 7.5. In order to improve the rate, it must <u>not</u> be possible to approximate $L(\theta; X)$ as a quadratic, i.e. $L(\theta; X)$ must <u>not</u> be twice differentiable at $\widehat{\theta}(X)$.

Example 7.6 (Linear Regression)**.** For each individual, we observe (X_i, y_i) where $X_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ which are modeled as

$$y_i = X_i^\top \theta + e_i$$

where e_i are i.i.d. mean zero and uncorrelated. Assume

(1) $-1 \le X_{ij} \le 1$ ($\implies \|X_i\|_\infty \le 1$) and $-1 \le y_i \le 1$.

(2) $\|\theta\|_1 \le B$.

The loss function is least squares:

$$L(\theta, D) = \sum_{i=1}^{n}(y_i - X_i^\top \theta)^2$$

For sensitivity, say D and D' differ in i-th entry, and the values are (X_1, y_1) and (X_2, y_2) respectively.

$$\begin{aligned}
\left|L(\theta, D) - L(\theta, D')\right| &= \left|(y_1 - X_1^\top\theta)^2 - (y_2 - X_2^\top\theta)^2\right| \\
&\le \sup_{X_1, y_1, \theta} \left(y_1 - X_1^\top\theta\right)^2 && (\because (y_2 - X_2^\top\theta)^2 \ge 0) \\
&\le \sup \left(1 + \left|X_1^\top\theta\right|\right)^2 && (\because |y_1| \le 1) \\
&\le \sup_\theta \left(1 + \|\theta\|_1\right)^2 && (\because \|X_1\|_\infty \le 1) \\
&\le (1 + B)^2 < \infty && (\because \|\theta\|_1 \le B)
\end{aligned}$$

So, we can employ the exponential mechanism

$$\widetilde{\theta} \sim \alpha \exp\left(-\frac{\varepsilon}{2(1+B)^2} \sum_{i=1}^n \left(y_i - X_i^\top\theta\right)^2\right)$$

w.r.t. the uniform measure on $\left\{\theta \,\middle|\, \|\theta\|_1 \le B\right\}$. However, $L(\theta, D)$ is strongly convex (and hence twice differentiable) unless $XX^\top$ is degenerate. Thus, Theorem 7.10 sugggests that the noise due to privacy is $O_p\left(\frac{1}{\sqrt{n}}\right)$.

7.5 K-Norm Gradient Mechanism

The main reference for this section is Reimherr and Awan [2019].

We saw that if the loss function L in the exponential mechanism is twice differentiable, then approximating L with a 2-term Taylor expansion results in

$$L_n(\theta) = L_n(\widehat{\theta}) + \underbrace{L_n'(\widehat{\theta})}_{=0}(\theta - \widehat{\theta}) + (\theta - \widehat{\theta})^\top L_n''(\widehat{\theta})(\theta - \widehat{\theta}) + \text{error}$$

Then,

$$\begin{aligned}
f_n(\theta) &\propto \exp\left(-\frac{\varepsilon n}{2\Delta} L_n(\theta)\right) \\
&= \exp\left(-\frac{\varepsilon n}{2\Delta}\left(L_n(\widehat{\theta}) + (\theta - \widehat{\theta})^\top \underbrace{L_n''}_{\to \Sigma^{-1}}(\widehat{\theta})(\theta - \widehat{\theta})\right)\right) + \text{error}
\end{aligned}$$

and the n in front becomes incorporated in the variance of the Gaussian, giving variance $\frac{2\Delta}{n}\Sigma$.

Remark 7.6. The idea was that we alter the loss function so that it can no longer be approximated as a quadratic function such as an absolute value function (e.g., $f(x) = |x|$ cannot be approximated with a 2-term Taylor expansion at its minimizer). The K-Norm gradient mechanism (kNG) applies the exponential mechanism to the altered loss function $\left\|\nabla L_n(\theta, X)\right\|$ where $\|\cdot\|$ is any norm in $\mathbb{R}^k$.

Theorem 7.11. Let $\{L_n(\theta \mid X) \mid X \in \mathcal{X}^n\}$ be a collection of measurable functions, which are differentiable w.r.t θ almost everywhere. Assume that

$$\left\|\nabla L_n(\theta \mid X) - \nabla L_n(\theta \mid X')\right\| \leq \Delta < \infty$$

for all $H(X, X') \leq 1$ and almost all θ. Then the density (with respect to a base measure)

$$\propto \exp\left(-\frac{\varepsilon}{2\Delta}\left\|\nabla L(\theta \mid X)\right\|\right)$$

satisfies ε-DP.

Proof of Theorem 7.11. Let $\widetilde{\mathcal{L}}(\theta \mid X) = \left\|\nabla L(\theta \mid X)\right\|$, then $\widetilde{\mathcal{L}}$ has sensitivity Δ by the triangle inequality. The result follows by ExpMech. ■

Remark 7.7. kNG is similar to Objective Perturbation without regularizer and it also removes the determinant appearing from the change of variable. This allows less strict assumptions on the loss.

Theorem 7.12. Let $L_n(\theta) = L_n(\theta \mid X)$ be a sequence of loss functions satisfying the assumptions of the previous result (Theorem 7.11) with sensitivity Δ. Assume further that

1. $\frac{1}{n}L_n(\theta)$ are twice differentiable (almost everwhere) and α-strongly convex functions.
2. The minimizers satisfy $\widehat{\theta} \to \theta^*$ and $\frac{1}{n}\nabla^2 L_n(\widehat{\theta}) \to \Sigma^{-1}$ where Σ is positive definite.
3. Assume the base measure is Lebesgue.

Let $\widetilde{\theta}$ be the sample drawn from kNG with privacy parameter ε, then

$$z = n(\widetilde{\theta} - \widehat{\theta}) \xrightarrow{d} \propto \exp\left(-\frac{\varepsilon}{2\Delta}\left\|\Sigma^{-1}z\right\|\right),$$

the K-norm distribution. In particular, $\widetilde{\theta} - \widehat{\theta} = \Theta_p\left(\frac{\Delta}{\varepsilon n}\right)$

Proof of Theorem 7.12. The density of kNG is

$$f_n(\widetilde{\theta}) = c_n^{-1}\exp\left(-\frac{\varepsilon}{2\Delta}\left\|\nabla L_n(\widetilde{\theta})\right\|\right)$$

Define $z = n(\widetilde{\theta} - \widehat{\theta})$, which has the density

$$f_n(z) = c_n^{-1}n^{-1}\exp\left(-\frac{\varepsilon}{2\Delta}\left\|\nabla L_n\left(\widehat{\theta} + \frac{z}{n}\right)\right\|\right)$$

By assumption 2, we can write

$$\begin{aligned}\nabla L_n\left(\widehat{\theta}+\frac{z}{n}\right) &= \underbrace{\nabla L_n(\widehat{\theta})}_{=0}+\nabla^2 L_n(\widehat{\theta})\frac{z}{n}+o_p(1)\\ &= \nabla^2 L_n(\widehat{\theta})\frac{z}{n}+o_p(1)\\ &\to \Sigma^{-1}z+o_p(1)\end{aligned}$$

Note that

$$c_n n = \int \exp\left(-\frac{\varepsilon}{2\Delta}\left\|\nabla L_n\left(\widehat{\theta}+\frac{z}{n}\right)\right\|\right)dz$$

By assumption 1, L_n is strongly convex, so

$$\left\langle \nabla L_n\left(\widehat{\theta}+\frac{z}{n}\right)-\underbrace{\nabla L_n(\widehat{\theta})}_{=0},\frac{z}{n}\right\rangle \geq n\alpha\left\|\frac{z}{n}\right\|_2^2$$

Then by Cauchy-Schwarz,

$$\left\|\nabla L_n\left(\widehat{\theta}+\frac{z}{n}\right)\right\|_2 \geq n\alpha\left\|\frac{z}{n}\right\|_2$$

By the equivalence of norms on $\mathbb{R}^d$, we have

$$-\left\|\nabla L_n\left(\widehat{\theta}+\frac{z}{n}\right)\right\| \leq -c\alpha\|z\|_2$$

for some constant c (it may depend on dimension d). Since $\exp\left(-\frac{\varepsilon c\alpha}{2\Delta}\|z\|_2\right)$ is integrable, by the dominated convergence theorem, the constants $c_n n$ converge to a non-zero finite quantity. So, the density converges

$$f_n(z)\to f(z)\propto \exp\left(-\frac{\varepsilon}{2\Delta}\left\|\Sigma^{-1}z\right\|\right)$$

■

Remark 7.8. If we replace the base measure with some probability measure, we can drop strong convexity, resulting in unique solution to some equation. The base measure will appear in the final result.

Example 7.7 (Linear Regression)**.** Observe (X_i, y_i) where $X_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Set

$$y_i = X_i^\top \theta + e_i$$

where $e_i \overset{\text{i.i.d.}}{\sim}$ mean zero, uncorrelated. Assume for all $i = 1, \cdots, n$ and $j = 1, \cdots, d$,

$$\begin{aligned}-1 \leq X_{ij} \leq 1\\ -1 \leq y_i \leq 1\\ \|\theta^*\|_1 \leq B,\end{aligned}$$

where θ^* is the true value. For least squares, we set $L(\theta \mid D) = \sum_{i=1}^{n}(y_i - X_i^\top\theta)^2$. kNG requires a bound on the gradient's sensitivity

$$\begin{aligned}\left\|\nabla L_n(\theta \mid D) - \nabla L_n(\theta \mid D')\right\| &\leq \sup_{y_1, X_1, \theta} 4\left\|(y_1 - X_1^\top\theta)X_1\right\| \\ &\leq \sup_{X_1} 4(1+B)\|X_1\|\end{aligned}$$

by using $\|\cdot\|_\infty$ and the fact that $\|X_1\|_\infty \leq 1$, thereby giving $\Delta = 4(1+B)$. Then kNG samples from

$$f_n(\theta) \propto \exp\left(-\frac{\varepsilon}{8(1+B)}\left\|\sum_{i=1}^{n}(y_i - X_i^\top\theta)X_i^\top\right\|_\infty\right)$$

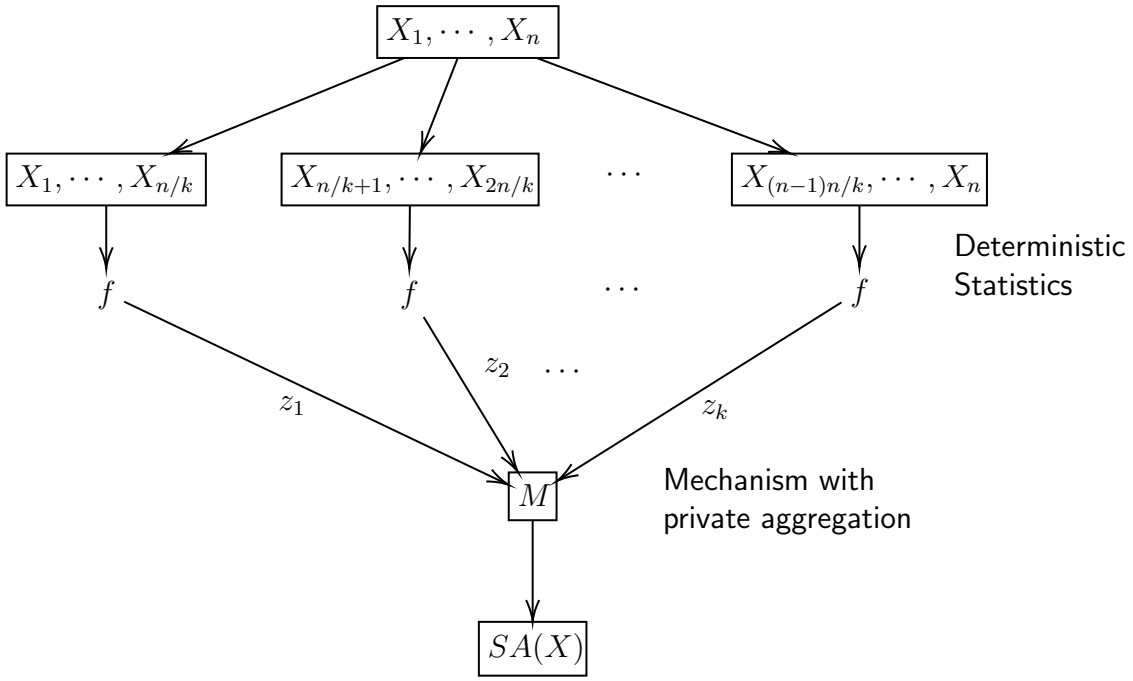
w.r.t the uniform measure on $\Theta = \left\{\theta \,\middle|\, \|\theta\|_1 \leq B\right\}$. Then, Theorem 7.12 says that the noise due to kNG is only $\Theta_p\left(\frac{1}{n}\right)$.

8 Subsample and Aggregate

The main reference for this chapter is Smith [2011].

What to do when we have a function that has arbitrary or difficult to analyze sensitivity? We would still want to use a function that is "usually" insensitive in practice. This is also referred to be the Black Box approach.

8.1 Algorithm of Subsample and Aggregate



In Subsample and Aggregate [Smith, 2011], the n rows of X are partitioned into k blocks $B_1, \cdots, B_k$, each of size $\approx n/k$. The function f is computed exactly on each block. Then the intermediate results

$$(z_1, \cdots, z_k) = (f(B_1), \cdots, f(B_k))$$

are combined in a DP aggregation mechanism (e.g. DP version of α-trimmed, winsorized mean and median).

Remark 8.1 (Key Observation)**.**

- One person can affect only one block and therefore only one $f(B_i)$. Even if f is arbitrary, the analyst can choose a DP aggression algorithm (independent of database).
- Privacy is easy: if the aggregation mechanism M is (ε, δ)-DP (change on one $f(B_i)$), then so is $SA(X)$. However, in general, it can be difficult to analyze the utility of $SA(X)$, since we still need some type of worst case sensitivity calculation.

- With iid data and "generally normal" Statistics, privacy can be achieved at no asymptotic cost.

Theorem 8.1 (Privacy Preserving Statistical Estimation with Optimal Convergence Rates: Smith [2011])**.** Informally: if $X_1, \cdots, X_n$ are i.i.d. and if $T(X_1, \cdots, X_n)$, appropriately rescaled and converges to a normal distribution as $n \to \infty$, then one can design a DP mechanism $M_T(X_1, \ldots, X_n)$ which converges to the same asymptotic distribution.

8.2 Asymptotic Analysis

Definition 8.1 (Generic Asymptotic Normality)**.** A statistic $T\colon \mathcal{X}^n \to \mathbb{R}$ is generally asymptotically normal at distribution P if there exists $T(P)$ and variance $\sigma_p^2 > 0$ such that if $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P$,

1. **Normality**: $\dfrac{T(X) - T(P)}{\sigma_p/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$.
2. **Linear Bias**: $\mathbb{E}\big[T(X) - T(P)\big] = O(1/n)$.
3. **Bounded Third Moment**: $\mathbb{E}\dfrac{|T(X) - T(P)|^3}{\sigma_p/\sqrt{n}} = O(1)$

Remark 8.2.

- We can generalize Definition 8.1 to $\mathbb{R}^d$ $(d > 1)$, by replacing σ_p^2 with a positive definite matrix Σ_p and updating the conditions analogously.
- Properties 2 and 3 of Definition 8.1 ensure that functional of $T(X)$, such as the mean and variance converge to what we want. For example, MLE, Sample mean of function, Regression estimators, m-estimators, etc.

Theorem 8.2. Given $T\colon \mathcal{X}^n \to \left[-\dfrac{\Lambda}{2}, \dfrac{\Lambda}{2}\right]^d$, there exists an ε-DP mechanism $M_{T,\varepsilon,\Lambda}$ such that if T is generally asymptotically normal at P, then

$$\sqrt{n}\Sigma_p^{-\frac{1}{2}}\big(T(X) - T(P)\big) - \sqrt{n}\Sigma_p^{-\frac{1}{2}}\big(M(X) - T(P)\big) \xrightarrow{d} 0$$

i.e., $M(X)$ has the same asymptotic distribution as $T(X)$. This also implies

$$M(X) - T(X) = o_p\left(\frac{1}{\sqrt{n}}\right)$$

Furthermore, there exists $c > 0$ such that convergence still holds if d, ε, Λ change with n, so long as $d, 1/\varepsilon$, and $\log \Lambda$ are all at most n^c.

Remark 8.3. If no Λ is known, we can change the result to (ε, δ)-DP.

Algorithm 1: The algorithm for Theorem 8.2

Input: $X = (X_1, \cdots, X_n)$ and function $T\colon \mathcal{X}^n \to \mathbb{R}^d$.

Set $k = n^{\frac{1}{2}-\eta}$ where $\eta = 1/10$ /* so $k = o(\sqrt{n})$ */

Randomly divide X into k blocks $X^{(1)}, \cdots, X^{(k)}$ of size n/k.

Compute $z_i = T(X^{(i)})$ for each block.

for *each dimension* $j = 1, \cdots, d$ **do**

  Call $z|_j$ the projection of $z = (z_1, \cdots, z_k)$.

  Call $M_j = W(z|_j, \varepsilon/d)$ where W is the noisy widened winsorized mean (DP step).

end for

Output: $M = (M_1, \cdots, M_d)$

Algorithm 2: The algorithm for Noisy Widened Winsorized Mean

Input: $z = (z_1, \cdots, z_k) \in \mathbb{R}^k$, $\varepsilon > 0$, $\Lambda > 0$ (will clamp to $[0, \Lambda]$).

Set $\mathrm{rad} = k^{\frac{1}{3}+\eta}$ where $\eta = 1/10$ /* so $\mathrm{rad} \approx \Omega(k^{1/3})$ */

/* First estimate the range of z, $[l, u]$, and use exp. mech. to estimate quantiles */

/* PrivateQuantile is an exponential mechanism for a quantile. z is for values, $1/4$ is for the target quantile, $\varepsilon/4$ is for privacy budget, and Λ is for bound. */

$\widehat{a} = \mathrm{PrivateQuantile}\left(z, \frac{1}{4}, \frac{\varepsilon}{4}, \Lambda\right)$ and $\widehat{b} = \mathrm{PrivateQuantile}\left(z, \frac{3}{4}, \frac{\varepsilon}{4}, \Lambda\right)$

$M_{\mathrm{crude}} = \dfrac{\widehat{a} + \widehat{b}}{2}$ and $\mathrm{Iqr}_{\mathrm{crude}} = |\widehat{b} - \widehat{a}|$ /* $\mathrm{Iqr}_{\mathrm{crude}} \approx \Theta\left(\frac{1}{\sqrt{n/k}}\right)$, interquartile range */

$u \leftarrow M_{\mathrm{crude}} + 4 \cdot \mathrm{rad} \cdot \mathrm{Iqr}_{\mathrm{crude}}$

$l \leftarrow M_{\mathrm{crude}} - 4 \cdot \mathrm{rad} \cdot \mathrm{Iqr}_{\mathrm{crude}}$

/* u and l forms an approximate range of z. */

/* Next, compute winsorized/clamped mean for range $[l, u]$ */

Let $\widehat{\mu} = \dfrac{1}{k}\displaystyle\sum_{i=1}^{k} z_i]_l^u$, where $x]_l^u$ is the clamp function.

Sample $L \sim \mathrm{Lap}\left(0, \dfrac{2(u-l)}{k\varepsilon}\right)$ /* sensitivity $= (u-l)/k$, budget $= \varepsilon/2$. */

Output: $W(z) = \widehat{\mu} + L$

Recall that the clamp function was defined to be

$$x]_l^u = \begin{cases} l & \text{if } x < l \\ x & \text{if } l \le x \le u \\ u & \text{if } x > u \end{cases}$$

Algorithm 3: Algorithm for $\mathrm{PrivateQuantile}(z, \alpha, \varepsilon)$

Input: $z = (z_1, \cdots, z_k) \in \mathbb{R}^k$, quantile $\alpha = (0, 1)$, $\varepsilon > 0$, and bound Λ

Output: X drawn from the density proportional to

$$\exp\left(-\frac{\varepsilon}{2}\left|\alpha k - \#\{i \mid z_i \le X\}\right|\right)\mathbb{1}\left(-\frac{\Lambda}{2} \le X \le \frac{\Lambda}{2}\right)$$

Algorithm 4: Sampling Algorithm

Input: $z = (z_1, \cdots, z_k) \in \mathbb{R}^k$, quantile $\alpha = (0, 1)$, $\varepsilon > 0$, and bound Λ

1 Sort z_i in ascending order.

2 Replace $z_i < -\frac{\Lambda}{2}$ with $-\frac{\Lambda}{2}$ and $z_i > \frac{\Lambda}{2}$ with $\frac{\Lambda}{2}$.

3 Define $z_0 = -\frac{\Lambda}{2}$ and $z_{k+1} = \frac{\Lambda}{2}$. **for** $i = 0, \cdots, k$ **do**

4 $\quad y_i = (z_{i+1} - z_i)\exp\left(-\varepsilon|i - \alpha k|\right)$

5 **end for**

6 Sample an integer $i \in \{0, \cdots, k\}$ with probability $\frac{y_i}{\sum_{i=0}^k y_i}$.

Output: a uniform draw from $z_{i+1} - z_i$.

Remark 8.4. Consider the rate of $L \sim \mathrm{Lap}\left(0, \frac{2(u-l)}{k\varepsilon}\right)$ from Algorithm 2,

$$\frac{|u-l|}{k} = O\left(\frac{k^{-(\frac{1}{3}+\eta)}}{k}\right) = O(k^{-1-\frac{1}{3}-\eta})$$

Plugging $k = n^{1/2-\eta}$ with $\eta = 0.1$, we get

$$\frac{|u-l|}{k} = O(n^{-0.57}),$$

which is barely $o(n^{-\frac{1}{2}})$. So the noise does go away asymptotically but slowly.

9 Technique of (ε, δ)-DP Proof

Recall that our (ε, δ)-DP definition relies on the two random variables $M(X)$ and $M(X')$. Let $f_{M(X)}$ and $f_{M(X')}$ be the densities/PMFs of $M(X)$ and $M(X')$ respectively, then

$$L_{M(X)\|M(X')} = \log\left(\frac{f_{M(X)}(y)}{f_{M(X')}(y)}\right),$$

where $y \sim M(X)$, is the privacy loss random variable (PLRV) for X and X' under M.

It is easy to see that M satisfies ε-DP if and only if

$$\Pr\left[L_{M(X)\|M(X')} \leq \varepsilon\right] = 1,$$

for all $H(X, X') \leq 1$. In fact, it can be seen that (ε, δ)-DP can be written in terms of the privacy loss random variable.

9.1 Hockey-Stick Divergence

Definition 9.1. The Hockey-Stick divergence for $\alpha > 0$,

$$H_\alpha(P\|Q) = \mathbb{E}_{y\sim Q}\left[\frac{dP}{dQ}(y) - \alpha\right]_+$$

where $[x]_+ = x\mathbb{1}(x \geq 0)$ and $\frac{dP}{dQ}$ is the Radon-Nikodym derivative (or ratio of densities) of P with respect to Q.

Theorem 9.1. M satisfies (ε, δ)-DP if and only if

$$\sup_{H(X,X')\leq 1} H_{e^\varepsilon}(M(X')\|M(X)) \leq \delta$$

Proof. Let $S := \left\{y \;\middle|\; \frac{dM(X')}{dM(X)}(y) \geq e^\varepsilon\right\}$.

($\Longrightarrow$) We prove for the continuous case. Consider

$$\begin{aligned}
H_{e^\varepsilon}(M(X')\|M(X)) &= \mathbb{E}_{y\sim M(X)}\left[\frac{dM(X')}{dM(X)}(y) - e^\varepsilon\right]_+ \\
&= \int_{\frac{dM(X')}{dM(X)}(y)\geq e^\varepsilon}\left(\frac{dM(X')}{dM(X)}(y) - e^\varepsilon\right)dM(X)(y) \\
&= \int_{\frac{dM(X')}{dM(X)}(y)\geq e^\varepsilon}\left(dM(X')(y) - e^\varepsilon dM(X)(y)\right) \\
&= \Pr[M(X') \in S] - e^\varepsilon \Pr[M(X) \in S] \\
&\leq \delta.
\end{aligned}$$

($\Longleftarrow$) Let T be an arbitrary set. Consider $T_1 = T \cap S \subseteq S$ and $T_2 = T \cap S^c \subseteq S^c$. Then,

$$\begin{aligned}\Pr[M(X') \in T_1] - e^{\varepsilon} \Pr[M(X) \in T_1] &= \int_{T_1} \left(\frac{dM(X')}{dM(X)}(y) - e^{\varepsilon} \right) dM(X)(y) \\ &\leq \int_{S} \left(\frac{dM(X')}{dM(X)}(y) - e^{\varepsilon} \right) dM(X)(y) \\ &= H_{e^{\varepsilon}}(M(X') \| M(X)) \\ &=\leq \delta,\end{aligned}$$

by the Hockey-Stick Divergence. Moreover,

$$\Pr[M(X') \in T_2] = \int_{T_2} dM(X')(y) \leq \int_{T_2} e^{\varepsilon} dM(X)(y) = e^{\varepsilon} \Pr[M(X) \in T_2]$$

Therefore, combining the two inequalities we get

$$\begin{aligned}\Pr[M(X') \in T] &= \Pr[M(X') \in T_1] + \Pr[M(X') \in T_2] \\ &\leq e^{\varepsilon} \Pr[M(X) \in T_1] + \delta + e^{\varepsilon} \Pr[M(X) \in T_2] \\ &= e^{\varepsilon} \Pr[M(X) \in T] + \delta\end{aligned}$$

■

Remark 9.1. Note that the Hockey-Stick divergence only depends on $\dfrac{dM(X')}{dM(X)}(y)$, which is $\exp\left(-L_{M(X') \| M(X)}\right)$, a function of the privacy loss random variables.

9.2 A Lemma for (ε, δ)-DP Proof Technique

Lemma 6. If $L_{M(X') \| M(X)}$ satisfies

$$\Pr[L_{M(X') \| M(X)} \geq \varepsilon] \leq \delta$$

for all $H(X, X') \leq 1$, then M satisfies (ε, δ)-DP.

Proof. Let $R = \left\{ y \;\middle|\; \log\left(\dfrac{f_{M(X)}(y)}{f_{M(X')}(y)} \right) \leq \varepsilon \right\}$ and S be given. Then, we can split S into two regions with R and R^c and consider their probabilities

$$\Pr[M(X) \in S] = \Pr[M(X) \in S \cap R] + \underbrace{\Pr[M(X) \in S \cap R^c]}_{\leq \delta} \tag{9.2}$$

Note that

$$\begin{aligned}\Pr[M(X) \in S \cap R] &= \int_{S \cap R} f_{M(X)}(y) dy \\ &\leq \int_{S \cap R} e^{\varepsilon} f_{M(X')}(y) dy \qquad (\because S \cap R \subseteq R) \\ &= e^{\varepsilon} \Pr[M(X') \in S \cap R]\end{aligned}$$

and $\Pr[M(X) \in S \cap R^c] \leq \delta$. Hence, Eq. (9.2) becomes

$$\begin{aligned}\Pr[M(X) \in S] &= \Pr[M(X) \in S \cap R] + \Pr[M(X) \in S \cap R^c] \\ &\leq e^{\varepsilon} \Pr[M(X') \in S \cap R] + \delta \\ &\leq e^{\varepsilon} \Pr[M(X') \in S] + \delta\end{aligned}$$

■

Remark 9.2. It is a relatively crude bound (i.e. not tight), so it is useful on proving the rate but not so on improving its constant.

9.3 Gaussian Mechanism

Let $f\colon \mathcal{X}^n \to \mathbb{R}^d$ be a d-dimensional function, and let

$$\Delta_2 f = \max_{H(X,X')\leq 1} \left\| f(X) - f(X') \right\|_2$$

be its ℓ_2-sensitivity. The Gaussian mechanism with parameter σ adds independent noise $\sim \mathcal{N}\left(0, \sigma^2\right)$ to each of the d-components of the output of f.

Theorem 9.3. Let $\varepsilon \in (0,1)$ and $\delta \in (0,1]$ for $c^2 > 2\log\left(\frac{1.32}{\delta}\right)$, then the Gaussian mechanism with parameter $\sigma \geq \frac{c\Delta_2 f}{\varepsilon}$ is (ε, δ)-DP.

Remark 9.3. Instead of 1.32, 1.25 is the constant usually used in the literature but there might be an error in Dwork and Roth [2014].

Proof. We will prove the one dimensional case. The PLRV in the worst case is

$$\pm \log\left(\frac{\exp\left(-\frac{y^2}{2\sigma^2}\right)}{\exp\left(-\frac{(y+\Delta f)^2}{2\sigma^2}\right)} \right),$$

where $y \sim \mathcal{N}\left(0, \sigma^2\right)$.

We will take the absolute value. The goal is to find a set S with probability (at least) $1-\delta$ such that $\left| \log\left(\frac{\exp\left(-\frac{y^2}{2\sigma^2}\right)}{\exp\left(-\frac{(y+\Delta f)^2}{2\sigma^2}\right)} \right) \right| \leq \varepsilon$. Consider

$$\left| \log\left(\frac{\exp\left(-\frac{y^2}{2\sigma^2}\right)}{\exp\left(-\frac{(y+\Delta f)^2}{2\sigma^2}\right)} \right) \right| = \left| -\frac{1}{2\sigma^2}(y^2 - (y+\Delta f)^2) \right| = \left| -\frac{1}{2\sigma^2}(2y\Delta f + (\Delta f)^2) \right|,$$

which is bounded by ε when $|y| < \frac{\sigma^2\varepsilon}{\Delta f} - \frac{\Delta f}{2}$.

We want to show $\Pr\left[|y| \geq \frac{\sigma^2\varepsilon}{\Delta f} - \frac{\Delta f}{2}\right] < \delta$ when $y \sim \mathcal{N}\left(0, \sigma^2\right)$. It suffices to find σ^2 such that

$$\Pr\left[y \geq \frac{\sigma^2\varepsilon}{\Delta f} - \frac{\Delta f}{2}\right] < \frac{\delta}{2}.$$

We assume that $\varepsilon \leq 1 \leq \Delta f$. Recall that

$$\Pr[X > t] \leq \frac{\sigma}{\sqrt{2\pi}} \frac{1}{t} e^{-\frac{t^2}{2\sigma^2}}$$

which holds because

$$\Pr[X > t] = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-\frac{x^2}{2}} dx \leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \frac{1}{t} e^{-\frac{t^2}{2}}$$

and we want

$$\frac{\sigma}{\sqrt{2\pi}} \frac{1}{t} e^{-\frac{t^2}{2\sigma^2}} < \frac{\delta}{2},$$

where $t = \frac{\sigma^2\varepsilon}{\Delta f} - \frac{\Delta f}{2}$. In other words,

$$\begin{aligned}
\frac{\sigma}{\sqrt{2\pi}} \frac{1}{t} e^{-\frac{t^2}{2\sigma^2}} < \frac{\delta}{2} &\iff \frac{\sigma}{t} e^{-\frac{t^2}{2\sigma^2}} < \sqrt{2\pi}\frac{\delta}{2} \\
&\iff \log\left(\frac{t}{\sigma}\right) + \frac{t^2}{2\sigma^2} > \log\left(\frac{2}{\sqrt{2\pi}\delta}\right) \\
&\iff \underbrace{\log\left(\frac{\frac{\sigma^2\varepsilon}{\Delta f} - \frac{\Delta f}{2}}{\sigma}\right)}_{=:A} + \underbrace{\frac{\left(\frac{\sigma^2\varepsilon}{\Delta f} - \frac{\Delta f}{2}\right)^2}{2\sigma^2}}_{=:B} > \log\left(\frac{2}{\sqrt{2\pi}\delta}\right) = \log\left(\sqrt{\frac{2}{\pi}}\frac{1}{\delta}\right).
\end{aligned}$$

We write $\sigma = \frac{c\Delta f}{\varepsilon}$ in such a from based on our previous experience. We want to bound c.

For the A term, we want to find conditions such that A is non-negative.

$$\frac{\frac{\sigma^2\varepsilon}{\Delta f} - \frac{\Delta f}{2}}{\sigma} = \frac{1}{\sigma}\left(\frac{c^2\Delta f^2}{\varepsilon^2}\frac{\varepsilon}{\Delta f} - \frac{\Delta f}{2}\right) = \frac{\varepsilon}{c\Delta f}\left(c^2\frac{\Delta f}{\varepsilon} - \frac{\Delta f}{2}\right) = c - \frac{\varepsilon}{2c}$$

Assume $c \geq 1$, then since $\varepsilon \leq 1$, we have

$$c - \frac{\varepsilon}{2c} \geq c - \frac{1}{2}.$$

So, $A > 0$ provided that $c \geq \frac{3}{2}$ (because $c - \frac{\varepsilon}{2c} \geq c - \frac{1}{2} \geq 1$).

For the B term,

$$\begin{aligned}
\frac{1}{2\sigma^2}\left(\frac{\sigma^2\varepsilon}{\Delta f} - \frac{\Delta f}{2}\right)^2 = \frac{1}{2\sigma^2}\left(\Delta f\left(\frac{c^2}{\varepsilon} - \frac{1}{2}\right)\right)^2 &= \frac{1}{2}\frac{\varepsilon^2}{c^2\Delta f^2}\Delta f^2\left(\frac{c^2}{\varepsilon} - \frac{1}{2}\right)^2 \\
&= \frac{1}{2}\left(c^2 - \varepsilon + \frac{\varepsilon^2}{4c^2}\right)
\end{aligned}$$

$c^2 - \varepsilon + \frac{\varepsilon^2}{4c^2}$ is quadratic in ε with minimum at $\varepsilon = 2c^2 \geq 9/2 > 1$. So, subject to $\varepsilon \leq 1$ and $c \geq 3/2$, it is minimized at $\varepsilon = 1$

$$c^2 - \varepsilon + \frac{\varepsilon^2}{4c^2} \geq c^2 - 1 + \frac{1}{4c^2} \geq c^2 - 1$$

So it suffices to show

$$\begin{aligned} & c^2 - 1 \geq 2\log\left(\sqrt{\frac{2}{\pi}}\frac{1}{\delta}\right) \\ \iff & c^2 \geq \log\left(\sqrt{\frac{2}{\pi}}\frac{1}{\delta}\right) + 2\log(\sqrt{e}) = 2\log\left(\sqrt{\frac{2e}{\pi}}\frac{1}{\delta}\right) \geq 2\log\left(\frac{1.32}{\delta}\right) \end{aligned}$$

because $\sqrt{2e/\pi} \leq 1.32$. So we conclude that when $c^2 = \max\left\{2\log\left(\frac{1.32}{\delta}\right), \frac{3}{2}\right\}$,

$$\Pr\left[|y| \geq \frac{\sigma^2\varepsilon}{\Delta f} - \frac{\Delta f}{2}\right] < \delta,$$

where $\sigma = \frac{c\Delta f}{\varepsilon}$ and conclude that the PLRV is bounded by ε w.p. $1 - \delta$. So by Lemma 6, Gaussian mechanism satisfies (ε, δ)-DP. ■

Remark 9.4. Claim from Dwork and Roth [2014]:

$$c^2 - \varepsilon + \frac{\varepsilon^2}{4c^2} \geq c^2 - \frac{8}{9},$$

which is false $c = 100$ and $\varepsilon = 1$.

Example 9.1. Suppose that $T\colon \mathcal{X}^n \to \mathbb{Z}^d$ consists of several (d) count statistics, then

$$\begin{aligned} \Delta_1 T &= d \\ \Delta_2 T &= \sqrt{d} \end{aligned}$$

- Laplace mechanism adds $O_p(d/\varepsilon)$ noise to get ε-DP.
- Gaussian mechanism adds $O_p(\sqrt{d\log(1/\delta)}/\varepsilon)$ noise to get (ε, δ)-DP.

The improved dependence on d is a key difference between ε-DP and (ε, δ)-DP.

10 Composition in Approximate DP

10.1 Pure versus Approximate DP

ε-DP is:

Pros:

- Original DP definition
- Relatively easy to interpret and prove
- Strong(er) privacy guarantee

Cons:

- Requires relatively heavy tailed noise
- **Poor scaling with dimension**

On the other hand, (ε, δ)-DP is:

Pros:

- Allows subgaussian noise
- Better scaling with dimension

Cons:

- (Generally) harder to interpret and prove
- **Ignores events with small probability**

Example 10.1 ("Between pure and approximate DP": Steinke and Ullman [2016])**.** For a database $D \in \{\pm 1\}^{n \times d}$, the $\overline{D}$, one-way marginals are the means of the bits in the d columns

$$\overline{D} = \frac{1}{n} \sum_{i=1}^{n} D_i \in [\pm 1]^d$$

where D_i is the i-th row of D. We call a mechanism M accurate on input D if its output is "close" to $\overline{D}$, i.e.

$$\left\| M(D) - \overline{D} \right\|_1 \leq \alpha d$$

Under ε-DP, to achieve $\mathbb{E}_M \left\| M(D) - \overline{D} \right\|_1 \leq \alpha d$, we require $n = \Omega\left(\frac{d}{\alpha \varepsilon} \right)$. But under (ε, δ)-DP, we require $n = \Omega\left(\frac{\sqrt{d \log(1/\delta)}}{\alpha \varepsilon} \right)$. Hence, if we have large d,

- For ε-DP, we need $n = \Omega(d)$.

- But for (ε, δ)-DP, we only need $n = \Omega(\sqrt{d})$.

Remark 10.1 (Connection to Composition)**.**

- Recall that when composing k ε-DP mechanism, joint release satisfies $(k\varepsilon, 0)$-DP.
- We learned that composing k (ε, δ)-DP mechanisms gives $(k\varepsilon, k\delta)$-DP.

10.2 Advanced Composition

The main reference for this section is Dwork and Roth [2014].

First, we need to clarify what is meant by "composition."

(1) Repeated used of DP mechanism on the **same dataset** (could be repeated use of the same mechanism or a modular constructor).

(2) Repeated use of DP mechanism on **different databases** that may contain information on the same individual.

- Reason about cumulative privacy loss of an individual across datasets.
- The adversary could influence the makeup of future datasets.

The composition model allow the adversary to

(1) adaptively affect the database input in the mechanisms.

(2) choose the queries based on past outputs.

Let f be a family of databases access mechanisms (e.g. set of all ε-DP mechanisms). For a probabilistic adversary $\mathcal{A}$, consider experiment 0 and experiment 1.

Algorithm 5: Experiment b for family f and adversary $\mathcal{A}$

for $i = 1, \cdots, k$ **do**
  $\mathcal{A}$ outputs two adjacent databases X_i^0 and X_i^1, a mechanism $M_i \in f$ and possible parameters w_i.
  $\mathcal{A}$ receives $y_i \in M_i(X_i^b, w_i)$.
end for

Note that the adversary A gets to choose the adjacent **databases** (that they want to distinguish from), **mechanisms**, and **parameters** based on previous outputs.

Call $(y_1, y_2, \cdots, y_k)$ to be $\mathcal{A}$'s **view** of the experiment (the X_i^j's, M_i's, and w_i's are all a function of $(y_1, y_2, \cdots, y_k)$).

Definition 10.1. We say that the family f satisfies (ε, δ)-DP under k-**fold adaptive composition** if for every $\mathcal{A}$,

$$\Pr[\underline{y_0} \in S] \leq e^{\varepsilon} \Pr[\underline{y_1} \in S] + \delta \qquad \forall S$$

where $\underline{y_0}$ and $\underline{y_1}$ are $\mathcal{A}$'s view under experiment 0 and 1 respectively.

Theorem 10.1 (Advanced Composition)**.** For all $\varepsilon, \delta, \delta' \geq 0$, the class of (ε, δ)-DP mechanisms satisfies $(\varepsilon', k\delta + \delta')$-DP under k-fold adaptive composition where δ' is arbitrary and

$$\varepsilon' = \sqrt{2k \log(1/\delta')}\varepsilon + k\varepsilon \underbrace{(e^{\varepsilon} - 1)}_{\approx \varepsilon \text{ when } \varepsilon \text{ is small}}$$

Remark 10.2. The (ε, δ) in the class of (ε, δ)-DP is fixed across iterations. Essentially, by sacrificing a proper amount of δ', we are able to improve the privacy from $k\varepsilon$ to $O(\sqrt{k}\varepsilon)$.

Corollary 10.1. Given a target $0 < \varepsilon' < 1$ and $\delta' > 0$, to ensure $(\varepsilon', k\delta + \delta')$ cumulative privacy loss over k mechanisms, it suffices that each of the k mechanisms is (ε, δ)-DP where

$$\varepsilon = \frac{\varepsilon'}{2\sqrt{2k \log(1/\delta')}}$$

Theorem 10.2 (Murtagh and Vadhan [2016])**.** Given k arbitrary $(\varepsilon_1, \delta_1), \cdots, (\varepsilon_k, \delta_k)$-DP mechanisms, computing

$$\inf_{\varepsilon} \{\varepsilon \mid (M_1, \cdots, M_k) \text{ is } (\varepsilon, \delta)\text{-DP}\}$$

for a given δ is #P-Complete.

Remark 10.3. Note that

- NP refers to "are there any solution with respect to certain constraints?";
- #P refers to "how many solutions are there?"

The bed new is NP is already "hard", while #P is even harder and #P-Complete is even worse!

11 DP SGD

11.1 Privacy Amplification

Main references for this lecture are Smith [2009] and Kasiviswanathan et al. [2008].

Subsampling is a very powerful method in DP. In pure-DP, we (mostly) work in the framework of unbounded-DP (add/delete-DP). Suppose M is a 1-DP mechanism and $M'(\cdot\,;\varepsilon)$ which operates as follows:

Algorithm 6: $M'(\cdot\,;\varepsilon)$ – Amplification via Subsampling

Input: 1-DP mechanism M, database X, $\varepsilon \in (0,1]$

1 Construct $T \subseteq X$ by selecting each element of X independently with probability ε.

Output: $M(T)$

Proposition 11.1 (Amplification via Subsampling)**.** If M is 1-DP then for $\varepsilon \in (0,1]$, $M'(\cdot\,;\varepsilon)$ in Algorithm 6 is 2ε-DP (technically, we can get $(e-1)\varepsilon$-DP).

Proof of Proposition 11.1. Fix an event S in the output sapce of M' and two adjacent databases X and X', where we assume the individual i is in X but not in X'.

Consider a run of M' on X.

(1) If i is not included in T.

Then the output is distributed the same as a run of M' on $X' = X \setminus \{i\}$ (conditioning on the event $i \notin T$) since the inclusion of i in T is independent of the inclusion of the other elements.

Let p_X be the distribution of $M'(X)$ and $p_{X'}$ be the distribution of $M'(X')$, then

$$p_X(S \mid i \notin T) = p_{X'}(S)$$

(2) If i is in T.

Then the behavior of M is within a factor of e^1 from the behavior of M on $T \setminus \{i\}$. Furthermore, by independence, the distribution of $T \setminus \{i\}$ is the same distribution of T conditioned on the omission of $\{i\}$.

$$e^{-1}p_{X'}(S) \le p_X(S \mid i \in T) \le e^1 p_{X'}(S)$$

Then, an upper bound of $p_X(S)$ is

$$\begin{aligned} p_X(S) &= (1-\varepsilon)p_X(S \mid i \notin T) + \varepsilon p_X(S \mid i \in T) \\ &\le (1-\varepsilon)p_{X'}(S) + \varepsilon e^1 p_{X'}(S) \\ &= [1+\varepsilon(e-1)]p_{X'}(S) \end{aligned} \tag{11.1}$$

At $x = 0$, tangent line of $\exp(2\varepsilon)$ is $1+2\varepsilon$ and is a lower bound, so

$$1 + (e-1)\varepsilon \le 1 + 2\varepsilon \le \exp(2\varepsilon)$$

because $e - 1 \leq 2$. So $1 + (e-1)\varepsilon$ is a weaker lower bound and hence

$$\text{Eq. (11.1)} \leq e^{2\varepsilon} p_{X'}(S);$$

a tighter result is $\exp((e-1)\varepsilon)$.

For a lower bound, we can do the following

$$\begin{aligned} p_X(S) &= (1-\varepsilon)p_X(S \mid i \notin T) + \varepsilon p_X(S \mid i \in T) \\ &\geq (1-\varepsilon)p_{X'}(S) + \varepsilon e^{-1} p_{X'}(S) \\ &= (1 - \varepsilon(1-e^{-1}))p_{X'}(S) \\ &\geq \exp(-\varepsilon) p_{X'}(S) \end{aligned}$$

because from the secant line of $\exp(-\varepsilon)$ at points $(0,1)$ and $(1, e^{-1})$, we can see that it is upper bound to $e^{-\varepsilon}$ by convexity. Thus,

$$y - 1 = \frac{1 - e^{-1}}{0 - 1}(\varepsilon - 0)$$

gives $y = 1 - \varepsilon(1 - e^{-1})$ as its upper bound.

In conclusion, we get $(e-1)\varepsilon$-DP, thereby implying 2ε-DP. ■

Remark 11.1. We can also start with c-DP, for any c, but retain elements with probability $\approx \varepsilon e^{-c}$ and $c \geq 1$.

Remark 11.2 (Two Interpretations of Proposition 11.1)**.**

(1) **Design of Mechanism**: By ignoring some data, we boost the privacy of the mechanism at the cost of some utility. With more advanced sampling and composition results, this can be powerful.

(2) **Interpretation of survey design**: If our data were collected in a simple random sample from some population of size N and if it is reasonable to model the inclusion of an individual as independent Bernuolli and the sample itself will remain secret, then we get privacy amplification "for free" called "secrecy of the sample".

This type of sampling where each individual is included independently is sometimes called Poisson sampling. More generally, subsampling with (ε, δ)-DP results in $(O(q\varepsilon), q\delta)$-DP, where $q = \frac{L}{N}$ is the subsampling rate.

11.2 Algorithm of DP SGD

The motivating question is – *where to put the noise?* There have been a few options:

- Perturb data/sufficient statistics: e.g. regression, exponential form, etc.
- Alter the objective functions: exponential mechanism (and KNG), optimize in a noisy way (functional mechanism, objective perturbation).

- Build a mechanism density based on objective.

DP Stochastic Gradient Descent (DP SGD) introduces noise into steps of an algorithm! In this way, we can handle very complicated objectives and high dimensional parameteres (e.g. deep neural network).

Algorithm 7: DP SGD (with bounded DP)

Input: Dataset $X = (X_1, \cdots, X_n)$ and loss function $L(\theta, X)$ differentiable in θ.
Parameter: Initial state θ_0, learning rate η_t, batch size m, time horizon T, noise scale σ, and gradient clipping bound C.

for $t = 1, \cdots, T$ **do**
  Sub(set)sampling: Take $I_t \subseteq \{1, \cdots, n\}$ of size m uniformly at random.
  for $i \in I_t$ **do**
    Compute gradient: $V_t^{(i)} = \nabla_\theta L(\theta_t, X_i)$.
    Clip gradient: $\overline{V}_t^{(i)} = \dfrac{V_t^{(i)}}{\max\left\{1, \frac{\|V_t^{(i)}\|_2}{C}\right\}}$.
    Average, perturb, and descend: $\theta_{t+1} = \theta_t - \eta_t \left(\dfrac{1}{m} \sum_{i=1}^{t} \overline{V}_t^{(i)} + \mathcal{N}\left(0, \dfrac{4\sigma^2 C^2}{m^2} I\right) \right)$.
    /* Sensitivity of $\frac{1}{m}\sum_{i=1}^{t} \overline{V}_t^{(i)}$ is $\frac{2C}{m}$. */
  end for
end for

Output: θ_T or the whole $(\theta_0, \cdots, \theta_T)$ for analysis.

Remark 11.3. 1. DP SGD was first proposed by Song et al. [2013]. It

- was limited by ε-DP, so it cannot afford many accurate iterations
- did not leverage privacy amplification by subsampling

2. In Bassily et al. [2014]

- DP SGD improved using (ε, δ)-DP and subsampling
- It matches lower bounds
- But its constraints/logarithmic factors were not optimized

3. Abadi et al. [2016]

- first practical implementation of DP SGD
- Most significant contribution is "Moments Accountant" for tighter composition

The only downside of the DP SGD is that it has a bunch of parameters.

11.3 Moments Accountant

The main reference for this section is Abadi et al. [2016].

Recall that (ε, δ)-DP is implied by a tail bound on the PLRV for mechanism M

$$\mathsf{PLRV}(y; M, \mathsf{aux}, D, D') = \log \frac{P(M(\mathsf{aux}, D) = y)}{P(M(\mathsf{aux}, D') = y)}$$

is a RV where $y \sim M(\mathsf{aux}, D)$ usually for adjacent D, D'.

Note that the composition of mechanisms results in the sum of the PLRVs: let $M_{1:i}$ denote $(M_1, \ldots, M_i)$ and $y_{1:i} = (y_1, \ldots, y_i)$

$$\begin{aligned}
\mathsf{PLRV}(y_{1:k}; M_{1:k}, y_{1:k-1}, D, D') &= \log \frac{P[M_{1:k}(D; y_{1:(k-1)}) = y_{1:k}]}{P[M_{1:k}(D'; y_{1:(k-1)}) = y_{1:k}]} \\
&= \log \prod_{i=1}^{k} \frac{P[M_i(D) = y_i \mid M_{1:(i-1)}(D) = y_{1:(i-1)}]}{P[M_i(D') = y_i \mid M_{1:(i-1)}(D') = y_{1:(i-1)}]} \\
&= \sum_{i=1}^{k} \mathsf{PLRV}(y_i; M_i, y_{1:(i-1)}, D, D')
\end{aligned}$$

Remark 11.4. Two observations: Moment generating functions (MGF)

1. behave nicely for sums $M_{X+Y}(t) = M_X(t) M_Y(t)$
2. can give fairly tight tail bounds via Markov/Chernoff bounds

Q: Are PLRVs independent with sequential compositions?

Definition 11.1. Define

1. the λ^{th} moment as the the cumulant generating function of the PLRV, that is,

$$\alpha_M(\lambda; \mathsf{aux}, D, D') = \log \mathbb{E}_{y \sim M(\mathsf{aux}, D)} \exp(\lambda \ \mathsf{PLRV}(y; M, \mathsf{aux}, D, D'))$$

2. its max over all possible aux and neighboring datasets D, D'

$$\alpha_M(\lambda) = \max_{\mathsf{aux}, D, D'} \alpha_M(\lambda; \mathsf{aux}, D, D').$$

This addresses possible concern with sequential composition.

Theorem 11.2. 1. Composition: Suppose M consists of a sequence of adaptive mechanisms $M_1, \ldots, M_k$, where $M_i : (\prod_{j=1}^{i-1} \mathcal{Y}_j) \times \mathcal{D} \to \mathcal{Y}_i$, then

$$\alpha_M(\lambda) \leq \sum_{i=1}^{k} \alpha_{M_i}(\lambda)$$

2. Tail bound: for any $\varepsilon > 0$, M is (ε, δ)-DP for

$$\delta = \min_{\lambda} \exp\left(\alpha_M(\lambda) - \lambda_\varepsilon\right)$$

Proof. 1. is a standard property of cumulant generating functions because the PLRV is a sum

2.

$$\begin{aligned} P_{y\sim M(D)}(\text{PLRV}(y) \geq \varepsilon) &= P(\exp(\lambda \text{ PLRV}) \geq \exp(\lambda\varepsilon)) \\ &\leq \frac{\mathbb{E}\exp\left(\lambda \text{ PLRV}\right)}{\exp(\lambda\varepsilon)} \\ &\leq \exp(\alpha - \lambda\varepsilon) \end{aligned}$$

from the previous lemma for (ε, δ)-DP completes the proof.

■

Lemma 7. Suppose $f : D \to \mathbb{R}^p$ with $\|f(\cdot)\|_2 \leq 1$. Let $\sigma \geq 1$ and let J be a sample from $[n]$ where each $i \in [n]$ is chosen independently with probability $q \leq \frac{1}{16\sigma}$. Then, for any $0 < \lambda < \sigma^2 \log \frac{1}{q\sigma}$, the mechanism $M(D) = \sum_{i\in J} f(d_i) + N(0, \sigma^2 I)$ satisfies

$$\alpha_M(\lambda) \leq \frac{q^2\lambda(\lambda+1)}{(1-q)\sigma^2} + O\left(\frac{q^3\lambda^3}{\sigma^3}\right).$$

Theorem 11.3. [Abadi et al., 2016] There exists constants c_1, c_2 given the sampling probability $q = \frac{L}{N}$ and # of iterations T, such that for any $\varepsilon < c_1 q^2 T$, DP SGD is (ε, δ)-DP for any $\delta > 0$ if (we choose)

$$\sigma \geq \frac{c_2 q\sqrt{T\log(1/\delta)}}{\varepsilon}.$$

Using Advanced Composition, we can get $\sigma = \Omega(\frac{q\sqrt{T\log(1/\delta)\log(T/\delta)}}{\varepsilon})$. So, Moments Accountant has a tighter results and, in fact, Moments Accountant motivated Rényi-DP.

Remark 11.5 (Analysis of Algorithm 7)**.** Let V be the vector space for θ and $M\colon \mathcal{X}^n \times V \to V$ from lines 3-6 in Algorithm 7. Then, in iteration t, $M(X_{I_t}, \theta_t) = \theta_{t+1}$ where X_{I_t} denotes the subset of X indexed by I_t.
Combining M with the subsampling step (line 2 in Algorithm 7), then Noisy SGD

$$\begin{aligned} \text{Noisy SGD}\colon \mathcal{X}^n &\to V \times \cdots \times V \\ X &\mapsto (\theta_1, \cdots, \theta_T) \end{aligned}$$

which is the composition of T copies of $\widetilde{M}$, since,

$$\theta_{j+1} = \widetilde{M}(X, \theta_j) \text{ for } j = 0, 1, \cdots, T-1$$

If M satisfies f-DP, then $\widetilde{M}$ is $C_{m/n}(f)$-DP. So, the composition is $\left[C_{m/n}(f)\right]^{\otimes T}$-DP.
Note that M satisfies $\frac{1}{\sigma}$-GDP, since $\frac{1}{m}\sum_{i=1}^{t} \overline{V}_t^{(i)}$ has sensitivity $\frac{2c}{m}$ (called change-DP).

Remark 11.6. However, composition with tradeoff functions is hard. Let us try *dominating pairs*. $P = \mathcal{N}(0, \sigma^2)$ and $Q = \mathcal{N}(1, \sigma^2)$ is a tightly dominating pair for M (under change-DP). Then by the previous result (Theorem 14.3),

$$\begin{aligned} &(P, (1-\gamma)P + \gamma Q) \text{ is dominating for } \widetilde{M} \text{ under add} \\ &((1-\gamma)P + \gamma Q, Q) \text{ is dominating for } \widetilde{M} \text{ under remove} \end{aligned}$$

For composition, we can use characteristic unctions of the PLRVs

$$\begin{aligned} \text{PLRV}_1 &= \log\left(\frac{\phi\left(\frac{X}{\sigma}\right)}{(1-\gamma)\phi\left(\frac{X}{\sigma}\right) + \gamma\phi\left(\frac{X-1}{\sigma}\right)}\right), \; X \sim \mathcal{N}(0, \sigma^2) \\ \text{PLRV}_2 &= \log\left(\frac{(1-\gamma)\phi\left(\frac{X}{\sigma}\right) + \gamma\phi\left(\frac{X-1}{\sigma}\right)}{\phi\left(\frac{X}{\sigma}\right)}\right), \; X \sim (1-\gamma)\mathcal{N}(0, \sigma^2) + \gamma\mathcal{N}(1, \sigma^2) \end{aligned}$$

12 Renyi Differential Privacy

The main references for this lecture are Bun and Steinke [2016] and Mironov [2017].

For the two main purposes of (ε, δ)-DP (Gaussian mechanism and composition), the analysis of (ε, δ)-DP is cumbersome and not tight. Potentially, an alternative framework can achieve these with a tighter privacy guarantee.

Recall that DP measures whether the distribution $M(X)$ and $M(X')$ are "close". We could use different measure of this "closeness."

Definition 12.1 (Rényi-Divergence)**.** For two probability distributions P and Q over the same space $\mathcal{Y}$, the Rényi divergence of order $\alpha > 1$ is

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \left(\mathbb{E}_{X\sim Q} \left[\left(\frac{p(X)}{q(X)} \right)^\alpha \right] \right)$$

where $p(X)$ and $q(X)$ are the densities at X.

Remark 12.1.

- $$D_1(P\|Q) = \lim_{\alpha\to 1} D_\alpha(P\|Q) = \mathbb{E}_{X\sim P}\left[\frac{p(X)}{q(X)}\right],$$

  which is the KL-divergence, and

- $$D_\infty(P\|Q) = \sup_{X\in \mathrm{Supp}(Q)} \left[\log \frac{P(X)}{Q(X)}\right],$$

  which is related to ε-DP in the following way: M satisfies $(\varepsilon, 0)$-DP if and only if

  $$D_\infty(M(X)\|M(X')) \le \varepsilon$$

  for all $H(X, X') \le 1$.

Remark 12.2 (Fact)**.** $D_{\alpha_1}(P\|Q) \le D_{\alpha_2}(P\|Q)$ for $\alpha_1 \le \alpha_2$.

The following definition is an relaxation of $(\varepsilon, 0)$-DP:

Definition 12.2 ((α, ε)-Renyi-DP, or (α, ε)-RDP)**.** A mechanism M satisfies (α, ε)-RDP if

$$D_\alpha(M(X)\|M(X')) \le \varepsilon$$

for all $H(X, X') \le 1$.

Usually, people give a function $(\alpha, \varepsilon(\alpha))$-DP.

12.1 Interpreting Rényi DP

Let S be a measurable subset of outcomes. Recall that under ε-DP,

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

So, the probability of S goes up by at most a factor of e^{ε}.

Proposition 12.1. If M is (α, ε)-RDP,

$$e^{-\varepsilon}\big(\Pr[M(X') \in S]\big)^{\frac{\alpha}{\alpha-1}} \leq \Pr[M(X) \in S] \leq \big(e^{\varepsilon} \Pr[M(X') \in S]\big)^{\frac{\alpha-1}{\alpha}}$$

We still get a bound on how much the probability of S increases or decreases, but it is not as simple.

Theorem 12.1 (Conversion from ε-DP to (α, ε)-RDP)**.** If M satisfies ε-DP then it satisfies $\left(\alpha, \frac{\alpha\varepsilon^2}{2}\right)$-RDP for all $1 \leq \alpha < \infty$.

Theorem 12.2 (Conversion from (α, ε)-RDP to ε-DP)**.** If M satisfies (α, ε)-RDP then it satisfies $\left(\varepsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta\right)$-DP for any $0 < \delta < 1$.

RDP implies a family of (ε, δ)-DP guarantee, avoiding the possibility of catastrophic failure at small probability. So overall, we have a sort of "sandwich"

$$\varepsilon\text{-DP} \implies \text{RDP} \implies \underbrace{\text{family of } (\varepsilon, \delta)\text{-DP}}_{\text{where } \varepsilon < \infty \text{ for all } \delta \in (0,1)} \implies (\varepsilon, \delta)\text{-DP}$$

Remark 12.3. Just as (ε, δ)-DP, RDP is robust against post-processing.

Theorem 12.3 (Adaptive Composition)**.** Let $M_1 \colon \mathcal{X}^n \to \mathcal{Y}$ be (α, ε_1)-RDP mechanism, and $M_2 \colon (\mathcal{Y}, \mathcal{X}^n) \to \mathcal{Z}$ be (α, ε_2)-RDP. Then M_3, which jointly release the output of M_1 and M_2, is $(\alpha, \varepsilon_1 + \varepsilon_2)$-RDP.

Proof. Fix $H(D, D') = 1$ for the databases D and D', write

$$\begin{aligned} &Y(y) \text{ as the distribution of } M_1(D) \\ &Z(z \mid y) \text{ as the (conditional) distribution of } M_2(D \mid y) \\ &W(y, z) \text{ as the joint distribution of } M_3(D) \end{aligned}$$

Write Y', Z', W' when D changes to D'. Then we can write joint distribution in terms of marginal

and conditional distribution as follows

$$\begin{aligned}
\exp\big((\alpha-1)D_\alpha(M_3(D)\|M_3(D'))\big) &= \mathbb{E}_{W'}\left(\frac{W}{W'}\right)^\alpha \\
&= \int_{\mathcal{Y}\times\mathcal{Z}} \big[W(y,z)\big]^\alpha \big[W'(y,z)\big]^{1-\alpha}\, dy\, dz \\
&= \int_{\mathcal{Y}}\int_{\mathcal{Z}} \big(Y(y)Z(z\mid y)\big)^\alpha \big(Y'(y)Z'(z\mid y)\big)^{1-\alpha}\, dz\, dy \\
&= \int_{\mathcal{Y}} Y(y)^\alpha Y'(y)^{1-\alpha}\left(\int_{\mathcal{Z}} Z(z\mid y)^\alpha Z'(z\mid y)^{1-\alpha}\, dz\right) dy \\
&\le \int_{\mathcal{Y}} Y(y)^\alpha Y'(y)^{1-\alpha} \exp\big((\alpha-1)\varepsilon_2\big)\, dy \\
&\le \exp\big((\alpha-1)\varepsilon_1\big)\exp\big((\alpha-1)\varepsilon_2\big) = \exp\big((\alpha-1)(\varepsilon_1+\varepsilon_2)\big)
\end{aligned}$$

since M_2 satisfies (α,ε_2)-RDP for all y. Taking log on both sides and dividing by $\alpha-1$ gives

$$D_\alpha(M_3(D)\|M_3(D')) \le \varepsilon_1+\varepsilon_2$$

and hence M_3 is $(\alpha,\varepsilon_1+\varepsilon_2)$-RDP. ■

Proposition 12.2 (Analyzing Gaussian DP). $D_\alpha\Big(\mathcal{N}\big(0,\sigma^2\big)\|\,\mathcal{N}\big(\mu,\sigma^2\big)\Big) = \dfrac{\alpha\mu^2}{2\sigma^2}$; in particular, if f has sensitivity 1, then the Gaussian mechanism $\mathcal{N}\big(0,\sigma^2\big)$ satisfies $\left(\alpha, \dfrac{\alpha}{2\sigma^2}\right)$-RDP.

Proof. Note that

$$\mathbb{E}_Q\left[\left(\frac{P}{Q}\right)^\alpha\right] = \int\left(\frac{P}{Q}\right)^\alpha Q(X)\,dx = \int \frac{P(X)^\alpha}{Q(X)^{\alpha-1}}dX$$

Hence,

$$\begin{aligned}
&D_\alpha\Big(\mathcal{N}\big(0,\sigma^2\big)\|\,\mathcal{N}\big(\mu,\sigma^2\big)\Big) \\
&= \frac{1}{\alpha-1}\log\left(\int \frac{1}{\sigma\sqrt{2\pi}}\frac{\exp\Big(-\frac{\alpha x^2}{2\sigma^2}\Big)}{\exp\Big(\frac{(1-\alpha)(x-\mu)^2}{2\sigma^2}\Big)}\,dx\right) \\
&= \frac{1}{\alpha-1}\log\left(\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{\alpha x^2}{2\sigma^2}+\frac{(1-\alpha)(x^2-2x\mu+\mu^2)}{2\sigma^2}\right)dx\right) \\
&= \frac{1}{\alpha-1}\log\left(\underbrace{\frac{1}{\sigma\sqrt{2\pi}}\int\exp\left(-\frac{(x-(1-\alpha)\mu)^2}{2\sigma^2}\right.}_{=1}\left.+\frac{(1-\alpha)^2\mu^2}{2\sigma^2}-\frac{(1-\alpha)\mu^2}{2\sigma^2}\right)dx\right) \\
&= \frac{1}{\alpha-1}\log\left(\exp\left(\frac{(\alpha^2-\alpha)\mu^2}{2\sigma^2}\right)\right) = \frac{\alpha\mu^2}{2\sigma^2}
\end{aligned}$$

as desired. Often with RDP, ε is treated as a function of α. ■

13 f-Differential Privacy

Main reference for this lecture is Dong et al. [2022].

13.1 Hypothesis Testing Formulation of Differential Privacy

Recall that if M satisfies (ε, δ)-DP, then when testing

$$H_0 \colon X \text{ or } H_1 \colon X'$$

for $H(X, X') \leq 1$ at type I error, the power is upper bounded by

$$\min\left\{e^{\varepsilon}\alpha + \delta, e^{-\alpha}(\alpha - 1 + \delta) + 1\right\}$$

We could generalize this by using other bounds on the power of such tests. Denote by

$$P \text{ the distribution of } M(X)$$
$$Q \text{ the distribution of } M(X')$$

Let $0 \leq \phi \leq 1$ be a rejection rule given y from either P or Q, we reject with probability $\phi(y)$. The type I and II errors are

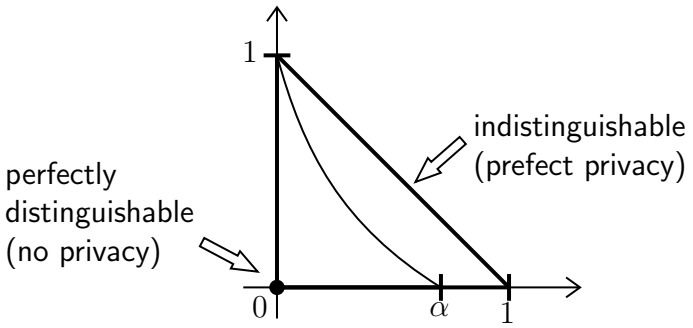
$$\text{Type I Error} = \alpha_\phi = \mathbb{E}_P[\phi]$$
$$\text{Type II Error} = \beta_\phi = 1 - \mathbb{E}_Q[\phi]$$

Definition 13.1 (Tradeoff Function/ROC)**.** For any two probability distributions P and Q on the same space, the tradeoff function $T(P, Q) \colon [0, 1] \to [0, 1]$ is defined as

$$T(P, Q)(\alpha) = \inf_{\phi}\left\{\beta_\phi \;\middle|\; \alpha_\phi \leq \alpha\right\},$$

which takes in type I errors and outputs type II errors.

Note that we are taking the infimum over all rejection rules. Geometrically, any tradeoff function will be inside the triangle:



Closer the curve is to the $y = -x + 1$, more indistinguishable P and Q are. Conversely, closer the curve is to the origin, more distinguishable they are.

Example 13.1. Consider a trivial one $T(P, P)$. Let R be any rejection set, then

$$\alpha = P(R) \text{ and } \beta = 1 - P(R)$$

So $T(P, P)(\alpha) = 1 - \alpha$.

Example 13.2. Suppose that P and Q have disjoint support:

$$R_1 \subseteq \text{Supp}(P)$$
$$R_2 = \text{Supp}(Q)$$

Let $R = R_1 \cup R_2$ be our rejection set, then

$$\alpha = P(R) = P(R_1) + P(R_2) = P(R_1)$$
$$\beta = 1 - Q(R) = 1 - (Q(R_1) + Q(R_2)) = 1 - (0 + 1) = 0$$

Example 13.3. Consider $T(\mathcal{N}(0, 1), \mathcal{N}(\mu, 1))(\alpha)$ where $\mu > 0$. Recall from Neyman-Pearson Lemma (NPL) that the optimal rejection region is of the form $[t, \infty)$.

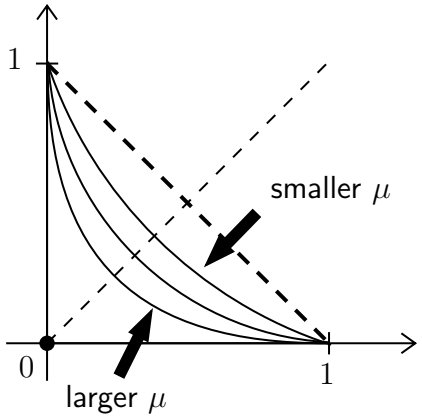
$$\text{Type I} = \Pr[Z \geq t] = 1 - \Phi(t)$$
$$\text{Type II} = \Pr[Z' \leq t] = \Pr[Z + \mu \leq t] = \Phi(t - \mu),$$

where $Z' \sim \mathcal{N}(\mu, 1)$ and $Z \sim \mathcal{N}(0, 1)$. Using the equation of type I error:

$$\alpha = 1 - \Phi(t) \implies t = \Phi^{-1}(1 - \alpha),$$

we substitute $\Phi^{-1}(1 - \alpha)$ for t to get the expression of its corresponding type II error:

$$T(\mathcal{N}(0, 1), \mathcal{N}(\mu, 1))(\alpha) = \Phi(\Phi^{-1}(1 - \alpha) - \mu).$$



The following proposition is characterization of a tradeoff function, which can be used to verify if a function is a tradeoff function or not.

Proposition 13.1. A function $f\colon [0,1] \to [0,1]$ is a tradeoff function if and only if it is:

- Convex,
- Continuous,
- Non-increasing (weakly decreasing), and
- $f(x) \leq 1 - x$ for $x \in [0,1]$.

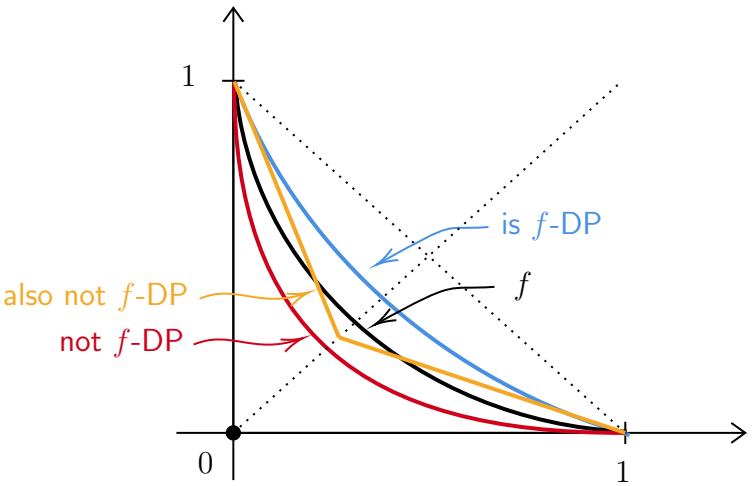
Remark 13.1 (Observation)**.** If f and g are tradeoff functions, then $\max\{f, g\}$ is a tradeoff function.

13.2 f-Differential Privacy

Definition 13.2 (f-DP)**.** Let f be a tradeoff function. A mechanism M satisfies f-Differential privacy (f-DP) if

$$T(M(X), M(X'))(\alpha) \geq f(\alpha)$$

for all $\alpha \in [0,1]$ and for all $H(X, X') \leq 1$.



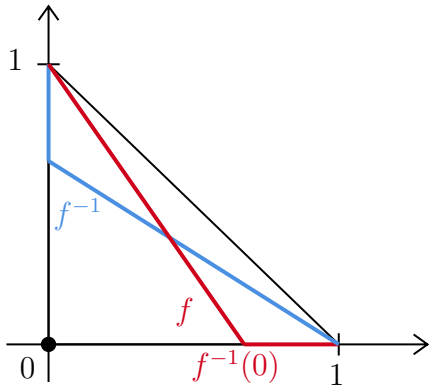
Remark 13.2. Notice that the f-DP definition is symmetric in the sense that the neighboring relation is symmetric. We also have $T(M(X'), M(X)) \geq f$. So, we can restrict our attention to symmetric tradeoff functions.

Proposition 13.2. Let M be f-DP, then M is

$$\max\{f, f^{-1}\}\text{-DP},$$

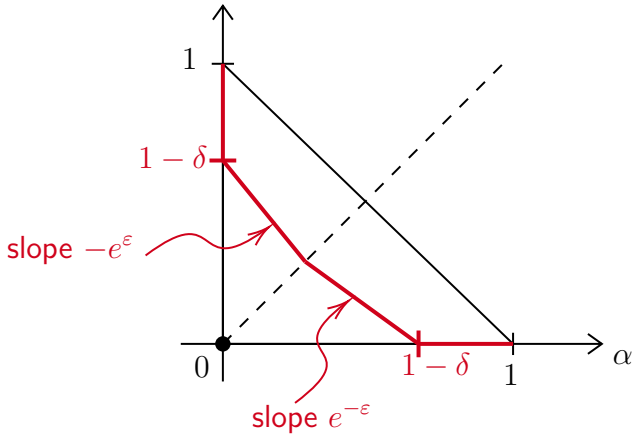
where $f^{-1}(\alpha) := \inf\{t \in [0,1] \mid f(t) \leq \alpha\}$ for all $\alpha \in [0,1]$.

This provides a natural tighter lower bound when f is not symmetric. The tradeoff function $f^S = \max\{f, f^{-1}\}$ is symmetric in that $f^S = (f^S)^{-1}$.



Remark 13.3. Finally, note that f-DP is a generalization of (ε, δ)-DP.

$$f_{\varepsilon,\delta}(\alpha) = \max\begin{cases} 0 \\ 1-\delta-e^{\varepsilon}\alpha \\ e^{-\varepsilon}(1-\delta-\alpha) \end{cases}$$



13.3 Gaussian Differential Privacy

While f-DP is a powerful privacy framework, it is nice to summarize the privacy guarantee in a single number. In pure DP, ε was our privacy parameter. An alternative is μ-GDP. Let

$$G_\mu := T(\mathcal{N}(0,1), \mathcal{N}(\mu,1)) = \Phi(\Phi^{-1}(1-\alpha) - \mu)$$

for $\mu \geq 0$. It is easy to see that $G_\mu \leq G_{\mu'}$ when $\mu \geq \mu'$.

Remark 13.4. This works out so nicely because Gaussian is log-concave and hence it has monotone likelihood ratio, which gives us a rejection region in the form of $[t, \infty)$.

Definition 13.3 (GDP)**.** A mechanism M satisfies μ-Gaussian differential privacy (μ-GDP) if it is G_μ-DP, that is

$$T(M(X), M(X')) \geq G_\mu$$

for all $H(X, X') \leq 1$.

Theorem 13.1. Let $f\colon \mathcal{X}^n \to \mathbb{R}$ has sensitivity Δ, then

$$M(X) = f(X) + z\ ,\ z \sim \mathcal{N}\left(0, \frac{\Delta^2}{\mu^2}\right)$$

satisfies μ-GDP.

Proof of Theorem 13.1. For any two X and X' s.t. $H(X, X') \leq 1$, note that

$$M(X) \sim \mathcal{N}\left(f(X), \frac{\Delta^2}{\mu^2}\right)$$
$$M(X') \sim \mathcal{N}\left(f(X'), \frac{\Delta^2}{\mu^2}\right)$$

Then, since T is preserved under invertible transformations,

$$\begin{aligned}
T(M(X), M(X')) &= T\left(\mathcal{N}\left(f(X), \frac{\Delta^2}{\mu^2}\right), \mathcal{N}\left(f(X'), \frac{\Delta^2}{\mu^2}\right)\right) \\
&= T\left(\mathcal{N}\left(0, \frac{\Delta^2}{\mu^2}\right), \mathcal{N}\left(f(X') - f(X), \frac{\Delta^2}{\mu^2}\right)\right) \\
&= T\left(\mathcal{N}(0,1), \mathcal{N}\left(\frac{\big(f(X') - f(X)\big)\mu}{\Delta}, 1\right)\right) \\
&= T\left(\mathcal{N}(0,1), \mathcal{N}\left(\frac{\big|f(X') - f(X)\big|\mu}{\Delta}, 1\right)\right) \\
&= G_{\frac{|f(X') - f(X)|\mu}{\Delta}} \\
&\geq G_\mu
\end{aligned}$$

since $\big|f(X') - f(X)\big| \leq \Delta$. ■

13.4 Post-Processing and Informativeness of f-DP

Let $\mathrm{Proc}\colon \mathcal{Y} \to \mathcal{Z}$ be a (randomized) algorithm. If $M\colon \mathcal{X}^n \to \mathcal{Y}$ is a mechanism, $\mathrm{Proc} \circ M\colon \mathcal{X}^n \to \mathcal{Z}$ is the post-processed mechanism.

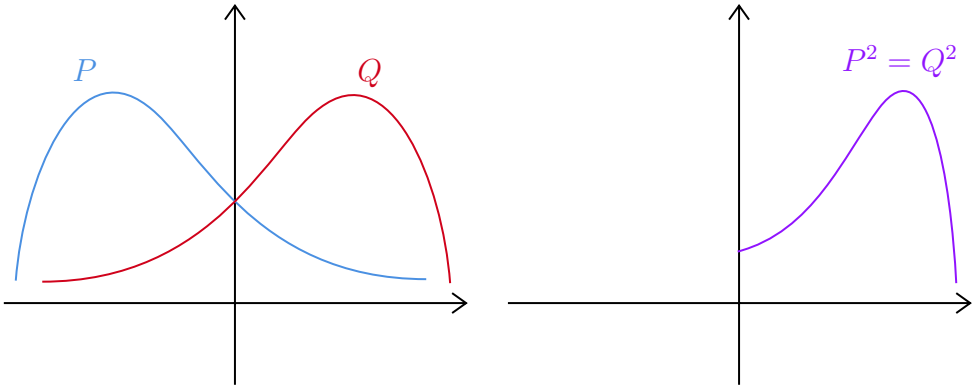
Lemma 8. For any two distributions P and Q, $T(\mathrm{Proc}(P), \mathrm{Proc}(Q)) \geq T(P, Q)$.

Proof. Let $\Phi = \{\phi : \mathcal{Y} \to [0,1]\}$ and $\Phi^{\mathrm{Proc}} = \{\phi : \mathcal{Z} \to [0,1]\}$. Given $\phi \in \Phi^{\mathrm{Proc}}$, then $\phi \circ \mathrm{Proc} \in \Phi$. Consider

$$\begin{aligned} T(M(X), M(X')) &= \inf_{\phi \in \Phi} \{1 - \mathbb{E}_{M(X')}\phi \mid \mathbb{E}_{M(X)}\phi \leq \alpha\} \\ &\leq \inf_{\phi \in \Phi^{\mathrm{Proc}}} \{1 - \mathbb{E}_{M(X')}\phi \circ \mathrm{Proc} \mid \mathbb{E}_{M(X)}\phi \circ \mathrm{Proc} \leq \alpha\} \\ &= \inf_{\phi \in \Phi^{\mathrm{Proc}}} \{1 - \mathbb{E}_{\mathrm{Proc} \circ M(X')}\phi \mid \mathbb{E}_{\mathrm{Proc} \circ M(X)}\phi \leq \alpha\} \\ &= T(\mathrm{Proc}(P), \mathrm{Proc}(Q)) \end{aligned}$$

■

Example 13.4. Basically, post-processing only makes the testing problem harder. For example, if $P = -Q$ then $P^2 = Q^2$.



Proposition 13.3. If M is f-DP, then $\mathrm{Proc} \circ M$ is also f-DP.

Remark 13.5. ε-DP, (ε, δ)-DP, Renyi-DP all have post-processing properties. The following theorem indicates that the tradeoff function is the most informative measure of indistinguishability.

Theorem 13.2 (Blackwell [1951])**.** Let P and Q be probability distribution on $\mathcal{Y}$, and P' and Q' are probability distribution on $\mathcal{Z}$. The following statements are equivalent:

(1) $T(P, Q) \leq T(P', Q')$.

(2) There exists a (randomized) function $\mathrm{Proc}\colon \mathcal{Y} \to \mathcal{Z}$ such that $\mathrm{Proc}(P) = P'$ and $\mathrm{Proc}(Q) = Q'$.[a]

[a] Probably a better notation is $\mathrm{Proc}(X) \sim P'$ such that $X \sim P$, and so on.

There is nothing new about (2) $\implies$ (1). Post-processing induces on order on pairs of distributions, called the **Blackwell order**. If (2) holds, we write

$$(P, Q) \preccurlyeq_{\text{Blackwell}} (P', Q')$$

and read it as (P, Q) **is easier to distinguish than** (P', Q') **in the Blackwell sense**. Similarly, if $T(P, Q) \leq T(P', Q')$, we write

$$(P, Q) \preccurlyeq_{\text{tradeoff}} (P', Q')$$

and read it as (P, Q) **is easier to distinguish than** (P', Q') **in the testing sense**.

For any privacy notion, we get an order $\preccurlyeq$ on pairs of distributions. If the privacy measure has post-processing property, then $\preccurlyeq$ must be consistent with $\preccurlyeq_{\text{Blackwell}}$, i.e.

$$(P, Q) \preccurlyeq_{\text{Blackwell}} (P', Q') \implies (P, Q) \preccurlyeq (P', Q')$$

Denote $\text{Ineq}(\preccurlyeq) = \left\{(P, Q; P', Q') \;\middle|\; (P, Q) \preccurlyeq (P', Q')\right\}$ be the set of all comparable pairs under the order $\preccurlyeq$. A privacy notion satisfies post-processing if and only if the induced order $\preccurlyeq$ satisfies

$$\text{Ineq}(\preccurlyeq) \supseteq \text{Ineq}(\preccurlyeq_{\text{Blackwell}})$$

Hence, a reasonable privacy definition must have $\text{Ineq}(\preccurlyeq)$ large enough to contain $\text{Ineq}(\preccurlyeq_{\text{Blackwell}})$, but we do not want $\text{Ineq}(\preccurlyeq)$ to be "too large."

Example 13.5. Consider the privacy notion based on a trivial divergence

$$D_0(P\|Q) = 0 \text{ for all } P \text{ and } Q$$

$\text{Ineq}(\preccurlyeq_{D_0})$ is the largest possible but is not at all informative about indistinguishability.

Remark 13.6. Theorem 13.2 stated that $\text{Ineq}(\preccurlyeq_{f\text{-diff}}) = \text{Ineq}(\preccurlyeq_{\text{Blackwell}})$. So f-DP is the most informative.

13.5 Conversion of f-DP to Divergence-based DP

Let $D(\cdot\|\cdot)$ be a "divergence" that takes in two probability distributions on a common space and outputs a number. We say that D has the data processing inequality (DPI) if

$$D(\text{Proc}(P)\|\text{Proc}(Q)) \leq D(P\|Q)$$

Proposition 13.4. If $D(\cdot\|\cdot)$ satisfies DPI, then there exists functional $\ell_D : \mathcal{T} \to \mathbb{R}$ such that $D(\cdot\|\cdot) = \ell_D(T(P, Q))$ for every P and Q, where $\mathcal{T}$ is a set of tradeoff functions.

Lemma 9. If $T(P', Q') \geq T(P, Q)$, then $D(P'\|Q') \leq D(P\|Q)$. In particular, if $T(P', Q') = T(P, Q)$, then $D(P'\|Q') = D(P\|Q)$.

Returning to the proposition, we define $\ell_D(T(P, Q)) = D(\widetilde{P}, \widetilde{Q})$ with any pair $(\widetilde{P}, \widetilde{Q})$ such that $T(\widetilde{P}, \widetilde{Q}) = T(P, Q)$.

Example 13.6 (F-divergence)**.** Let P and Q be distributions with densities P and Q (with respect to a common base measure). For a convex function $F\colon (0,\infty) \to \mathbb{R}$ such that $F(1) = 0$, the F-divergence $D_F(P\|Q)$ is

$$D_F(P\|Q) = \int_{\{p,q>0\}} F\left(\frac{p}{q}\right) dQ + F(0)Q[p=0] + \tau_F P[q=0]$$

where $F(0) = \lim_{x\to 0^+} F(t)$, $\tau_F = \lim_{t\to\infty} \frac{F(t)}{t}$, and we set $F(0)\cdot 0 = \tau_F \cdot 0 = 0$ even if $F(0)$ or τ_F are ∞. Note that all F-divergence satisfy DPI, i.e. post-processing. For example,

- Total variation: $F(t) = |t-1|/2$,
- KL variation: $t\log t$.

Proposition 13.5. Let $f = T(P,Q)$ and $z_f = \inf\{x \in [0,1] \mid f(x) = 0\}$ be the first zero of f. The functional $\ell_F\colon \mathcal{T} \to \mathbb{R}$ that computes the F-divergence of $D_F(P\|Q)$ is

$$\ell_F(f) = \int_0^{z_f} F(|f'(x)|^{-1})\left|f'(x)\right| dx + F(0)(1-f(0)) + \tau_F(1-z_f)$$

Example 13.7 (Proposition 13.5)**.**

- $\ell_{\mathsf{TV}}(f) = \frac{1}{2}\int_0^1 \left|1+f'(x)\right| dx$,
- $\ell_{\mathrm{KL}}(f) = -\int_0^1 \log\left|f'(x)\right| dx$, and
- Renyi-divergence can be expressed as a function of an F-divergence but we know ℓ has post-processing

$$\ell_\alpha^{\mathrm{Renyi}}(f) = \begin{cases} \frac{1}{\alpha-1}\log\int_0^1 \left|f'(x)\right|^{1-\alpha} dx & \text{, if } z_f = 1 \\ \infty & \text{, if } z_f < 1 \end{cases}$$

  Note the **exact equalities** here – no inequalities here!

Proposition 13.6. If M is f-DP then it is $(\alpha, \ell_\alpha^{\mathrm{Renyi}}(f))$-RDP for any $\alpha > 1$.

Corollary 13.1 (zCDP)**.** If M is μ-GDP then it is $\left(\alpha, \frac{1}{2}\mu^2\alpha\right)$-RDP for any $\alpha > 1$. And this is called $\left(P = \frac{1}{2}\mu^2\right)$-zCDP (zero concentrated DP).

We can convert f-DP to any divergence-based DP guarantee. It is not always possible to convert them back. Even when $\{D_\alpha(P\|Q); \alpha > 1\}$ is considered as an infinite-dimensional object, it still does not induce Blackwell order, that is,

$$\mathrm{Ineq}(\preccurlyeq_{\mathrm{Renyi}}) \supsetneq \mathrm{Ineq}(\preccurlyeq_{\mathrm{Blackwell}})$$

So there exists two pairs of distributions, where one is easier to distinguish in the Rényi sense, but not in the Blackwell sense.

Let P_ε and Q_ε denote Bernoulli distribution with success probabilities $\frac{e^\varepsilon}{1+e^\varepsilon}$ and $\frac{1}{1+e^\varepsilon}$ respectively.

Proposition 13.7. There exists $\varepsilon > 0$ such that both of the following statements are true:

(a) For all $\alpha > 1$, $D_\alpha(P_\varepsilon\|Q_\varepsilon) \le D_\alpha(\mathcal{N}(0,1)\|\mathcal{N}(\varepsilon,1))$.

(b) $\mathrm{TV}(P_\varepsilon, Q_\varepsilon) > \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(\varepsilon,1))$.

In fact (a) holds for all α.

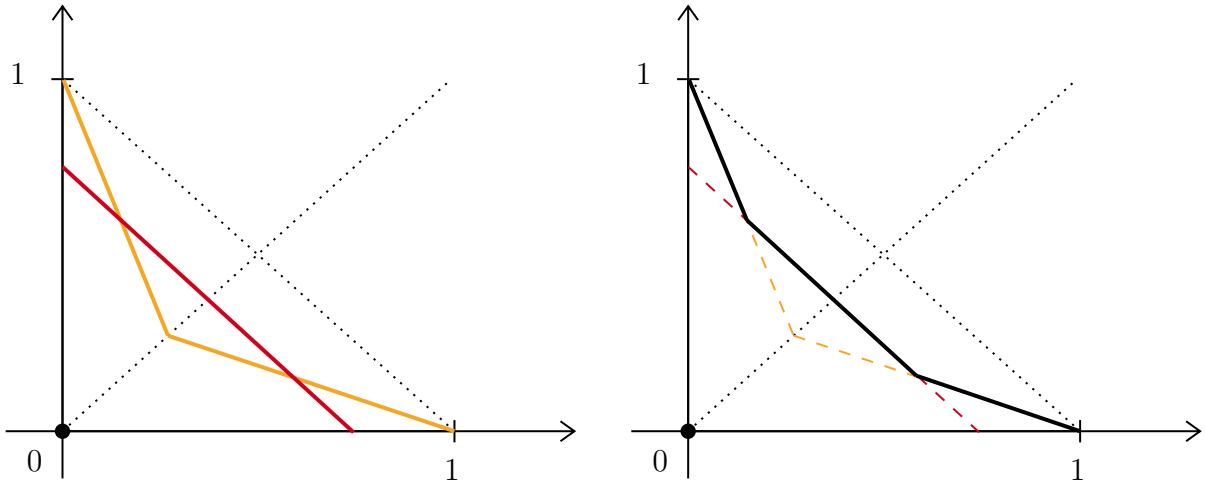
(a) says that $(\mathcal{N}(0,1), \mathcal{N}(\varepsilon,1)) \preccurlyeq_{\mathrm{Renyi}} (P_\varepsilon, Q_\varepsilon)$.

(b) excludes the possibility $(\mathcal{N}(0,1), \mathcal{N}(\varepsilon,1)) \preccurlyeq_{\mathrm{Blackwell}} (P_\varepsilon, Q_\varepsilon)$, since otherwise DPI of TV would imply $\mathrm{TV}(P_\varepsilon, Q_\varepsilon) \le \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(\varepsilon,1))$.

13.6 Primal-Dual Connection to (ε, δ)-DP

Similar to RDP, we can convert f-DP to an infinite collection of (ε, δ)-DP guarantees. Unlike RDP, we can also convert back loss-lessly.

Proposition 13.8 (Dual to Primal)**.** Let I be an arbitrary index set such that for each $i \in I$, we have $\varepsilon_i \in [0, \infty)$ and $\delta_i \in [0, 1]$. A mechanism is $(\varepsilon_i, \delta_i)$-DP for all $i \in I$ if and only if it is f-DP with $f(\alpha) = \sup_i f_{\varepsilon_i, \delta_i}(\alpha)$.



The black curve on the right is the f from the proposition. Note that the construction f is a symmetric tradeoff function.

Proposition 13.9 (Primal to Dual)**.** For a symmetric tradeoff function f, the following are equivalent:

(a) M is f-DP.

(b) M is (ε, δ)-DP for every (ε, δ) such that $-e^{\varepsilon}\alpha + (1-\delta)$ is tangent to $f(\alpha)$ at some point.

(c) M is $(\varepsilon, \delta(\varepsilon))$-DP for all $\varepsilon \geq 0$ with $\delta(\varepsilon) = 1 + f^*(e^{-\varepsilon})$, where $f^*(y) = \sup_{0\leq x\leq 1} yx - f(x)$ is the convex conjugate of f.

Note that $(\varepsilon, \delta(\varepsilon))$ and $(\varepsilon(\delta), \delta)$ are called privacy profiles on equivalent framework to f-DP.

13.7 Group Privacy

We say D and D' are k-neighbors if there exists a sequence of databses

$$D = D_0, D_1, D_2, \cdots, D_k = D'$$

such that D_i and D_{i+1} are neighbors (or identical) for all $i = 0, 1, \cdots, k-1$.

Definition 13.4. A mechanism is f-DP for groups of size k if

$$T(M(D), M(D')) \geq f$$

for all k-neighbors D and D'

Let us denote $f^{\circ k} = \underbrace{f \circ f \circ \cdots \circ f}_{k \text{ times}}$. And let $f \hat{\circ} g(x) = f(1 - g(x))$ and $f^{\hat{\circ} k} = \underbrace{f \mathbin{\hat{\circ}} f \mathbin{\hat{\circ}} \cdots \mathbin{\hat{\circ}} f}_{k \text{ times}}$ so that $f \mathbin{\hat{\circ}} g = 1 - (1-f) \circ (1-g)$ and $f^{\hat{\circ} k} = 1 - (1-f)^{\circ k}$.

Lemma 10. (1) $f \mathbin{\hat{\circ}} g$ is a tradeoff function if f and g are tradeoff functions.

(2) $(f \mathbin{\hat{\circ}} g)^{-1} = (g^{-1}) \mathbin{\hat{\circ}} (f^{-1})$ and if f is symmetric then so is $f^{\hat{\circ} k}$.

Lemma 11. Let f and g be tradeoff functions. Suppose $T(P, Q) \geq f$ and $T(Q, R) \geq g$, then

$$T(P, R) \geq g \mathbin{\hat{\circ}} f.$$

Proof of Lemma 11. Fix $\alpha \in [0, 1]$ and let ϕ be the optimal testing function for

$$H_0 \colon P \text{ v.s. } H_1 \colon R \text{ at type I error } \alpha$$

then $\mathbb{E}_P[\phi] = \alpha$ (type I) and $1 - \mathbb{E}_R[\phi] = T(P, R)(\alpha)$ (type II). Note that ϕ is suboptimal for $H_0 : Q$ v.s. $H_1 : R$ with type I error $\mathbb{E}_Q[\phi]$, the type I and type II errors must be above the tradeoff function g.

$$g(\mathbb{E}_Q[\phi]) \leq T(Q, R)(\mathbb{E}_Q[\phi]) \leq 1 - \mathbb{E}_R[\phi]$$

Similarly, ϕ is suboptimal for $H_0 : P$ v.s. $H_1 : Q$. So,

$$1 - \mathbb{E}_Q[\phi] \geq T(P, Q)(\mathbb{E}_P[\phi]) = T(P, Q)(\alpha) \geq f(\alpha),$$

which gives $\mathbb{E}_Q[\phi] \leq 1 - f(\alpha)$.

Together, we have

$$T(P, Q)(\alpha) = 1 - \mathbb{E}_R[\phi] \geq g(\mathbb{E}_Q[\phi]) \geq g(1 - f(\alpha)) = g \mathbin{\widehat{\circ}} f(\alpha)$$

since g is decreasing for the last inequality. ■

Theorem 13.3. if M is f-DP (for groups of size 1) then it is $\left(1 - (1 - f)^{\circ k}\right)$-DP for groups of size k. If M is μ-GDP, then it is $k\mu$-GDP for groups of size k.

Proof of Theorem 13.3. For f-DP part, one can apply Lemma 11 $k-1$ times. For the GDP claim, note that $G_\mu \mathbin{\widehat{\circ}} G_{\mu'} = G_{\mu+\mu'}$ because

$$G_\mu \mathbin{\widehat{\circ}} G_{\mu'}(\alpha) = G_\mu(1 - G_{\mu'}(\alpha)) = \Phi(\Phi^{-1}(G_{\mu'}(\alpha)) - \mu) = G_{\mu+\mu'}(\alpha)$$

since $G_{\mu'}(\alpha) = \Phi(\Phi^{-1}(1 - \alpha) - \mu')$. ■

Remark 13.7. If we take $\widetilde{f}(\alpha) = f(1 - \alpha)$ as in Awan and Dong [2022], then for group of size k, we get $\widetilde{f}^{\circ k}$.

13.8 Composition in f-DP

Let $M_1 : \mathcal{X} \to \mathcal{Y}_1$ and $M_2 : \mathcal{X} \to \mathcal{Y}_1 \to \mathcal{Y}_2$. The joint mechanism $M : \mathcal{X} \to \mathcal{Y}_1 \times \mathcal{Y}_2$ is defined as

$$M(X) = (M_1(X), M_2(X, M_1(X)))$$

Remark 13.8. Composition is closed and tight in the f-DP framework!

Definition 13.5 (Tensor Product)**.** The **tensor product** of two tradeoff functions $f = T(P, Q)$ and $g = T(P', Q')$ is defined as

$$f \otimes g = T(P \times P', Q \times Q')$$

and we write $f^{\otimes k} = \underbrace{f \otimes f \otimes \cdots \otimes f}_{k \text{ times}}$.

First, we need to check that $\otimes$ is well-defined.

Lemma 12 ($\otimes$ is well-defined)**.**

(1) If f is $T(P, Q) = T(\widetilde{P}, \widetilde{Q})$ then $T(P \times P', Q \times Q') = T(\widetilde{P} \times P', \widetilde{Q} \times Q')$.

(2) $\otimes$ is commutative and associative.

(3) If $g_1 \geq g_2$ then $f \otimes g_1 \geq f \otimes g_2$.

(4) $f \otimes \mathrm{Id} = \mathrm{Id} \otimes f = f$ where $\mathrm{Id}(\alpha) = 1 - \alpha$.

(5) $(f \otimes g)^{-1} = f^{-1} \otimes g^{-1}$, which implies that if f and g are symmetric then $f \otimes g$ is symmetric.

Theorem 13.4. Let $M_i(\,\cdot\,; y_1, \cdots, y_{i-1})$ be f_i-DP for all $y_1 \in \mathcal{Y}_1, \cdots, y_{i-1} \in \mathcal{Y}_{i-1}$ then the k-fold composed mechanism $M \colon \mathcal{X} \to \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_k$ is $f_1 \otimes f_2 \otimes \cdots \otimes f_k$-DP.

This theorem is **tight** in that it cannot be improved in general. For example, if M_2 does not depend on the output of M_1,

$$\begin{aligned} T(M(X), M(X')) &= T(M_1(X) \times M_2(X), M_1(X') \times M_2(X')) \\ &= T(M_1(X), M_1(X')) \otimes T(M_2(X), M_2(X')) \end{aligned}$$

If X and X' are neighboring datasets such that

$$T(M_1(X), M_1(X')) = f_1 \text{ and } T(M_2(X), M_2(X')) = f_2,$$

we conclude that $f_1 \otimes f_2$ is the tightest possible bound on their composition.

In the case of GDP,

$$G_{\mu_1} \otimes \cdots \otimes G_{\mu_k} = G_{\sqrt{\mu_1^2 + \cdots + \mu_k^2}}$$

So, k-fold composition of μ-GDP mechanisms is $\sqrt{\mu_1^2 + \cdots + \mu_k^2}$-GDP.

13.9 Central Limit Theorem for Composition

The moral here is that when composing many privacy mechanisms, each with a small privacy loss budget (close to perfect indistinguishability) in the limit, the privacy guarantee approaches μ-GDP for some μ.

The μ parameter depends on certain functionals of the tradeoff functions:

$$\begin{aligned} kl(f) &= -\int_0^1 \log\left|f'(x)\right| dx \\ k_2(f) &= \int_0^1 \log^2\left|f'(x)\right| dx \\ k_3(f) &= \int_0^1 \left|\log\left|f'(x)\right|\right|^3 dx \\ \overline{k_3}(f) &= \int_0^1 \left|\log\left|f'(x)\right| - kl(f)\right|^3 dx \end{aligned}$$

all of which take values in $[0, \infty)$.

Those functionals calculate moments of PLRV (i.e. moments of the log likelihood ratio of P and Q such that $f = T(P, Q)$). We write

$$\underline{kl} = (kl(f_1), \cdots, kl(f_k))$$

and similarly for other functionals.

Theorem 13.5. Let $f_1, \cdots, f_k$ be symmetric tradeoff functions such that $k_3(f_i) < \infty$ for all $i = 1, \cdots, k$. Denote

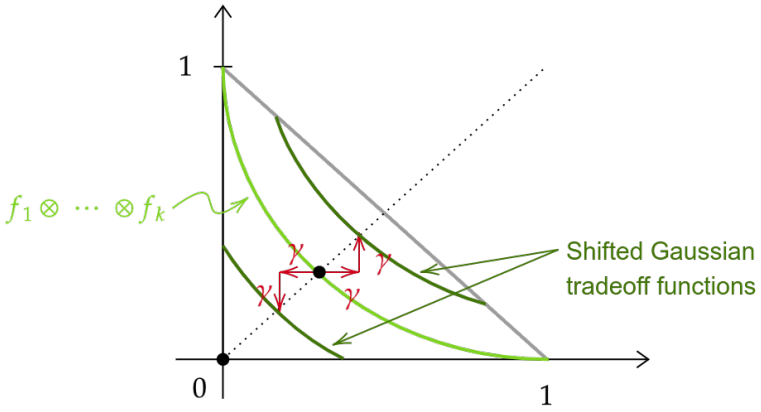
$$\mu = \frac{2\|\underline{kl}\|_1}{\sqrt{\|k_2\|_1 - \|kl\|_2^2}}$$

$$\gamma = \frac{0.56\left\|\overline{\underline{k_3}}\right\|_1}{\left(\|k_2\|_1 - \|kl\|_2^2\right)^{\frac{3}{2}}}$$

Assume that $\gamma < 1/2$, then for all $\alpha \in [\gamma, 1-\gamma]$,

$$G_\mu(\alpha + \gamma) - \gamma \leq f_1 \otimes \cdots \otimes f_k \leq G_\mu(\alpha - \gamma) + \gamma$$

This is a Berry-Esseen type result, as the figure below suggests:



Theorem 13.6 (Berry-Esseen result)**.** Let $\left\{f_{k_i} \;\middle|\; 1 \le i \le k\right\}_{k=1}^{\infty}$ be a triangular array of symmetric tradeoff functions, and assume there exists $K \ge 0$ and $s > 0$ such that as $k \to \infty$:

(1) $\sum_{i=1}^{k} kl(f_{k_i}) \to K$

(2) $\max_{1 \le i \le k} kl(f_{k_i}) \to 0$

(3) $\sum_{i=1}^{k} k_2(f_{k_i}) \to s^2$

(4) $\sum_{i=1}^{k} k_3(f_{k_i}) \to 0$

then $\lim_{k\to\infty} f_{k_1} \otimes f_{k_2} \otimes \cdots \otimes f_{k_k}(\alpha) = G_{\frac{2K}{s}}(\alpha)$ uniformly for all $\alpha \in [0,1]$.

Remark 13.9. CLT is theoretically interesting; however, asymptotic privacy guarantees are not accepted in practice. The Berry Esseen result gives a lower bound, but it has a "delta." In general, evaluating the exact tensor product of tradeoff functions is a difficult testing problem.

13.10 Subset Sampling in f-DP

Let $1 \le m \le n$ and a dataset $X \in \mathcal{X}^n$ define $\mathrm{Sample}_m(X)$ to be a subset of X, chosen uniformly among all subsets of size m.

For a mechanism M defined on $\mathcal{X}^n$, call $M \circ \mathrm{Sample}_m(X) = M(\mathrm{Sample}_m(X))$ the mechanism which applies M to the subsampled dataset. Note that the subsample itself is not released.

Let f be a tradeoff function. Let $0 \le \rho \le 1$ and define

$$f_\rho = \rho f + (1-\rho)\mathrm{Id}$$

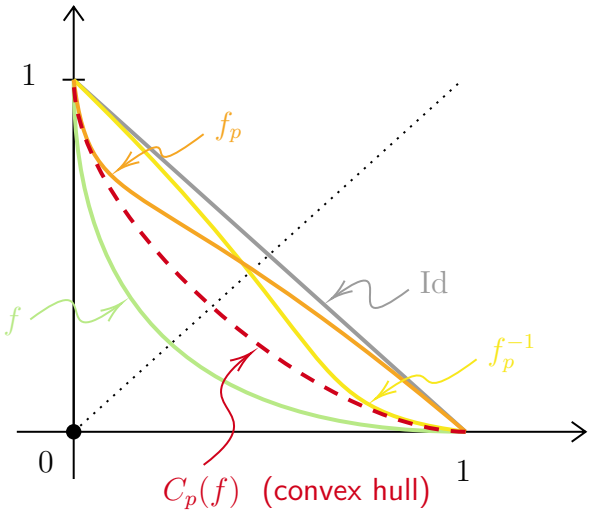
where $\mathrm{Id}(x) = 1 - x$. Note that f_ρ is asymmetrical in general.

Definition 13.6 (Subsampling Operator)**.** For any $0 \le p \le 1$, define the C_p acting on tradeoff function as

$$C_p(f) = \mathrm{ConvexHull}\left(\min\left\{f_p, f_p^{-1}\right\}\right)$$

and call C_p the subsampling operator.

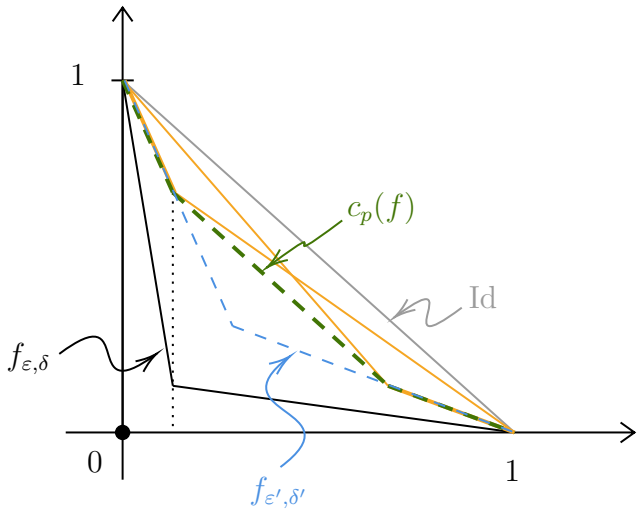
Theorem 13.7. If M is f-DP on $\mathcal{X}^n$ then $M \circ \mathrm{Sample}_m(X)$ is $C_p(f)$-DP on $\mathcal{X}^n$ where $p = m/n$.



Corollary 13.2. If M is (ε, δ)-DP on $\mathcal{X}^m$, then the subsampled mechanism $M \circ \mathrm{Sample}_m(X)$ is $C_p(f_{\varepsilon,\delta})$-DP where

$$C_p(f_{\varepsilon,\delta})(\alpha) = \max \begin{cases} f_{\varepsilon',\delta'}(\alpha) \\ 1 - p\delta - p\left(\frac{e^\varepsilon - 1}{e^\varepsilon + 1}\right) - \alpha \end{cases}$$

where $\varepsilon' = \log(1 - p + pe^\varepsilon)$, $\delta' = p\delta$, and $p = m/n$.



Remark 13.10 (Takeaway). (ε, δ)-DP is not expressive enough to communicate subsampling in a single (ε, δ) pair. Note if $\varepsilon = 1$, $\delta = 1$, and $p = \varepsilon \leq 1$ (target), then $\varepsilon' = \log(1 - \varepsilon(1 + e)) \leq e$, which is one of the quantities in the pure-DP Poisson subsampling result. Stay tuned!

14 Dominating Pairs

The main reference for this lecture is Zhu et al. [2022].

14.1 Privacy Loss Random Variable

Definition 14.1 (PLRV)**.** Let X and Y be two random variables on the same space with densities P and Q respectively. The privacy loss random variable is

$$\mathrm{PLRV}(X \mid Y) = \log \frac{P(X)}{Q(X)} \text{ where } X \sim P \text{ and } Y \sim Q$$

Lemma 13 (PLRV is sufficient)**.** Let $X \sim P$ and $Y \sim Q$ be two r.v.s on $\mathcal{X}$. Define $L(X) \colon \mathcal{X}^n \to \mathbb{R}$ by $L(x) = \log \frac{Q(x)}{P(x)}$ the log-likelihood ratio statistics. Note that

$$L(X) \overset{d}{=} -\mathrm{PLRV}(X \mid Y)$$
$$L(Y) \overset{d}{=} \mathrm{PLRV}(Y \mid X)$$

Then, $T(X, Y) = T(L(X), L(Y)) = T(-\mathrm{PLRV}(X \mid Y), \mathrm{PLRV}(Y \mid X))$.

Proof of Lemma 13. First, by post-processing

$$T(X, Y) \leq T(L(X), L(Y))$$

and for the other direction, we use Neyman-Pearson Lemma. The optimal test for $H_0 \colon X$ v.s. $H_1 \colon Y$ of size α and of the form

$$\phi(X) = \begin{cases} 1 & \text{where } L(X) > t \\ C & \text{where } L(X) = t \\ D & \text{where } L(X) < t \end{cases}$$

where C and t are chosen s.t. $\mathbb{E}_{X \sim P}[\phi(X)] = \alpha$. The type I error is

$$\begin{aligned} \mathbb{E}_{X \sim P}[\phi(X)] &= \mathbb{E}_{X \sim P}\big[\mathbb{1}(L(X) > t) + c\mathbb{1}(L(X) = t)\big] \\ &= \Pr_{X \sim P}\big[L(X) > t\big] + c \Pr_{X \sim P}\big[L(X) = t\big] \end{aligned}$$

which only depends on $L(X)$. And the type II error is

$$\begin{aligned} 1 - \mathbb{E}_{Y \sim Q}[\phi(Y)] &= 1 - \mathbb{E}_{Y \sim Q}\big[\mathbb{1}(L(Y) > t) + c\mathbb{1}(L(Y) = t)\big] \\ &= \Pr_{Y \sim Q}\big[L(Y) \leq t\big] - c \Pr_{Y \sim Q}\big[L(Y) = t\big] \end{aligned}$$

which only depends on $L(Y)$.

When testing $H_0 \colon L(X)$ v.s. $H_1 \colon L(Y)$, we can consider the particular test:

$$\psi(L) = \mathbb{1}(L > t) + c\mathbb{1}(L = t)$$

(which may be suboptimal) which has the same type I and type II errors as above. So,

$$T(L(X), L(Y)) \leq T(X, Y)$$

Together, $T(X,Y) = T(L(X), L(Y)) = T(-\mathrm{PLRV}(X \mid Y), \mathrm{PLRV}(Y \mid X))$. ■

Corollary 14.1. If $\mathrm{PLRV}(X \mid Y)$ and $\mathrm{PLRV}(Y \mid X)$ are continuous, with CDFs F and G, then

$$T(X,Y)(\alpha) = G(F^{-1}(1-\alpha))$$

Remark 14.1. We can recover $T(P,Q)$ if we know $\mathrm{PLRV}(P \mid Q)$ and $\mathrm{PLRV}(Q \mid P)$. So, they are also sufficient for Renyi divergence, f-divergence, and ε, (ε, δ)-DP.

Lemma 14. (Assume that P and Q have same support, i.e. PLRV's are finite.) The distribution of $\mathrm{PLRV}(Y \mid X)$ can be expressed as a function of the distribution of $\mathrm{PLRV}(X \mid Y)$. In particular, if

$$\mathrm{PLRV}(X \mid Y) \sim F \text{ and } \mathrm{PLRV}(Y \mid X) \sim G$$

then $G(x) = \displaystyle\int_{-\infty}^{x} e^t dF(t)$, i.e. there exists P and Q such that these are PLRVs.

Proof of Lemma 14. By Fundamental theorem of calculus and Radon-Nikodym theorem,

$$\begin{aligned} G(x) = \int_{-\infty}^{x} dG(t) &= \int_{-\infty}^{\infty} \mathbb{1}\left(\log \frac{dQ}{dP}(w) \leq x\right) dQ(w) \\ &= \int_{-\infty}^{\infty} \mathbb{1}\left(\log \frac{dQ}{dP}(w) \leq x\right) \frac{dQ}{dP}(w) dP(w) \\ &= \int_{-\infty}^{x} e^t dF(t) \end{aligned}$$

■

Lemma 15. The PLRV's are uniquely determined by $f = T(P,Q)$

$$F(x) = 1 - \inf\{t \mid -e^x \text{ is the slope a tangent line to } f \text{ at } t\},$$

where $F(x)$ is the CDF of $\mathrm{PLRV}(X \mid Y)$ with $X \sim P$ and $Y \sim Q$.

Remark 14.2. Note that using f, we have

$$\begin{aligned} f(\alpha) = G(F^{-1}(1-\alpha)) &\implies G(x) = f(1 - F(x)) && (\because \text{Continuity}) \\ &\implies G(x) = \int_{-\infty}^{x} e^t dF(t) && (\because \text{Finite PLRV}) \end{aligned}$$

Proof of Lemma 15 (Sketch). Assume F and G are continuous and $1 - F(x)$ is a point where f is differentiable. Start with $f(1 - F(x)) = G(x)$, take derivative of both sides.

$$\begin{aligned} f'(1 - F(x))(-F'(x)) = G'(x) &\implies f'(1 - F(x)) = -\frac{G'(x)}{F'(x)} = -e^x \\ &\implies 1 - F(x) = (f')^{-1}(-e^x) \end{aligned}$$

■

Corollary 14.2 (Symmetry)**.** If f is symmetric, i.e. $f = T(X, Y) = T(Y, X)$, then

$$\mathrm{PLRV}(X \mid Y) \stackrel{d}{=} \mathrm{PLRV}(Y \mid X)$$

Proposition 14.1 (Composition)**.** Let $(X_1, X_2) \sim P_1 \times P_2$ and $(Y_1, Y_2) \sim Q_1 \times Q_2$, then

$$\mathrm{PLRV}((X_1, X_2) \mid (Y_1, Y_2)) = \mathrm{PLRV}(X_1 \mid Y_1) + \mathrm{PLRV}(X_2 \mid Y_2)$$

Proof of Proposition 14.1. Let $X_1 \sim P_1$ and $X_2 \sim P_2$, then

$$\begin{aligned} \mathrm{PLRV}((X_1, X_2) \mid (Y_1, Y_2)) &= \log \frac{P_1(X_1)P_2(X_2)}{Q_1(X_1)Q_2(X_2)} \\ &= \log \frac{P_1(X_1)}{Q_1(X_1)} + \log \frac{P_2(X_2)}{Q_2(X_2)} \\ &= \mathrm{PLRV}(X_1 \mid Y_1) + \mathrm{PLRV}(X_2 \mid Y_2) \end{aligned}$$

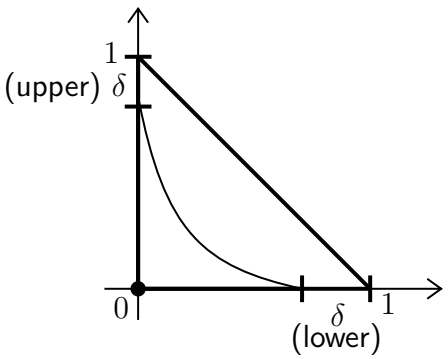
■

Remark 14.3 (Tensor Product)**.** For tensor product of tradeoff functions (composition) $f \otimes g$ where $f = T(P_1, Q_1)$ and $g = T(P_2, Q_2)$, the PLRV for $f \otimes g$ is the same as the sum of the PLRV's for f and g.

So, composition of mechanism is equivalent to tensor product of tradeoff functions or to convolution of PLRVs. Convolution is still difficult, but is easier with Fourier transforms.

Remark 14.4 (Idea)**.** We have 3 ways of calculating δ for an ε

1. $\delta(\varepsilon) = 1 + f^*(-e^\varepsilon))$ is the upper δ
2. $\delta(\varepsilon) = 1 - e^\varepsilon + e^\varepsilon F(\varepsilon)$ is the upper δ, where F is the cdf of PLRV(P, Q)
3. $\delta(\varepsilon) = H_{e^\varepsilon}(P||Q)$ is the lower δ using slope $-e^{-\varepsilon}$



This is sometimes called the privacy profile $(\varepsilon, \delta(\varepsilon))$

14.2 Characteristic Functions

For a real-valued random variable $X \sim F$, its characteristic function is $\varphi_X(t) = \mathbb{E}[e^{itX}] = \int_{-\infty}^{\infty} e^{-itx} dF(x)$.

Remark 14.5. If X and Y are independent, $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$.

Proposition 14.2. Let $f_1, f_2, \cdots, f_k$ be tradeoff functions, and let $(P_1, Q_1), \cdots, (P_k, Q_k)$ be the PLRVs for $f_1, \cdots, f_k$. Let $\varphi_{P_i}(t)$ and $\varphi_{Q_i}(t)$ be the characteristic functions for P_i and Q_i. Then $\varphi_{\sum P_i}(t)$ and $\varphi_{\sum Q_i}(t)$ are the characteristic functions for the PLRVs of $f_1 \otimes \cdots \otimes f_k$.

14.3 Dominating Pairs

Definition 14.2 (Dominating Pair)**.** Let M be a mechanism. We say (P, Q) is a dominating pair for M if

$$T(M(X), M(X'))(\alpha) \geq T(P, Q)(\alpha)$$

for all $H(X, X') \leq 1$. In other words, (P, Q) is a dominating pair if M satisfies $T(P, Q)$-DP.

Equivalently, for all $\alpha \geq 0$,

$$H_\alpha(M(X)||M(X')) \leq H_\alpha(P||Q)$$

for all $H(X, X') \leq 1$, where $H_\alpha(P||Q) = \mathbb{E}_{X\sim Q}[\frac{dP}{dQ}(X) - \alpha]_+$ and setting $\alpha = e^\varepsilon$, H_α returns δ.

Theorem 14.1. If (P, Q) dominates M and (P', Q') dominates M', then $(P \times P', Q \times Q')$ dominates the composed mechanism.

We previously defined a dominating pair (P, Q) for a mechanism M as any pair satisfying $T(P, Q) \leq T(M(D), M(D'))$ for adjacent D and D'. Equivalently, for all $\alpha \geq 0$,

$$H_\alpha(M(D)\|M(D')) \leq H_\alpha(P\|Q)$$

for all adjacent D and D'. Recall that $H_\alpha(P\|Q) = \mathbb{E}_{X\sim Q}\left[\frac{dP}{dQ}(X) - \alpha\right]_+$ and setting $\alpha = e^\varepsilon$, H_α returns "δ."

14.4 Tight Dominating Pairs

We say that (P, Q) is tightly dominating for M if for all lower bounds,

$$f \leq T(M(D), M(D'))$$
$$f \leq T(P, Q) \leq T(M(D), M(D'))$$

Equivalently, for all $\alpha \geq 0$,

$$\sup_{\substack{D \text{ and } D' \\ \text{adjacent}}} H_\alpha(M(D), M(D')) = H_\alpha(P\|Q)$$

Here is a result that seems obvious, and even if it is not, it is good to know.

Proposition 14.3. Any mechanism has a tight dominating pair of distributions. For a tradeoff function, note that supremum of all tradeoff function lower bounds is a tradeoff function, and any tradeoff function can be realized by some pair of distribution.

Example 14.1. Let $P = \mathcal{U}(0,1)$ and Q has a CDF

$$\begin{cases} f(1-x) & \text{when } 0 \leq x < 1 \\ 1 & \text{when } x = 1 \end{cases}$$

So, Q has density $f(1-x)$ and atom at 1, namely $Q(\{1\}) = 1 - f(0)$.

Lemma 16. For $H\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, there exists P and Q s.t. $H(\alpha) = H_\alpha(P\|Q)$ iff

$$H \in \left\{ H\colon \mathbb{R}_{\geq 0} \to \mathbb{R} \;\middle|\; H \text{ is convex, decreasing, } H(0) = 1, \text{ and } H(x) \geq (1-x)_+ \right\}$$

Moreover, we can construct P and Q

$$P \text{ has CDF } 1 + H^*(X - 1)$$
$$Q \sim \mathcal{U}(0,1)$$

based on duality between tradeoff functions and Hockey-stick divergence. H^* here denotes the convex conjugate of H.

A related concept is *worst-case pair of datasets* D and D', which satisfies that $(M(D), M(D'))$ is tightly dominating for M. While there always is a tight dominating pair, there does not always exist a worst-case pair of datasets.

14.5 Rephrasing with Dominating Pairs

Theorem 14.2 (Rephrasing Composition with Dominating Pairs)**.** If (P, Q) dominate M, and (P', Q') dominate M', then $(P \times P', Q \times Q')$ dominates the composed mechanism[a] (M, M').

[a] It may be adaptive. Need to be checked.

Definition 14.3 (Dominating Pairs and Subsampling – Notations)**.**

- **Poisson Subsampling**: Denote by $S^\gamma_{\text{Poisson}}$ which includes each datapoint independently w.p. $0 \leq \gamma \leq 1$.
- **Subset Sampling**: Denote by S^γ_{Subset}, which samples a dataset of size m ($\gamma = m/n$) uniformly at random.

The following result not only considers add/delete notion of adjacency, but treats add and delete separately. It is necessary to get a close-form expression.

Theorem 14.3. Let M be a mechanism.

(1) If (P, Q) dominates M for add neighbors, then
   - $(P, (1-\gamma)P + \gamma Q)$ dominates $M \circ S^{\gamma}_{\text{Poisson}}$ for add neighbors, and
   - $((1-\gamma)Q + \gamma P, Q)$ dominates $M \circ S^{\gamma}_{\text{Poisson}}$ for removing neighbors.

(2) If (P, Q) dominates M for replacing neighbors, then
   - $(P, (1-\gamma)P + \gamma Q)$ dominates $M \circ S^{\gamma}_{\text{Subset}}$ for add neighbors, and
   - $((1-\gamma)Q + \gamma P, Q)$ dominates $M \circ S^{\gamma}_{\text{Subset}}$ for removing neighbors.

To get (ε, δ) guranatee for "add/remove" for k-fold composition of a subsampled mechanism, we take the point-wise maximum

$$\max\left\{ H_{e^\varepsilon}(P_1^k \| Q_1^k), H_{e^\varepsilon}(P_2^k, Q_2^k) \right\}$$

where (P_1, Q_1) is the **remove only** pair and (P_2, Q_2) is the **add only** pair.

Remark 14.6. Some existing literature has mistakenly evaluated the privacy cost of $(\gamma P + (1-\gamma)Q, Q)$ for Poisson subsampled mechanism. However, by doing this, they are essentially only providing privacy guarantees for "remove only" neighbors.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. 67, 68, 69

Jordan Awan and Jinshuo Dong. Log-concave and multivariate canonical noise distributions for differential privacy. *Advances in Neural Information Processing Systems*, 35:34229–34240, 2022. 85

Jordan Awan and Aleksandra Slavković. Structure and sensitivity in differential privacy: Comparing k-norm mechanisms. *Journal of the American Statistical Association*, 116(534):935–954, 2021. 31, 34

Jordan Awan, Ana Kenney, Matthew Reimherr, and Aleksandra Slavković. Benefits and Pitfalls of the Exponential Mechanism with Applications to Hilbert Spaces and Functional PCA. In *International Conference on Machine Learning*, pages 374–384. PMLR, 2019. http://proceedings.mlr.press/v97/awan19a/awan19a.pdf. 46

Rina Foygel Barber and John C. Duchi. Privacy and Statistical Risk: Formalisms and Minimax Bounds. arXiv:1412.4451, 12 2014. https://arxiv.org/abs/1412.4451. 25

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv preprint arXiv:1405.7085*, 2014. 67

Yvonne M. M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Springer New York, NY, 1 edition, 2007. ISBN 978-0-387-72805-6. https://doi.org/10.1007/978-0-387-72806-3. 7

David Blackwell. Comparison of Experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 93–102. University of California Press, 07-08 1951. First appeared as a technical report in August 1950, https://apps.dtic.mil/sti/citations/AD1028649. 80

Mark Bun and Thomas Steinke. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016. https://doi.org/10.1007/978-3-662-53641-4_24. 71

Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. *Advances in neural information processing systems*, 21, 2008. 31

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011. 31

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, page 202–210, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136706. 8, 9

Jinshuo Dong, Aaron Roth, and Weijie Su. Gaussian Differential Privacy. *Journal of the Royal Statistical Society Series B*, 84(1):3–37, 02 2022. https://doi.org/10.1111/rssb.12454. 75

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local Privacy and Statistical Minimax Rates. In *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, FOCS '13, page 429–438, USA, 2013. IEEE Computer Society. ISBN 9780769551357. doi: 10.1109/FOCS.2013.53. https://doi.org/10.1109/FOCS.2013.53. 25

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax Optimal Procedures for Locally Private Estimation. *Journal of the American Statistical Association (JASA)*, 113(521): 182–201, 2018. doi: 10.1080/01621459.2017.1389735. https://doi.org/10.1080/01621459.2017.1389735. 25

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042. http://dx.doi.org/10.1561/0400000042. 11, 59, 61, 63

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006. 31

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What Can We Learn Privately? In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '08, pages 531–540, 2008. doi: 10.1109/FOCS.2008.27. https://doi.org/10.1109/FOCS.2008.27. Full version published in 2011 SIAM Journal on Computing: https://doi.org/10.1137/090756090. 65

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 25.1–25.40, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. 32

Jing Lei, Anne-Sophie Charest, Aleksandra Slavkovic, Adam Smith, and Stephen Fienberg. Differentially Private Model Selection with Penalized and Constrained Likelihood. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3):609–633, 2018. https://doi.org/10.1111/rssa.12324. 29

Ninghui Li, Min Lyu, Dong Su, and Weining Yang. *Differential privacy: From theory to practice*. Springer, 2017. 21

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, 2007. doi: 10.1109/FOCS.2007.66. 36

Ilya Mironov. Renyi Differential Privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017. doi: 10.1109/CSF.2017.11. https://doi.org/10.1109/CSF.2017.11. 71

Jack Murtagh and Salil Vadhan. The Complexity of Computing the Optimal Composition of Differential Privacy. In Eyal Kushilevitz and Tal Malkin, editors, *Theory of Cryptography*, pages 157–175, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg. ISBN 978-3-662-49096-9. https://doi.org/10.1007/978-3-662-49096-9_7. 64

Matthew Reimherr and Jordan Awan. Kng: The k-norm gradient mechanism. *Advances in neural information processing systems*, 32, 2019. 49

Adam Smith. Differential Privacy and the Secrecy of the Sample. "Oddly Shaped Pegs" (Personal webpage), 09 2009. https://adamdsmith.wordpress.com/2009/09/02/sample-secrecy/. 65

Adam Smith. Privacy-Preserving Statistical Estimation with Optimal Convergence Rates. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11, page 813–822, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306911. doi: 10.1145/1993636.1993743. https://doi.org/10.1145/1993636.1993743. 53, 54

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pages 245–248. IEEE, 2013. 67

Thomas Steinke and Jonathan Ullman. Between Pure and Approximate Differential Privacy. *Journal of Privacy and Confidentiality*, 7(2):3–22, 2016. https://doi.org/10.29012/jpc.v7i2.648. 62

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal Accounting of Differential Privacy via Characteristic Function . In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4782–4817. PMLR, 03 2022. https://proceedings.mlr.press/v151/zhu22c.html. 90