Privacy-Preserving Natural Language Processing

Lecture 6 – Approximate Differential Privacy and Gaussian Mechanism

Prof. Dr. Ivan Habernal

June 5, 2025

www.trusthlt.org
Chair of Trustworthy Human Language Technologies (TrustHLT)
Ruhr University Bochum & Research Center Trustworthy Data Science and Security

Recap

What we covered so far

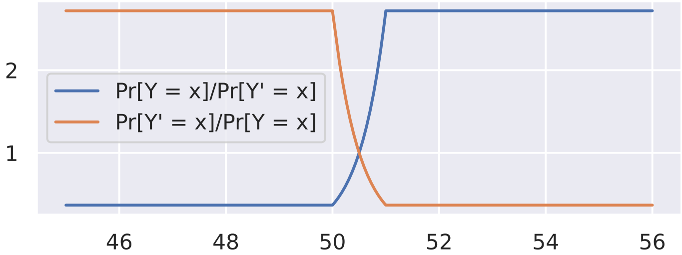
- For provably private data analysis we need randomized algorithms
- Central (with a trusted curator) pure $(\varepsilon, 0)$ differential privacy
- Laplace mechanism: numeric queries, ℓ_1 sensitivity,
- Exponential mechanism: ‘any-range’ queries (arbitrary sets), utility function and its sensitivity
- Local DP

Today

Approximate DP

Privacy Loss Random Variable

Previously: Can we generalize it for any observed x?



Seems like the maximum we can get is $2.718 = e = \exp(1)$

Previously: How does that relate to the maximum privacy loss?

Recall: likelihood of any output (x-axis) coming from D' as opposed do D (and vice versa)

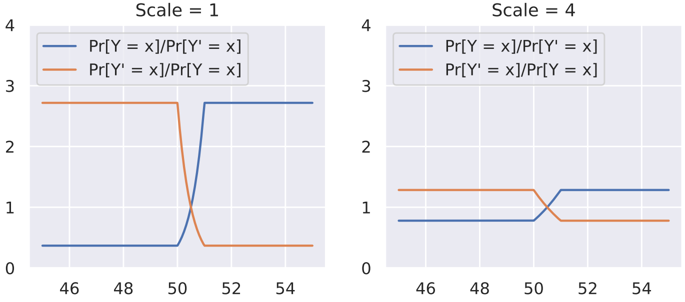


Figure 1: Privacy loss for two Laplace distributions for a counting query, varying scale b

Recall: $(\varepsilon, 0)$ differential privacy (aka. pure DP)

C. Dwork and A. Roth (2013). **"The Algorithmic Foundations of Differential Privacy".** In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4, pp. 211–407, Definition 2.4

A randomized algorithm (mechanism) $\mathcal{M}$ is $(\varepsilon, 0)$-**differentially private** if for any two neighboring datasets D, D' and any output $\mathcal{Y} \subseteq \mathcal{Z}$ this guarantee holds:

$$\Pr[\mathcal{M}(D) \in \mathcal{Y}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in \mathcal{Y}]$$

We bounded our 'privacy loss' by ε

$$\Pr[\mathcal{M}(D) \in \mathcal{Y}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in \mathcal{Y}]$$

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

What is $\mathcal{M}(D)$ (and also $\mathcal{M}(D')$)?

We bounded our ‘privacy loss’ by ε

$$\Pr[\mathcal{M}(D) \in \mathcal{Y}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in \mathcal{Y}]$$

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

What is $\mathcal{M}(D)$ (and also $\mathcal{M}(D')$)?

The private mechanism is randomized, so somewhere in the mechanism there is a random variable

- e.g., Laplace mechanism uses Laplace R.V.
- Randomized response uses Bernoulli R.V., etc.

In general, since the mechanism $\mathcal{M}(D)$ is a function of a random variable, it is also a random variable

Towards the 'privacy loss' random variable

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

$\mathcal{M}(D)$ (but also $\mathcal{M}(D')$) are random variables

Towards the 'privacy loss' random variable

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

$\mathcal{M}(D)$ (but also $\mathcal{M}(D')$) are random variables

In general, since the mechanism $\mathcal{M}(D)$ is a random variable, the entire left-hand side function $\ln \left(\frac{\Pr[\mathcal{M}(D) \in \mathcal{Y}]}{\Pr[\mathcal{M}(D') \in \mathcal{Y}]} \right)$ is again a random variable

(recall: random variables can be 'pushed through' functions. If X is a random variable, then $Y = g(X)$ is also a random variable. Here the g is a complicated function even including probability of X)

Step aside: You know functions of RV having similar form

If X is a discrete random variable

Expectation of X?

Step aside: You know functions of RV having similar form

If X is a discrete random variable

Expectation of X?

$$\mathbb{E}(X) = \sum_{x \in \text{Range}(X)} x \cdot \Pr[X = x]$$

Entropy of X? (notice lazy notation for $P[X]$ and $\sum_x$)

$$\mathbb{H}(X) = \mathbb{E}\left(\log \frac{1}{P[X]}\right) = -\sum_x x \cdot \log \Pr[X = x]$$

RUHR UNIVERSITÄT BOCHUM RUB

Step aside: You know functions of RV having similar form

KL-Divergence between X and Y (same range)?

$$\mathbb{D}(X||Y) = \mathbb{E}\left[\log \frac{P(X)}{P(Y)}\right]$$

Here we implicitly assume $\log \frac{P(X)}{P(Y)}$ is a function of X and is therefore distributed according to X

$$\mathbb{D}(X||Y) = \mathbb{E}\left[\log \frac{P(X)}{P(Y)}\right] = \mathbb{E}[g(X)] = \sum_{x \in \text{Range}(X)} \Pr[X = x] \cdot g(x)$$

$$= \sum_{x \in \text{Range}(X)} \Pr[X = x] \log \frac{\Pr[X = x]}{\Pr[Y = x]}$$

Privacy Loss Random Variable

$\mathcal{M}(D)$ and $\mathcal{M}(D')$ are two random variables

The privacy loss random variable is defined as

$$\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')} = \ln \left(\frac{\Pr[\mathcal{M}(D) = t]}{\Pr[\mathcal{M}(D') = t]} \right)$$

and is distributed by drawing $t \sim \mathcal{M}(D)$

(Sanity check: You should know how to compute the expectation of the privacy loss R.V. given the previous slides)

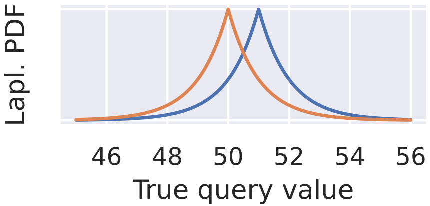
Example of Privacy Loss Random Variable

The privacy loss random variable

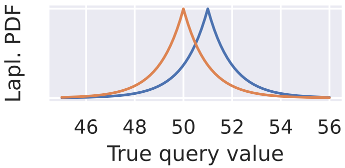
$$\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')} = \ln \left(\frac{\Pr[\mathcal{M}(D) = t]}{\Pr[\mathcal{M}(D') = t]} \right)$$

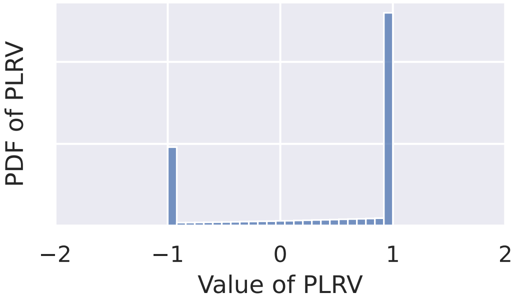
and is distributed by drawing $t \sim \mathcal{M}(D)$

How would the distribution of $\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}$ would look like for the Laplace mechanism?

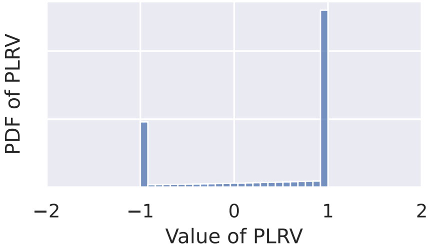


Example of $\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}$ for Laplace mechanism $\varepsilon = 1$





Values of $\left|\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}\right|$ are upper-bounded by ε in $(\varepsilon, 0)$-DP



This distribution demonstrates (not a proof!) that the probability the value of $\left|\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}\right|$ exceeds ε is zero

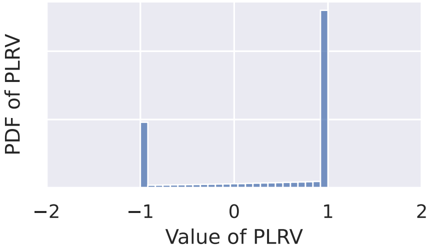
In other words

$$\Pr\left[\left|\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}\right| \leq \varepsilon\right] = 1$$

Approximate Differential Privacy

Maybe we don't need to always ensure the bound

What if we allow to exceed ε with some small probability δ?



In other words change $\Pr\left[\left|\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}\right| \leq \varepsilon\right] = 1$ into

$$\Pr\left[\left|\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}\right| \leq \varepsilon\right] \geq 1 - \delta$$

$$\Pr\left[\left|\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}\right| > \varepsilon\right] < \delta \qquad \text{(equivalent)}$$

Formalizing approximate (ε, δ)-DP

C. Dwork and A. Roth (2013). **"The Algorithmic Foundations of Differential Privacy".** In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4, pp. 211–407, Definition 2.4

A randomized algorithm (mechanism) $\mathcal{M}$ is (ε, δ)-**differentially private** if for any two neighboring datasets D, D' and any output $\mathcal{Y} \subseteq \mathcal{Z}$ this guarantee holds:

$$\Pr[\mathcal{M}(D) \in \mathcal{Y}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in \mathcal{Y}] + \delta$$

One immediate observation: for $\delta = 0$ we get our known 'pure' DP (that's why we called it (ε, δ)-DP)

Formalizing approximate (ε, δ)-DP

C. Dwork and A. Roth (2013). **"The Algorithmic Foundations of Differential Privacy".** In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4, pp. 211–407, Definition 2.4

$$\Pr[\mathcal{M}(D) \in \mathcal{Y}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in \mathcal{Y}] + \delta$$

This is equivalent to say[1] that the P.L.R.V. is bounded by ε with probability $1 - \delta$

$$\Pr\left[\left|\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}\right| \leq \varepsilon\right] \geq 1 - \delta$$

[1] The proof is lengthy and technical, see Dwork and Roth (2013, pp. 44–47)

What is this δ doing?

Extreme algorithm 1: When bad things are really bad

Our query is: Given a database of secrets, give me all rows

Name	Hospitalized in year	Age	Illegal drug use
Alice	2024	32	yes
Bob	2020	21	no
Charlie	2023	45	no
...			
Xander	2020	31	yes

Table 1: Example database D

Our goal is to have this algorithm (ε, δ)-DP

Given a database of secrets, give me all rows (part 1)

With probability $1 - \delta$, return completely random table

Name	Hospitalized in year	Age	Illegal drug use
Jim	2022	16	no
Dave	2011	71	yes
...			

Table 2: Example output of $\mathcal{M}(D)$ — completely random

Since the output is completely random, there would be no difference in outputs of any neighboring datasets D and D', therefore this is perfectly private algorithm $\varepsilon = 0$

Given a database of secrets, give me all rows (part 2)

With probability δ, return the **original** dataset in full

Name	Hospitalized in year	Age	Illegal drug use
Alice	2024	32	yes
Bob	2020	21	no
...			

Table 3: Example output of $\mathcal{M}(D)$ — returning the full original D

This part of the algorithm is purely deterministic, there is no randomness, therefore this would be $\varepsilon = \infty$

Why? Remove Alice to get D'. But this algorithm is never going to return D', so $\Pr[\mathcal{M}(D') = 0]$, which leads to $\frac{\Pr[\mathcal{M}(D)=x]}{\Pr[\mathcal{M}(D')=0]} \to \infty$

Given a database of secrets, give me all rows (part 3)

Summary of our algorithm:

- With prob. $1 - \delta$, return completely random table ($\varepsilon = 0$)
- With prob. δ, return the **original** dataset in full ($\varepsilon = \infty$)

Our algorithm is (ε, δ)-DP! (in fact $(0, \delta)$-DP)

$$\Pr\left[\left|\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}\right| > \varepsilon\right] < \delta$$

Very bad things can happen with δ, so it should be very small! But how small?

Extreme algorithm 2: Leak just a few rows

Our query is: Given a database of secrets, give me a few rows verbatim

Name	Hospitalized in year	Age	Illegal drug use
Alice	2024	32	yes
Bob	2020	21	no
Charlie	2023	45	no
...			
Xander	2020	31	yes

Table 4: Example database D

Our goal is to have this algorithm $(0, \delta)$-DP

Extreme algorithm 2: Leak just a few rows

Our algorithm 2:

- Iterate over all n rows
- For each row **independently**, with probability δ add this row to the output[2]

Name	Hospitalized in year	Age	Illegal drug use
Alice	2024	32	yes
Bob	2020	21	no
...			

Table 5: Example database D

[2] For instance by drawing from $\mathrm{Ber}(\delta)$

Extreme algorithm 2: Leak just a few rows (part 2)

Our algorithm 2:

- Iterate over all n rows
- For each row **independently**, with probability δ add this row to the output

This has again bad consequences! Probability that at least one person will be leaked?[3]

$$1 - (1 - \delta)^n \approx \delta n \qquad \text{for small } \delta$$

[3] Why? Pr. that a single person will not be leaked = $(1 - \delta)$. Pr. that no persons will be leaked = $(1 - \delta)^n$

Extreme algorithm 2: Leak just a few rows (part 3)

Probability that at least one person will be leaked?

$$1 - (1 - \delta)^n \approx \delta n \qquad \text{for small } \delta$$

General recommendation

We should therefore consider

$$\delta \ll \frac{1}{n}$$

(ie. very small; typically $\delta = 1 \times 10^{-6}$, aka 'cryptographically' small)

Gaussian mechanism

ℓ_2 sensitivity

Similar to ℓ_1 sensitivity of the query

ℓ_2 sensitivity

The ℓ_2-sensitivity of a function $f : D \to \mathbb{R}^k$:

$$\Delta_2 f = \max_{D,D'} \|f(D) - f(D')\|_2$$

Gaussian (Normal) random variable

The density (PDF) of a general univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Gaussian mechanism

Function (numeric query) $f : D \to \mathbb{R}^k$:

Very important constraints on ε!

For $\varepsilon \in (0, 1)$ and $\delta > 0$

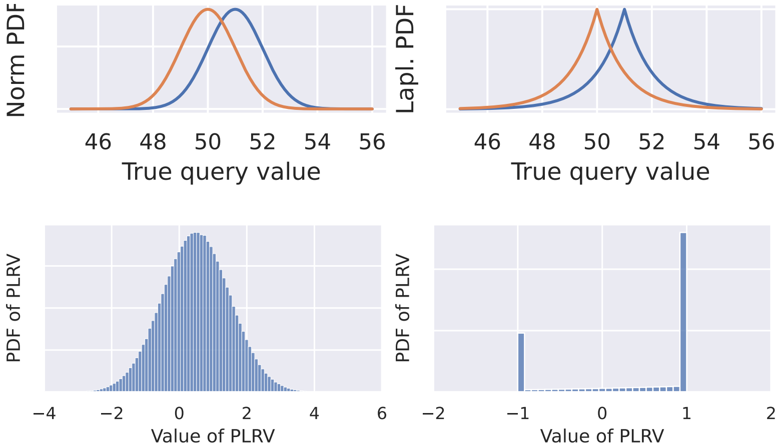
The Gaussian mechanism $\mathcal{M}(D)$ is defined as

$$f(D) + (Y_1, \ldots, Y_k)$$

where each Y_n is drawn **independently** from $\mathcal{N}(0, \sigma^2)$, such that

$$\sigma^2 > 2 \ln \left(\frac{1.25}{\delta} \right) \frac{(\Delta_2)^2}{\varepsilon^2}$$

Gaussian mechanism is (ε, δ)-DP[4]



[4]Non-trivial proof in Appendix A of Dwork and Roth (2013); also note that there are quite a few typos there

General properties of DP algorithms

Post-processing

Let $\mathcal{M}(D) \mapsto R$ be a (ε, δ)-DP algorithm

Let $f : R \mapsto S$ be an arbitrary (randomized) function

Then $f(\mathcal{M}(D))$ is (ε, δ)-DP

In words

Whatever you do with (ε, δ)-DP output, you cannot 'weaken' privacy

Group privacy

Let D and D' differ in k positions.

Let $\mathcal{M}(D)$ be (ε, δ)-DP

Then for any output T we have

$$\Pr[\mathcal{M}(D) \in T] \leq \exp(k\varepsilon) \Pr[\mathcal{M}(D') \in T] + k \exp(\varepsilon \cdot (k-1))\delta$$

Implications for large groups

If k grows, the privacy budget grows exponentially

Basic composition

Let $\mathcal{M} = (M_1, \ldots, M_k)$ be a sequence of mechanisms, where each M_i is $(\varepsilon_i, \delta_i)$-DP. (They might be adaptive)

Then $\mathcal{M}$ is $(\sum_{i=1}^{k} \varepsilon_i, \sum_{i=1}^{k} \delta_i)$-DP

In words

Overall privacy 'budget' can be spent for a sequence of private queries

License and credits

Credits

Ivan Habernal

Content from ACL Anthology papers licensed under CC-BY
`https://www.aclweb.org/anthology`

Partly inspired by lectures from Gautam Kamath