EACL 2023 Tutorial on Privacy-Preserving NLP

Block 2a, Part 1: Defence with formal guarantees

Dr. Ivan Habernal

May 6, 2023

Trustworthy Human Language Technologies
Department of Computer Science
Technical University of Darmstadt

TrustHLT
1001100110101

www.trusthlt.org

Differential privacy: Setup and terminology

- Differential privacy: Setup and terminology
- Basic DP mechanisms
- Approximate DP
    - Pure and approximate DP
    - Composition
- Some application of differential privacy in NLP
    - Publishing models trained with DP
    - Synthetic data generation with DP-trained models
    - Data publishing with local differential privacy

Trusted curator and dataset

The basic settings: A **trusted curator** holds a **database** or **dataset** (i.e., a table) where

- Each entry is linked to a single person
- Each person in the dataset is linked to a single entry

In other words, the rows are completely independent

Name	Income	Last facebook post	...
Alice	60,000	"I hate my job, ..."	...
Bob	90,000	"Spending time with my ..."	...
...	...	...	...

Datasets formally

Name	Income	Last facebook post	...
Alice	60,000	”I hate my job, ...”	...
Bob	90,000	”Spending time with my ...”	...
...	...	...	...

S. Vadhan (2017). “The Complexity of Differential Privacy”. In: *Tutorials on the Foundations of Cryptography*. Ed. by Y. Lindell. Information Security and Cryptography. Cham: Springer International Publishing. Chap. 7, pp. 347–450

Dataset x with n individuals as a n-tuple

$$x = (x_1, x_2, \ldots, x_n)$$

Note: Many different notations, e.g., D, X, ...

x_1 corresponds to Alice, x_2 to Bob, etc.

Neighboring datasets

Definition of neighboring datasets

Neighboring datasets x and x' differ on one row

- Variant 1: x and x' have the same size ($|x| = |x'|$), one person is replaced
- Variant 2: $|x'| = |x| \pm 1$, one person is removed or added

These two options will not protect the same thing.

- Variant 1: will protect the value of the record
- Variant 2: will also protect the presence of the record in the data

D. Desfontaines and B. Pejó (2020). "SoK: Differential privacies". In: *Proceedings on Privacy Enhancing Technologies* 2020.2, pp. 288–313

Some examples of neighboring datasets x and x' (Variant 2)

Alice	60,000	”I hate my job, ...”	...
Bob	90,000	”Spending time with ...”	...
Charlie	50,000	”Feeling great”	...

Table 1: Dataset x

Bob	90,000	”Spending time with ...”	...
Charlie	50,000	”Feeling great”	...

Table 2: Dataset x'

$$|x| = 3, |x'| = 2$$

Some examples of neighboring datasets x and x' (Variant 1)

Alice	60,000	”I hate my job, ...”	...

Table 3: Dataset x

Bob	90,000	”Spending time with ...”	...

Table 4: Dataset x'

$$|x| = 1, |x'| = 1$$

Note: Dataset of size 1 might look a bit extreme, but it follows the definition and will be useful for local DP.

Query

An **analyst** or **adversary** wants to get information about a dataset x

Numerical queries

- Counting query: $f(x) \to \mathbb{N}$
- Mean query: $f(x) \to \mathbb{R}$
- General numerical query: $f(x) \to \mathbb{R}^k$

Arbitrary queries

- $f(x) \to \mathcal{Y}$

Transformation of the dataset x to any object

Example of numeric queries

Alice	60,000	”I hate my job, ...”	...
Bob	90,000	”Spending time with ...”	...
Charlie	50,000	”Feeling great”	...

Table 5: Dataset x

- Counting query: How many person? $f(x) = 3$
- Mean query: Mean income? $f(x) = 6.667$
- General numerical query: Histogram of word counts in all Facebook posts: $f(x) = \{\text{my} = 5, \text{the} = 4, \ldots\}$
- Arbitrary query: Get copy of Facebook posts $f(x) = ($”I hate my job, ...”$, \ldots)$

Further example queries

Alice	60,000	"I hate my job, ..."	...
Bob	90,000	"Spending time with ..."	...
Charlie	50,000	"Feeling great"	...

Table 6: Dataset x

- Return Facebook posts $f(x) = (\text{"I hate my job, ..."}, \ldots)$
- Use the income as a predicted variable y_i, use the Facebook text of x_i as an input to a given neural network for income prediction, run a forward pass and backprop, and return the gradient: $f(x) = \left(\frac{\partial E}{\partial w_1} = 0.02, \frac{\partial E}{\partial w_2} = -24.9, \ldots \right)$

Pure differential privacy

The trusted curator **randomizes** the query result

- Counting query: How many person? $f(x) = 3 +$ value drawn from a certain probability distribution

This is called a (randomized) **mechanism** (algorithm) $\mathcal{M}$ (or $\mathcal{M}(f(x))$, $\mathcal{M}(x)$, $\mathcal{M}(f, x)$, etc.)

What pure differential privacy guarantees ($\varepsilon \geq 0$)

For every pair of neighboring datasets x, x' and for any output $y \in \mathcal{Y}$ of $\mathcal{M}$ and query f

$$\frac{\Pr\left[\mathcal{M}(f(x)) = y\right]}{\Pr\left[\mathcal{M}(f(x')) = y\right]} \leq \exp(\varepsilon)$$

Pure $(\varepsilon, 0)$ differential privacy

What pure differential privacy guarantees ($\varepsilon \geq 0$)

For every pair of neighboring datasets x, x' and for any output $y \in \mathcal{Y}$ of $\mathcal{M}$ and query f

$$\frac{\Pr\left[\mathcal{M}(f(x)) = y\right]}{\Pr\left[\mathcal{M}(f(x')) = y\right]} \leq \exp(\varepsilon)$$

Privately querying a dataset x with Alice and dataset x' without Alice will give me two results — and probabilities of these two results must be similar up to factor $\exp(\varepsilon)$

Why probabilities $\Pr$? Because of the randomness in $\mathcal{M}$

Basic DP mechanisms

- Differential privacy: Setup and terminology
- Basic DP mechanisms
- Approximate DP
    - Pure and approximate DP
    - Composition
- Some application of differential privacy in NLP
    - Publishing models trained with DP
    - Synthetic data generation with DP-trained models
    - Data publishing with local differential privacy

Laplace mechanism for numeric queries

PDF of the Laplace distribution:

$$\mathrm{Lap}(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

To privatize a query output by adding noise from the Laplace distribution, we need to know the **scale** b.

Global ℓ_1 sensitivity Δf

What is the maximum ℓ_1 difference of the query f for any two neighboring datasets x, x'

$$\Delta f = \max_{x,x'} |f(x) - f(x')|$$

Global sensitivity and Laplace mechanism

Example

Sensitivity of the counting query $\Delta f = 1$

The scale b of the Laplace noise is then proportional to the sensitivity and privacy budget ε:

$$b = \frac{\Delta f}{\varepsilon}$$

The Laplace mechanism

$$\mathcal{M}_{\text{Lap}}(f(x)) = f(x) + z \qquad z \sim \text{Lap}(\mu = 0; b = \Delta f / \varepsilon)$$

Example of the Laplace mechanism for a counting query

Laplace mechanism satisfies (ε)-DP[1]

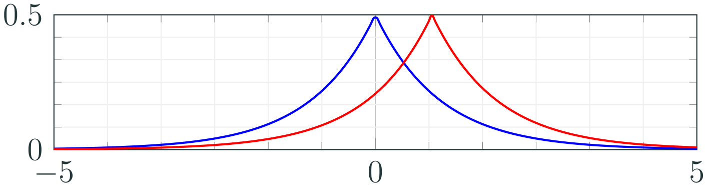

Figure 1: Laplace PDFs for two means: 0 and 1 corresponding to the max difference of two datasets for counting queries with $\Delta f = 1$. Plotted for $\varepsilon = 1.0$.

[1] Step-by-step proof at https://blog.ivanhabernal.com/2020-11-10-detailed-proof-of-laplace-mechanism-in-differential-privacy.html

Exponential mechanism for discrete outputs

Given some arbitrary range (set) $\mathcal{R}$, the **exponential mechanism** is defined to some utility function $u : x \times \mathcal{R} \to \mathbb{R}$, which maps database/output pairs to utility scores

Example

x is a word, $\mathcal{R}$ is a set of words, the utility u is some similarity or word x and words in $\mathcal{R}$, e.g., a similarity of word embeddings (the higher, the better)

We prefer that the mechanism outputs some element of $\mathcal{R}$ with the maximum possible utility score

Sampling from soft-max in language generation: J. Mattern, B. Weggenmann, and F. Kerschbaum (2022). "The Limits of Word Level Differential Privacy". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, WA, USA: Association for Computational Linguistics, pp. 867–881

Exponential mechanism is ε-DP

Sensitivity of the utility score Δu is defined

$$\Delta u = \max_{\forall r \in \mathcal{R}, x, x'} |u(x, r) - u(y, r)|$$

The exponential mechanism[2] $\mathcal{M}_E(x, u, \mathcal{R})$ selects and outputs an element $r \in \mathcal{R}$ with probability proportional to

$$\exp\left(\frac{\varepsilon \cdot u(x, r)}{2\Delta u}\right)$$

[2] For a step-by-step proof see https://blog.ivanhabernal.com/2021-06-30-detailed-proof-of-exponential-mechanism.html

Approximate DP

Differential privacy: Setup and terminology

Basic DP mechanisms

Approximate DP

- Pure and approximate DP
- Composition

Some application of differential privacy in NLP

- Publishing models trained with DP
- Synthetic data generation with DP-trained models
- Data publishing with local differential privacy

Approximate DP

Pure and approximate DP

Privacy loss random variable

Recall: Definition of pure differential privacy

$$\frac{\Pr\left[\mathcal{M}(f(x)) = y\right]}{\Pr\left[\mathcal{M}(f(x')) = y\right]} \leq \exp(\varepsilon)$$

The mechanism $\mathcal{M}(f(x))$ is in fact a **random variable** parametrized by the query f and the dataset x, and as such it has a certain **probability distribution**

We define the privacy loss random variable $L^{(y)}_{(\mathcal{M}(f(x)) \| \mathcal{M}(f(x')))}$ as a function of $\mathcal{M}(f(x))$

$$L^{(y)}_{(\mathcal{M}(f(x)) \| \mathcal{M}(f(x')))} = \ln\left(\frac{\Pr\left[\mathcal{M}(f(x)) = y\right]}{\Pr\left[\mathcal{M}(f(x')) = y\right]}\right) \leq \varepsilon$$

M. Bun and T. Steinke (2016). "Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds". In: *Proceedings of the 14th International Conference on Theory of Cryptography*. Ed. by M. Hirt and A. Smith. Beijing, China: Springer, Berlin, Heidelberg, pp. 635–658

From pure $(\varepsilon, 0)$-DP to approximate (ε, δ)-DP

Privacy loss random variable $L^{(y)}_{(\mathcal{M}(f(x)) \| \mathcal{M}(f(x')))}$, or simply L

$$L = \ln \left(\frac{\Pr\left[\mathcal{M}(f(x)) = y\right]}{\Pr\left[\mathcal{M}(f(x')) = y\right]} \right) \leq \varepsilon \quad \iff \quad \Pr\left[L \leq \varepsilon\right] = 1$$

The random variable L is strictly bounded by ε value

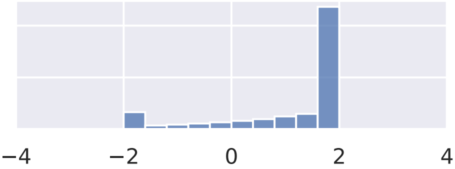

Figure 2: Example distribution of privacy loss random variable for counting query with Laplace mechanism and $\varepsilon = 2$ (x-axis)

From pure $(\varepsilon, 0)$-DP to approximate (ε, δ)-DP

Pure $(\varepsilon, 0)$-DP

$$\Pr\left[L \leq \varepsilon\right] = 1$$

Pure $(\varepsilon, 0)$-DP is very strict. Let's add some relaxation with a 'cryptographically' small $\delta \approx 10^{-6}$

Approximate (ε, δ)-DP

$$\Pr\left[L \leq \varepsilon\right] = 1 - \delta \quad \iff \quad \Pr\left[L \geq \varepsilon\right] = \delta$$

Privacy loss random variable now can have 'tails' beyond ε, but they must remain 'small'

The Gaussian mechanism for (ε, δ)-DP

Global sensitivity now uses the Euclidian (ℓ_2) norm:

$$\Delta f = \max_{x,x'} \|f(x) - f(x')\|$$

The Gaussian mechanism

$$\mathcal{M}_{\text{Gaus}}(f(x)) = f(x) + z \qquad z \sim \mathcal{N}(0; \sigma^2 I)$$

The ‘classical’ Gaussian mechanism for $\varepsilon \leq 1$ (!!) and $\delta \in (0, 1)$:

$$\sigma = \frac{\Delta}{\varepsilon} \sqrt{2 \ln \left(\frac{1.25}{\delta} \right)}$$

Proof in Appendix B of 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

Revisiting the Gaussian mechanism for any ε value

The Analytic Gaussian mechanism

$$\mathcal{M}_{\text{Gaus}}(f(x)) = f(x) + z \qquad z \sim \mathcal{N}(0; \sigma^2 I)$$

Where

- σ satisfies certain inequality (Balle and Wang, 2018, Eq. 6)
- Computed numerically using the error function of the Gaussian (Balle and Wang, 2018, Algorithm 1)

B. Balle and Y. X. Wang (2018). "Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Stockholm, Sweden: PMLR, pp. 678–692

Approximate DP

Composition

Running several mechanisms on the same data

Composition theorems: Running the same or various privacy mechanisms on the same data

Basic composition — “epsilons and deltas add up”

For $k \in \mathbb{N}$, the composition of k mechanisms (each of them is (ε, δ)-DP) gives $(k\varepsilon, k\delta)$-DP

k-fold adaptive composition of an (ε, δ)-DP mechanism

Advanced composition — using smaller overall budget

For $\delta' > 0$ and $\varepsilon' = \varepsilon\sqrt{2k \ln(1/\delta')} + k\varepsilon(\exp(\varepsilon) - 1)$ the composite mechanism is $(\varepsilon', k\delta + \delta')$-DP

Theorem III.3 in C. Dwork, G. N. Rothblum, and S. Vadhan (2010). “Boosting and Differential Privacy”. In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. Las Vegas, USA: IEEE, pp. 51–60

Some application of differential privacy in NLP

- Differential privacy: Setup and terminology
- Basic DP mechanisms
- Approximate DP
  - Pure and approximate DP
  - Composition
- Some application of differential privacy in NLP
  - Publishing models trained with DP
  - Synthetic data generation with DP-trained models
  - Data publishing with local differential privacy

Some application of differential privacy in NLP

Publishing models trained with DP

Stochastic graident descent with differential privacy

Setup: A set of labeled i.i.d. examples — like tabular data (each example = single person)

Privacy 'accountant' — utilizes composition of DP

- Computes the privacy cost at each access to the training data (gradient computation)
- Accumulates this cost as the training progresses

Tightest privacy by numerical integration to get bounds on the **moment generating function** of the **privacy loss random variable** for all moments ≤ 32

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang (2016). "Deep Learning with Differential Privacy". In: Vienna, Austria: ACM, pp. 308–318

DP-SGD algorithm

```
function DP-SGD($f(\boldsymbol{x}; \Theta)$, $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, $|L|$ — ‘lot’ size, $T$ — # of steps)
    for $t \in (1, 2, \dots, T)$ do
        Add each training example to a ‘lot’ $L_t$ with probability $|L|/n$
        for each example in the ‘lot’ $\boldsymbol{x}_i \in L_t$ do
            $\boldsymbol{g}(\boldsymbol{x}_i) \leftarrow \nabla \mathcal{L}(\theta_t, \boldsymbol{x}_i)$                ▷ Compute gradient
            $\bar{\boldsymbol{g}}(\boldsymbol{x}_i) \leftarrow \boldsymbol{g}(\boldsymbol{x}_i) / \max \left(1, \|\boldsymbol{g}(\boldsymbol{x}_i)\| / C\right)$    ▷ Clip gradient
            $\tilde{\boldsymbol{g}}(\boldsymbol{x}_i) \leftarrow \bar{\boldsymbol{g}}(\boldsymbol{x}_i) + \mathcal{N}(0, \sigma^2 C^2 \boldsymbol{I})$      ▷ Add noise
        $\hat{\boldsymbol{g}} \leftarrow \frac{1}{|L|} \sum_{k=1}^{|L|} \tilde{\boldsymbol{g}}(\boldsymbol{x}_k)$     ▷ Gradient estimate of ‘lot’ by averaging
        $\Theta_{t+1} \leftarrow \Theta_t - \eta_t \hat{\boldsymbol{g}}$     ▷ Update parameters by gradient descend
    return $\Theta$
```

Applications of DP-SGD in NLP

De-facto standard treatment of training neural nets with differential privacy

- Private fine-tuning of classifiers

How about LLM pre-training?

- Technically challenging — per-example clipping computationally prohibitive
- How is the next-word prediction objective compatible with the perspective of ‘persons = i.i.d. rows in a table’?

Some application of differential privacy in NLP

Synthetic data generation with DP-trained models

Fine-tune generative model to produce synthetic data

Method: fine-tune generative LM with DP-SGD

- GPT-2 with Opacus (DP), $\varepsilon = 4$
- Yelp reviews, generation conditioned on ‘control codes’ (e.g., review score)
- Downstream classification on synthetic data with RoBERTa

X. Yue, H. A. Inan, X. Li, G. Kumar, J. McAnallen, H. Sun, D. Levitan, and R. Sim (2022). “Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe”. In: *arXiv preprint*

Some application of differential privacy in NLP

Data publishing with local differential privacy

No trusted curator scenario (local differential privacy)

Main idea: Each individual gives already ‘privatized’ version of the example

Privatizing text

- Convert to a latent vector representation
- Add DP noise
- Decode back to text

Inherent challenges

- High dimensionality → very large noise required

Open question: Are meaningful ε values achievable?

T. Igamberdiev and I. Habernal (2023). “DP-BART for Privatized Text Rewriting under Local Differential Privacy”. In: *arXiv preprint*

Recap

- Differential privacy: Setup and terminology
- Basic DP mechanisms
- Approximate DP
  - Pure and approximate DP
  - Composition
- Some application of differential privacy in NLP
  - Publishing models trained with DP
  - Synthetic data generation with DP-trained models
  - Data publishing with local differential privacy

Take aways

- Differential privacy provides formal guarantees with exact probability bounds on the privacy loss
- Fully supervised training with DP-SGD is a go-to option
- LM pre-training, data generation, local DP, word-level privacy, etc. are very tricky to get right

License and credits

Credits

Ivan Habernal

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