Privacy

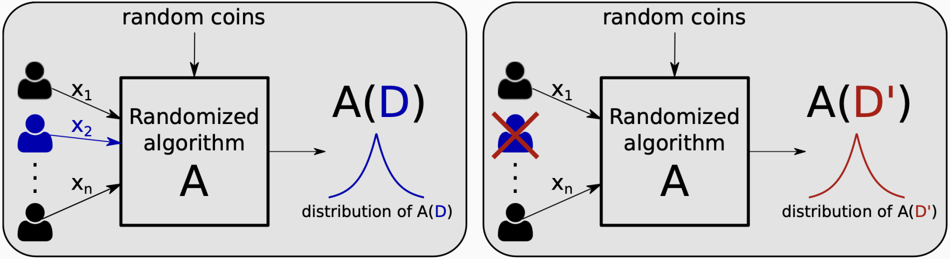
Differential Privacy (Part II)

Guillaume Raschia — Nantes Université

Last update: February 5, 2025

original slides from A. Bellet (Inria), M2DS Univ. Lilles
also from A. Machanavajjhala, M. Hay, X. He *Differential Privacy in the Wild*, VLDB'16 & SIGMOD'17 Tutorial
also from J.P. Near and C. Abuah *Programmming Differential Privacy*

(Figure inspired from R. Bassily)



Definition (Differential Privacy)

An algorithm $\mathcal{A}$ preserves differential privacy if for any pair of neighboring datasets D and D', and for all possible sets of outputs S:

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{A}(D') \in S], \quad \varepsilon > 0$$

REMINDER: GLOBAL SENSITIVITY

Definition (Global ℓ_1 sensitivity)

The global ℓ_1 sensitivity of a query (function) $f : \mathcal{X}^n \to \mathbb{R}$ is

$$\boldsymbol{\Delta}_1(f) = \max_{D,D' : D \Delta D' \leq 1} |f(D) - f(D')|_1$$

How much adding or removing a single record can change the value of the query, measured in ℓ_1 norm

REMINDER: THE LAPLACE MECHANISM

Algorithm: Laplace mechanism $\mathcal{A}_{\text{Lap}}(D, f : \mathcal{X}^n \to \mathbb{R}, \varepsilon)$

1. Compute $\boldsymbol{\Delta} = \boldsymbol{\Delta}_1(f)$, the sensitivity of function f
2. draw $Y \sim \text{Lap}(\boldsymbol{\Delta}/\varepsilon)$, the added noise
3. Output $f(D) + Y$, the noisy answer to query f over D

Theorem (DP guarantees for Laplace mechanism)
The Laplace mechanism $\mathcal{A}_{\text{Lap}}(D, f, \varepsilon)$ *satisfies* ε*-differential privacy*

LIMITATIONS OF OUTPUT PERTURBATION

- So far we have seen the Laplace mechanism, which is based on output perturbation

$$\mathcal{A}(D) = f(D) + Y$$

- Can you think of some intrinsic limitations?

LIMITATIONS OF OUTPUT PERTURBATION

- So far we have seen the Laplace mechanism, which is based on output perturbation

$$\mathcal{A}(D) = f(D) + Y$$

- Can you think of some intrinsic limitations?

First limitation

- It only works for numeric queries

For instance, what if the output is a label, a set or even a graph?

LIMITATIONS OF OUTPUT PERTURBATION

- So far we have seen the Laplace mechanism, which is based on output perturbation

$$\mathcal{A}(D) = f(D) + Y$$

- Can you think of some intrinsic limitations?

First limitation

- It only works for numeric queries

For instance, what if the output is a label, a set or even a graph?

Second limitation

- It is relevant only if the utility function is sufficiently regular

When perturbation leads to invalid outputs, e.g. how to ensure integrality or non-negativity of the output?

EXAMPLE QUERIES NOT WELL SUITED TO OUTPUT PERTURBATION

- Non-numeric queries
  - What is the most popular website among Firefox users?
  - What is the best set of hyperparameters to train my classifier on the dataset?

EXAMPLE QUERIES NOT WELL SUITED TO OUTPUT PERTURBATION

- Non-numeric queries
  - What is the most popular website among Firefox users?
  - What is the best set of hyperparameters to train my classifier on the dataset?
- Numeric queries for which two “similar” outputs can have very different utility
  - Which date works better for a set of people to meet?
  - Which price would make the most profit from a set of buyers?

Buyer	Offer
Alice	3.0€
Bob	4.0€

EXAMPLE QUERIES NOT WELL SUITED TO OUTPUT PERTURBATION

- Non-numeric queries
  - What is the most popular website among Firefox users?
  - What is the best set of hyperparameters to train my classifier on the dataset?
- Numeric queries for which two “similar” outputs can have very different utility
  - Which date works better for a set of people to meet?
  - Which price would make the most profit from a set of buyers?

Buyer	Offer
Alice	3.0€
Bob	4.0€

- Profit if we set price to 3€: 3€

EXAMPLE QUERIES NOT WELL SUITED TO OUTPUT PERTURBATION

- Non-numeric queries
  - What is the most popular website among Firefox users?
  - What is the best set of hyperparameters to train my classifier on the dataset?
- Numeric queries for which two “similar” outputs can have very different utility
  - Which date works better for a set of people to meet?
  - Which price would make the most profit from a set of buyers?

Buyer	Offer
Alice	3.0€
Bob	4.0€

- Profit if we set price to 3€: 3€
- Profit if we set price to 3.01€: 3.01€

EXAMPLE QUERIES NOT WELL SUITED TO OUTPUT PERTURBATION

- Non-numeric queries
    - What is the most popular website among Firefox users?
    - What is the best set of hyperparameters to train my classifier on the dataset?
- Numeric queries for which two “similar” outputs can have very different utility
    - Which date works better for a set of people to meet?
    - Which price would make the most profit from a set of buyers?

Buyer	Offer
Alice	3.0€
Bob	4.0€

- Profit if we set price to 3€: 3€
- Profit if we set price to 3.01€: 3.01€
- Profit if we set price to 4€: 4€

EXAMPLE QUERIES NOT WELL SUITED TO OUTPUT PERTURBATION

- Non-numeric queries
  - What is the most popular website among Firefox users?
  - What is the best set of hyperparameters to train my classifier on the dataset?
- Numeric queries for which two “similar” outputs can have very different utility
  - Which date works better for a set of people to meet?
  - Which price would make the most profit from a set of buyers?

Buyer	Offer
Alice	3.0€
Bob	4.0€

- Profit if we set price to 3€: 3€
- Profit if we set price to 3.01€: 3.01€
- Profit if we set price to 4€: 4€
- Profit if we set price to 4.01€: 0€

NON-NUMERIC QUERIES

- We will now consider queries $f : \mathcal{X}^n \to \mathcal{O}$ with an abstract output space $\mathcal{O}$
  - Example (dates): $\mathcal{O} = \{$'Monday', 'Tuesday', 'Wednesday', ...$\}$
  - Example (prices): $\mathcal{O} = \{3, 3.01, 4, 4.01, ...\}$
  - Example (hair color): $\mathcal{O} = \{$'dark', 'blond', 'brown', 'red'$\}$

NON-NUMERIC QUERIES

- We will now consider queries $f : \mathcal{X}^n \to \mathcal{O}$ with an abstract output space $\mathcal{O}$
  - Example (dates): $\mathcal{O} =$ {'Monday', 'Tuesday', 'Wednesday', ...}
  - Example (prices): $\mathcal{O} =$ {3, 3.01, 4, 4.01, ...}
  - Example (hair color): $\mathcal{O} =$ {'dark', 'blond', 'brown', 'red'}
- Associated to $\mathcal{O}$ we have a scoring function (or utility function)

$$s : \mathcal{X}^n \times \mathcal{O} \to \mathbb{R}$$

NON-NUMERIC QUERIES

- We will now consider queries $f : \mathcal{X}^n \to \mathcal{O}$ with an abstract output space $\mathcal{O}$
  - Example (dates): $\mathcal{O} = \{$'Monday', 'Tuesday', 'Wednesday', ...$\}$
  - Example (prices): $\mathcal{O} = \{3, 3.01, 4, 4.01, ...\}$
  - Example (hair color): $\mathcal{O} = \{$'dark', 'blond', 'brown', 'red'$\}$
- Associated to $\mathcal{O}$ we have a scoring function (or utility function)

$$s : \mathcal{X}^n \times \mathcal{O} \to \mathbb{R}$$

- For a dataset $D \in \mathcal{X}^n$ and an output $o \in \mathcal{O}$, $s(D, o)$ represents how good it is to return o when the query is $f(D)$

NON-NUMERIC QUERIES

- We will now consider queries $f : \mathcal{X}^n \rightarrow \mathcal{O}$ with an abstract output space $\mathcal{O}$
    - Example (dates): $\mathcal{O} = \{$'Monday', 'Tuesday', 'Wednesday', ...$\}$
    - Example (prices): $\mathcal{O} = \{3, 3.01, 4, 4.01, ...\}$
    - Example (hair color): $\mathcal{O} = \{$'dark', 'blond', 'brown', 'red'$\}$
- Associated to $\mathcal{O}$ we have a scoring function (or utility function)

$$s : \mathcal{X}^n \times \mathcal{O} \rightarrow \mathbb{R}$$

- For a dataset $D \in \mathcal{X}^n$ and an output $o \in \mathcal{O}$, $s(D, o)$ represents how good it is to return o when the query is $f(D)$
- The function s can be arbitrary: it should be designed according to the use-case

NON-NUMERIC QUERIES

- We will now consider queries $f : \mathcal{X}^n \rightarrow \mathcal{O}$ with an abstract output space $\mathcal{O}$
  - Example (dates): $\mathcal{O} =$ {'Monday', 'Tuesday', 'Wednesday', ...}
  - Example (prices): $\mathcal{O} = \{3, 3.01, 4, 4.01, ...\}$
  - Example (hair color): $\mathcal{O} =$ {'dark', 'blond', 'brown', 'red'}
- Associated to $\mathcal{O}$ we have a scoring function (or utility function)

$$s : \mathcal{X}^n \times \mathcal{O} \rightarrow \mathbb{R}$$

- For a dataset $D \in \mathcal{X}^n$ and an output $o \in \mathcal{O}$, $s(D, o)$ represents how good it is to return o when the query is $f(D)$
- The function s can be arbitrary: it should be designed according to the use-case
- Of course, $o = f(D)$ is usually assigned the maximum score

SENSITIVITY OF THE SCORING FUNCTION

Definition (Sensitivity of scoring function)

The sensitivity of $s : \mathcal{X}^n \times \mathcal{O} \to \mathbb{R}$ is

$$\boldsymbol{\Delta}(s) = \max_{o \in \mathcal{O}} \max_{D, D' : D \Delta D' \leq 1} |s(D, o) - s(D', o)|$$

SENSITIVITY OF THE SCORING FUNCTION

Definition (Sensitivity of scoring function)

The sensitivity of $s : \mathcal{X}^n \times \mathcal{O} \to \mathbb{R}$ is

$$\mathbf{\Delta}(s) = \max_{o \in \mathcal{O}} \max_{D, D' : D \Delta D' \leq 1} |s(D, o) - s(D', o)|$$

- Worst-case change of score of an output when adding or removing one record
- Note that sensitivity is only with respect to the dataset (scores can vary arbitrarily across outputs)

Algorithm: Exponential mechanism $\mathcal{A}_{\text{Exp}}(D, f : \mathcal{X}^n \to \mathcal{O}, s : \mathcal{X}^n \times \mathcal{O} \to \mathbb{R}, \varepsilon)$

1. Compute $\mathbf{\Delta} = \mathbf{\Delta}(s)$
2. Output $o \in \mathcal{O}$ with probability:

$$\Pr[o] = \frac{\exp\left(\frac{s(D,o)\cdot\varepsilon}{2\mathbf{\Delta}}\right)}{\sum_{o' \in \mathcal{O}} \exp\left(\frac{s(D,o')\cdot\varepsilon}{2\mathbf{\Delta}}\right)}$$

- Sample $o \in \mathcal{O}$ with probability proportional to its score (denominator: normalization)
- Make high quality outputs exponentially more likely, at a rate that depends on the sensitivity of the score and the privacy parameter

Theorem (DP guarantees for exponential mechanism)

Let $\varepsilon > 0$, $f : \mathcal{X}^n \to \mathcal{O}$ and $s : \mathcal{X}^n \times \mathcal{O} \to \mathbb{R}$. $\mathcal{A}_{\text{Exp}}(\cdot, f, s, \varepsilon)$ satisfies ε-DP.

THE EXPONENTIAL MECHANISM: UTILITY GUARANTEES

- Given a dataset D, let $s^*(D) = \max_{o \in \mathcal{O}} s(D, o)$, the best score for D
- One shows that it is unlikely that $\mathcal{A}_{\text{Exp}}$ returns a "bad" output, measured w.r.t. $s^*(D)$

THE EXPONENTIAL MECHANISM: UTILITY GUARANTEES

- Given a dataset D, let $s^*(D) = \max_{o \in \mathcal{O}} s(D, o)$, the best score for D
- One shows that it is unlikely that $\mathcal{A}_{\text{Exp}}$ returns a “bad” output, measured w.r.t. $s^*(D)$

Theorem (Utility guarantees for the Exponential mechanism)

Let $\varepsilon > 0$, $f : \mathcal{X}^n \to \mathcal{O}$ *and* $s : \mathcal{X}^n \times \mathcal{O} \to \mathbb{R}$. *Given a dataset* $D \in \mathcal{X}^n$, *let* $\mathcal{O} = \{o \in \mathcal{O} : s(D, o) = s^*(D)\}$. *Then:*

$$\Pr\left[s^*(D) - s(\mathcal{A}_{\text{Exp}}(D, f, s, \varepsilon) \leq \frac{2\Delta(s)}{\varepsilon} \ln\left(\frac{|\mathcal{O}|}{\beta|\mathcal{O}^*|}\right)\right] \geq 1 - \beta$$

THE EXPONENTIAL MECHANISM: UTILITY GUARANTEES

- Given a dataset D, let $s^*(D) = \max_{o \in \mathcal{O}} s(D, o)$, the best score for D
- One shows that it is unlikely that $\mathcal{A}_{\text{Exp}}$ returns a “bad” output, measured w.r.t. $s^*(D)$

Theorem (Utility guarantees for the Exponential mechanism)

Let $\varepsilon > 0$, $f : \mathcal{X}^n \to \mathcal{O}$ *and* $s : \mathcal{X}^n \times \mathcal{O} \to \mathbb{R}$. *Given a dataset* $D \in \mathcal{X}^n$, *let* $\mathcal{O} = \{o \in \mathcal{O} : s(D, o) = s^*(D)\}$. *Then:*

$$\Pr\left[s^*(D) - s(\mathcal{A}_{\text{Exp}}(D, f, s, \varepsilon) \leq \frac{2\Delta(s)}{\varepsilon} \ln\left(\frac{|\mathcal{O}|}{\beta|\mathcal{O}^*|}\right)\right] \geq 1 - \beta$$

- It is highly unlikely that we get utility score smaller than $s^*(D)$ by more than an additive factor of $O\left(\frac{\Delta(s)}{\varepsilon} \ln(|\mathcal{O}|)\right)$
- Guarantees are better if several outputs have maximal score (i.e., $|\mathcal{O}^*| \geq 1$)

THE EXPONENTIAL MECHANISM: UTILITY GUARANTEES

- Let $\mathcal{O} = \{\text{'dark', 'blond', 'brown', 'red'}\}$ and consider the query "What is the most common hair color?" with counts as scores

THE EXPONENTIAL MECHANISM: UTILITY GUARANTEES

- Let $\mathcal{O} = \{$'dark', 'blond', 'brown', 'red'$\}$ and consider the query "What is the most common hair color?" with counts as scores
- Suppose that the most common color is 'dark' (with count 500) and the second most common is 'brown' (with count 399)

THE EXPONENTIAL MECHANISM: UTILITY GUARANTEES

- Let $\mathcal{O} = \{$'dark', 'blond', 'brown', 'red'$\}$ and consider the query "What is the most common hair color?" with counts as scores
- Suppose that the most common color is 'dark' (with count 500) and the second most common is 'brown' (with count 399)

For $\varepsilon = 0.1$, what is the probability that $\mathcal{A}_{\text{Exp}}$ returns 'dark'?

THE EXPONENTIAL MECHANISM: UTILITY GUARANTEES

- Let $\mathcal{O} = \{$'dark', 'blond', 'brown', 'red'$\}$ and consider the query "What is the most common hair color?" with counts as scores
- Suppose that the most common color is 'dark' (with count 500) and the second most common is 'brown' (with count 399)

For $\varepsilon = 0.1$, what is the probability that $\mathcal{A}_{\text{Exp}}$ returns 'dark'?

- Note that $\mathbf{\Delta}(s) = 1$, $|\mathcal{O}| = 4$ and $|\mathcal{O}^*| = 1$

THE EXPONENTIAL MECHANISM: UTILITY GUARANTEES

- Let $\mathcal{O} = \{$'dark', 'blond', 'brown', 'red'$\}$ and consider the query "What is the most common hair color?" with counts as scores
- Suppose that the most common color is 'dark' (with count 500) and the second most common is 'brown' (with count 399)

For $\varepsilon = 0.1$, what is the probability that $\mathcal{A}_{\text{Exp}}$ returns 'dark'?

- Note that $\mathbf{\Delta}(s) = 1$, $|\mathcal{O}| = 4$ and $|\mathcal{O}^*| = 1$
- Applying the theorem, we know that the probability of returning an output whose score is larger than $400 = 500 - 20\ln(4/\beta)$ is at least $1 - \beta$

THE EXPONENTIAL MECHANISM: UTILITY GUARANTEES

- Let $\mathcal{O} = \{$'dark', 'blond', 'brown', 'red'$\}$ and consider the query "What is the most common hair color?" with counts as scores
- Suppose that the most common color is 'dark' (with count 500) and the second most common is 'brown' (with count 399)

For $\varepsilon = 0.1$, what is the probability that $\mathcal{A}_{\text{Exp}}$ returns 'dark'?

- Note that $\boldsymbol{\Delta}(s) = 1$, $|\mathcal{O}| = 4$ and $|\mathcal{O}^*| = 1$
- Applying the theorem, we know that the probability of returning an output whose score is larger than $400 = 500 - 20\ln(4/\beta)$ is at least $1 - \beta$
- This gives $\beta = 4e^{-5}$, hence the probability to get the correct answer is at least $1 - \beta = 0.973$

REPORT NOISY MAX

Implements the Exp mechanism in terms of the Lap mechanism for finite output $\mathcal{O}$

Algorithm: Report Noisy Max (RNM)

1. For each $o \in \mathcal{O}$, calculate a noisy score $s(D, o) + \text{Lap}(\frac{\mathbf{\Delta}(s)}{\varepsilon})$
2. Output the element o with the maximum noisy score

REPORT NOISY MAX

Implements the Exp mechanism in terms of the Lap mechanism for finite output $\mathcal{O}$

Algorithm: Report Noisy Max (RNM)

1. For each $o \in \mathcal{O}$, calculate a noisy score $s(D, o) + \text{Lap}(\frac{\boldsymbol{\Delta}(s)}{\varepsilon})$
2. Output the element o with the maximum noisy score

Pessimistic privacy/utility analysis:

- With sensitivity $\boldsymbol{\Delta}(s)$, each “query” in step 1 is ε-DP

REPORT NOISY MAX

Implements the Exp mechanism in terms of the Lap mechanism for finite output $\mathcal{O}$

Algorithm: Report Noisy Max (RNM)

1. For each $o \in \mathcal{O}$, calculate a noisy score $s(D, o) + \text{Lap}(\frac{\mathbf{\Delta}(s)}{\varepsilon})$
2. Output the element o with the maximum noisy score

Pessimistic privacy/utility analysis:

- With sensitivity $\mathbf{\Delta}(s)$, each "query" in step 1 is ε-DP
- If $|\mathcal{O}| = n$, then the overall RNM is $n\varepsilon$-DP by sequential composition

REPORT NOISY MAX

Implements the Exp mechanism in terms of the Lap mechanism for finite output $\mathcal{O}$

Algorithm: Report Noisy Max (RNM)

1. For each $o \in \mathcal{O}$, calculate a noisy score $s(D, o) + \text{Lap}(\frac{\boldsymbol{\Delta}(s)}{\varepsilon})$
2. Output the element o with the maximum noisy score

Pessimistic privacy/utility analysis:

- With sensitivity $\boldsymbol{\Delta}(s)$, each "query" in step 1 is ε-DP
- If $|\mathcal{O}| = n$, then the overall RNM is $n\varepsilon$-DP by sequential composition
- However, exponential mechanism would cost ε only!
  - it releases less information, only the output rather than the n noisy scores

Privacy Guarantees of RNM

Report Noisy Max satisfies ε-DP, no matter how large (but finite) is $\mathcal{O}$, since it releases only the identity of the element with the largest noisy score [McSherry and Talwar, 2007]

THE EXPONENTIAL MECHANISM: PRACTICAL CONSIDERATIONS

- The exponential mechanism is the natural building block for answering queries with arbitrary utilities and arbitrary non-numeric range
- As we have seen, it is often quite easy to analyze
- The set $\mathcal{O}$ of possible outputs should not be specific to the particular dataset!
    - Otherwise one violates DP
    - Example of violation: possible prices for items based on actual bids
- The exponential mechanism can define a complex distribution over an arbitrary large domain, so it is not always possible to implement it efficiently

A LOT MORE TO SAY

- Approximate DP
- Gaussian mechanism
- Advanced composition
- Variants
  - Rényi DP
  - zero-Concentrated DP
- Local DP
- Sparse Vector Technique
- DP for ML: Private Empirical Risk Minimization, ...
- *etc.*

References

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