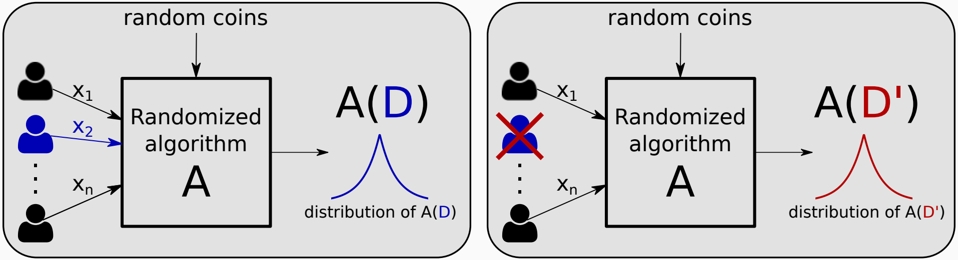
PRIVACY PRESERVING MACHINE LEARNING

LECTURE 3: THE EXPONENTIAL MECHANISM & ADVANCED COMPOSITION

Aurélien Bellet (Inria)

Master 2 Data Science, University of Lille

(Figure inspired from R. Bassily)



Definition (Differential privacy [Dwork et al., 2006])

Let $\varepsilon > 0$ and $\delta \in [0, 1)$. A randomized algorithm $\mathcal{A} : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{O}$ is (ε, δ)-differentially private (DP) if for all datasets $D, D' \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|D - D'\|_1 \leq 1$ and for all $\mathcal{S} \subseteq \mathcal{O}$:

$$\Pr[\mathcal{A}(D) \in \mathcal{S}] \leq e^{\varepsilon} \Pr[\mathcal{A}(D') \in \mathcal{S}] + \delta, \tag{1}$$

where the probability space is over the coin flips of $\mathcal{A}$.

Definition (Global ℓ_1 sensitivity)

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

$$\Delta_1(f) = \max_{D,D':\|D-D'\|_1 \leq 1} \|f(D) - f(D')\|_1$$

Definition (Global ℓ_2 sensitivity)

The global ℓ_2 sensitivity of a query (function) $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^K$ is

$$\Delta_2(f) = \max_{D,D':\|D-D'\|_1 \leq 1} \|f(D) - f(D')\|_2$$

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

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

1. Compute $\Delta = \Delta_1(f)$
2. For $k = 1, \ldots, K$: draw $Y_k \sim \text{Lap}(\Delta/\varepsilon)$ independently for each k
3. Output $f(D) + Y$, where $Y = (Y_1, \ldots, Y_K) \in \mathbb{R}^K$

Theorem (DP guarantees for Laplace mechanism)

Let $\varepsilon > 0$ and $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^K$. The Laplace mechanism $\mathcal{A}_{Lap}(\cdot, f, \varepsilon)$ satisfies ε-DP.

Algorithm: Gaussian mechanism $\mathcal{A}_{\text{Gauss}}(D, f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^K, \varepsilon, \delta)$

1. Compute $\Delta = \Delta_2(f)$
2. For $k = 1, \ldots, K$: draw $Y_k \sim \mathcal{N}(0, \sigma^2)$ independently for each k, where $\sigma = \frac{\sqrt{2\ln(1.25/\delta)}\Delta}{\varepsilon}$
3. Output $f(D) + Y$, where $Y = (Y_1, \ldots, Y_K) \in \mathbb{R}^K$

Theorem (DP guarantees for Gaussian mechanism)

Let $\varepsilon, \delta > 0$ and $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^K$. The Gaussian mechanism $\mathcal{A}_{\text{Gauss}}(\cdot, f, \varepsilon, \delta)$ is (ε, δ)-DP.

TODAY'S LECTURE

1. The exponential mechanism
2. Advanced composition results

The exponential mechanism

LIMITATIONS OF OUTPUT PERTURBATION

- So far we have seen the Laplace and Gaussian mechanisms, which are based on output perturbation: $\mathcal{A}(D) = f(D) + Y$
- Can you think of some intrinsic limitations?
- First limitation: they only work for numeric queries
- Second limitation: they are useful only if the utility function is sufficiently regular

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€
Bob	4€

- 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 $f : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{O}$ with an abstract output space $\mathcal{O}$
    - Example (websites): $\mathcal{O} = \{\text{'Google'}, \text{'Qwant'}, \text{'GitHub'}, \text{'La Quadrature du Net'}, \dots\}$
    - Example (prices): $\mathcal{O} = \{3, 3.01, 4, 4.01, \dots\}$
    - Example (hair color): $\mathcal{O} = \{\text{'dark'}, \text{'blond'}, \text{'brown'}, \text{'red'}\}$
- Associated to $\mathcal{O}$ we have a score function (or utility function)

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

- For a dataset $D \in \mathbb{N}^{|\mathcal{X}|}$ 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

Definition (Sensitivity of score function)

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

$$\Delta(s) = \max_{o \in \mathcal{O}} \max_{D, D' : \|D - D'\|_1 \leq 1} |s(D, o) - s(D', o)|$$

- 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 $\mathcal{A}_{\text{Exp}}(D, f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{O}, s : \mathbb{N}^{|\mathcal{X}|} \times \mathcal{O} \rightarrow \mathbb{R}, \varepsilon)$

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

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

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

Proof.

- For clarity, assume $\mathcal{O}$ is finite and let D, D' such that $\|D - D'\|_1 \leq 1$. For any $o \in \mathcal{O}$:

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

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

$$\leq \exp\left(\frac{\varepsilon}{2}\right) \cdot \exp\left(\frac{\varepsilon}{2}\right) \cdot \frac{\sum_{o'\in\mathcal{O}} \exp\left(\frac{s(D,o')\cdot\varepsilon}{2\Delta(s)}\right)}{\sum_{o'\in\mathcal{O}} \exp\left(\frac{s(D,o')\cdot\varepsilon}{2\Delta(s)}\right)} = e^{\varepsilon}$$

$\square$

- Fixing a dataset D, let $s^*(D) = \max_{o\in} s(D, o)$
- We show that it is unlikely that $\mathcal{A}_{\text{Exp}}$ returns a "bad" output, measured w.r.t. $s^*(D)$

Theorem (Utility guarantees for exponential mechanism)

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

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

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

Proof.

- We want to show that $\Pr[s(\mathcal{A}_{\text{Exp}}(D, f, s, \varepsilon)) \leq c] \leq \beta$ for $c = s^*(D) - \frac{2\Delta(s)}{\varepsilon} \ln\left(\frac{|\mathcal{O}|}{\beta|\mathcal{O}^*|}\right)$
- Think about "bad"outputs $o \in \mathcal{O}$ with $s(D, o) \leq c$
- Each such o has un-normalized probability mass at most $\exp(\varepsilon c/2\Delta(s))$, hence the entire set has total un-normalized probability mass at most $|\mathcal{O}| \exp(\varepsilon c/2\Delta(s))$
- In contrast, there is at least $|\mathcal{O}^*| \geq 1$ outputs o with $s(D, o) = s^*(D)$, therefore:

$$\begin{aligned}
\Pr[s(\mathcal{A}_{\text{Exp}}(D, f, s, \varepsilon)) \leq c] &\leq \frac{|\mathcal{O}| \exp(\varepsilon c/2\Delta(s))}{|\mathcal{O}^*| \exp(\varepsilon s^*(D)/2\Delta(s))} \\
&= \frac{|\mathcal{O}|}{|\mathcal{O}^*|} \exp\left(\frac{\varepsilon(c - s^*(D))}{2\Delta(s)}\right) \\
&= \beta
\end{aligned}$$

□

THE EXPONENTIAL MECHANISM: UTILITY GUARANTEES

- Let $\mathcal{O} = \{\text{'dark'}, \text{'blond'}, \text{'brown'}, \text{'red'}\}$ 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 $\Delta(s) = 1$, $|\mathcal{O}| = 4$ 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 $1 - \beta = 0.973$

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 we violate 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

Advanced composition results

Theorem (Simple composition)

Let $\mathcal{A}_1, \ldots, \mathcal{A}_K$ be K independently chosen algorithms where $\mathcal{A}_k$ satisfies $(\varepsilon_k, \delta_k)$-DP. For any dataset D, let $\mathcal{A}$ be such that

$$\mathcal{A}(D) = (\mathcal{A}_1(D), \ldots, \mathcal{A}_k(D)).$$

Then $\mathcal{A}$ is (ε, δ)-DP with $\varepsilon = \sum_{k=1}^{K} \varepsilon_k$ and $\delta = \sum_{k=1}^{K} \delta_k$.

- But data science is inherently an adaptive process: we would like to choose the next analysis to do based on previous results!

- Consider the following algorithm $\mathcal{A}_{\text{adap}}$ which takes as input a dataset D and runs K adaptively chosen DP mechanisms $\mathcal{A}_1, \ldots, \mathcal{A}_k$ on D

Algorithm $\mathcal{A}_{\text{adap}}(D)$

- Set initial state to s_0 (independent of D)
- For $k \in \{1, \ldots, K\}$:
  - $\mathcal{A}_k \leftarrow \text{Pick\_Alg}(s_0, \ldots, s_{k-1})$ // choose $\mathcal{A}_k$ based on previous outputs
  - $s_k \leftarrow \mathcal{A}_k(D)$
- Return $(s_1, \ldots, s_K)$

Theorem (Simple *adaptive* composition)

If at each round $k \in \{1, \ldots, K\}$, the selected algorithm $\mathcal{A}_k$ is guaranteed to satisfy $(\varepsilon_k, \delta_k)$-DP, then $\mathcal{A}_{adap}$ is (ε, δ)-DP with $\varepsilon = \sum_{k=1}^{K} \varepsilon_k$ and $\delta = \sum_{k=1}^{K} \delta_k$.

Proof.

- Let $D, D' \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|D - D'\|_1 \le 1$
- Let $S = (S_1, \ldots, S_K)$ (resp. S') be a random variable that denotes the vector of outputs of the K rounds when the input dataset is D (resp. D')
- Fix an output $s = (s_1, \ldots, s_K)$. Given $s_0, \ldots, s_{k-1}$, the algorithm $\mathcal{A}_k$ is determined by the (possibly randomized) algorithm Pick_Alg. Fix any internal randomness in Pick_Alg (i.e., we implicitly condition on fixed random coins of Pick_Alg)
- **Goal:** show that

$$\Pr[S = s] \le e^{\sum_{k=1}^{K} \varepsilon_k} \Pr[S' = s] + \sum_{k=1}^{K} \delta_k$$

$\square$

Proof.

- By the chain rule, we have

$$
\begin{aligned}
\Pr[S = s] &= \Pr[S = (s_1, \dots, s_K)] \\
&= \Pr[S_1 = s_1] \prod_{k=2}^{K} \Pr[S_k = s_k \mid S_1 = s_1, \dots, S_{k-1} = s_{k-1}] \\
&= \Pr[S_1 = s_1 | \mathcal{A}_1] \prod_{k=2}^{K} \Pr[S_k = s_k \mid S_1 = s_1, \dots, S_{k-1} = s_{k-1}, \mathcal{A}_k]
\end{aligned}
$$

- Since $S_k = \mathcal{A}_k(D)$, and S_k is independent of $S_1, \dots, S_{k-1}$ given $\mathcal{A}_k$, we have

$$
\Pr[S = s] = \prod_{k=1}^{K} \Pr[\mathcal{A}_k(D) = s_k | \mathcal{A}_k]
$$

$\square$

Proof.

- Consider the k-th term $\Pr[\mathcal{A}_k(D) = s_k | \mathcal{A}_k]$. Since $\mathcal{A}_k$ is $(\varepsilon_k, \delta_k)$-DP, we have

$$\begin{aligned}
\Pr[\mathcal{A}_k(D) = s_k | \mathcal{A}_k] &\leq e^{\varepsilon_k} \Pr[\mathcal{A}_k(D') = s_k | \mathcal{A}_k] + \delta_k \\
&\leq \min\left(e^{\varepsilon_k} \Pr[\mathcal{A}_k(D') = s_k | \mathcal{A}_k] + \delta_k, 1\right) \\
&\leq \min\left(e^{\varepsilon_k} \Pr[\mathcal{A}_k(D') = s_k | \mathcal{A}_k], 1\right) + \delta_k
\end{aligned}$$

- We can thus write:

$$\begin{aligned}
\prod_{k=1}^{K} \Pr[\mathcal{A}_k(D) = s_k | \mathcal{A}_k] &\leq \left(\min\left(e^{\varepsilon_1} \Pr[\mathcal{A}_1(D') = s_1 | \mathcal{A}_1], 1\right) + \delta_1\right) \prod_{k=2}^{K} \Pr[\mathcal{A}_k(D) = s_k | \mathcal{A}_k] \\
&\leq \min\left(e^{\varepsilon_1} \Pr[\mathcal{A}_1(D') = s_1 | \mathcal{A}_1], 1\right) \prod_{k=2}^{K} \Pr[\mathcal{A}_k(D) = s_k | \mathcal{A}_k] + \delta_1
\end{aligned}$$

□

Proof.

- Applying this recursively and using the conditional independence property used earlier on $\Pr[S' = s]$, we get

$$
\begin{aligned}
\Pr[S = s] &= \prod_{k=1}^{K} \Pr[\mathcal{A}_k(D) = s_k | \mathcal{A}_k] \\
&\le \prod_{k=1}^{K} \left(\min \left(e^{\varepsilon_k} \Pr[\mathcal{A}_k(D') = s_k | \mathcal{A}_k], 1 \right) + \sum_{k=1}^{K} \delta_k \right. \\
&\le e^{\sum_{k=1}^{K} \varepsilon_k} \prod_{k=1}^{K} \Pr[\mathcal{A}_k(D') = s_k | \mathcal{A}_k] + \sum_{k=1}^{K} \delta_k \\
&= e^{\sum_{k=1}^{K} \varepsilon_k} \Pr[S' = s] + \sum_{k=1}^{K} \delta_k
\end{aligned}
$$

□

- We can also prove another adaptive composition result known as advanced composition (see [Dwork and Roth, 2014] for the proof, which is more involved)

Theorem (Advanced composition)

Let $\epsilon, \delta, \delta' > 0$. If at each round $k \in \{1, \dots, K\}$, the selected algorithm $\mathcal{A}_k$ is guaranteed to satisfy (ε, δ)-DP, then $\mathcal{A}_{adap}$ is $(\varepsilon', K\delta + \delta')$-DP with

$$\varepsilon' = \sqrt{2K \ln(1/\delta')}\varepsilon + K\varepsilon(e^{\varepsilon} - 1)$$

- For small enough ϵ, the dominant term is $\sqrt{2K \ln(1/\delta')}\varepsilon$, which is much better than $K\varepsilon$ (simple composition) for large K!
- The result holds for $\delta = 0$ (composition of pure DP mechanisms) but requires $\delta' > 0$
- The two composition results do not conflict: they hold simultaneously

Corollary (see [Dwork and Roth, 2014])

Given target privacy parameters $0 < \varepsilon' < 1$ and $\delta' > 0$, to ensure $(\varepsilon', K\delta + \delta')$-DP for the composition of K mechanisms, it suffices that each mechanism is (ε, δ)-DP with

$$\varepsilon = \frac{\varepsilon'}{2\sqrt{2K\ln(1/\delta')}}$$

- We can fix the final privacy guarantee and use advanced composition to get much better utility by perturbing less each query (assuming we know K in advance)
- This corollary is convenient, but using the theorem directly yields tighter ε, which matters in practice!
- See [Kairouz et al., 2015] for slightly tighter (optimal) composition results that also hold when A_k is $(\varepsilon_k, \delta_k)$-DP

- These advanced composition results are not quite tight: they give somewhat loose upper bounds on the privacy cost
- Some variants of (ε, δ)-DP, such Rényi DP [Mironov, 2017] and zero-concentrated DP (zCDP) [Bun and Steinke, 2016], can enable tighter bounds
- In particular, they provide tighter composition results for the Gaussian mechanism
- Converting the privacy guarantees back to (ε, δ)-DP, this shaves off a logarithmic factor in δ and gives better constants

[Bun and Steinke, 2016] Bun, M. and Steinke, T. (2016).
Concentrated differential privacy: simplifications, extensions, and lower bounds.
In *TCC*.

[Dwork et al., 2006] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006).
Calibrating noise to sensitivity in private data analysis.
In *Theory of Cryptography (TCC)*.

[Dwork and Roth, 2014] Dwork, C. and Roth, A. (2014).
The Algorithmic Foundations of Differential Privacy.
Foundations and Trends in Theoretical Computer Science, 9(3–4):211–407.

[Kairouz et al., 2015] Kairouz, P., Oh, S., and Viswanath, P. (2015).
The Composition Theorem for Differential Privacy.
In *ICML*.

[Mironov, 2017] Mironov, I. (2017).
Renyi differential privacy.
In *CSF*.