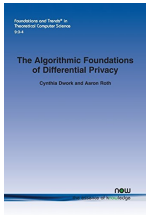
Differential Privacy for Machine Learning

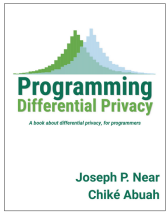
Master IASD, Université PSL

February 2024

Recommended Readings

Tutorials





- The Algorithmic Foundations of Differential Privacy – C. Dwork & A. Roth (2014)
- Programming Differential Privacy – J. P. Near & C. Abuah (2021)

Other References Used for this Course

- Simple Demographics Often Identify People Uniquely – L. Sweeney (2000)
- Near Instance-Optimality in Differential Privacy – H. Asi & J. C. Duchi (2020)
- Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising – B. Balle & Y-X. Wang (2018)
- Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences – B. Balle *et al.* (2018)
- Renyi Differential Privacy – I. Mironov (2017)
- Gaussian Differential Privacy – J. Dong *et al.* (2019)
- SoK: Differential Privacies – D. Desfontaines & B. Pejó (2020)
- Deep Learning with Differential Privacy – M. Abadi *et al.* (2016)

Table of Contents

Why Is Personal Data Protection in Machine Learning a Timely Topic?

From www.webfx.com

Let's look at the example of health data...

Because Personal Data Is Protected by the (EU) Legislation

(1) 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

(5) 'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;

(12) 'personal data breach' means a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed;


Text

ISSN 1977-0677

Official Journal **L 119**

of the European Union

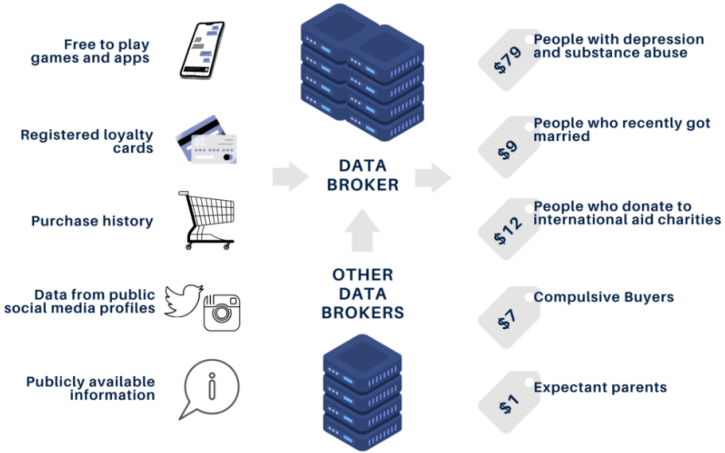
English edition | Legislation | Volume 59, 4 May 2016

Contents | *I Legislative acts* | page

REGULATIONS

* **Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)** ([1]) 1

Because Personal Data Collection Is Everywhere



And Not Always Where You Expect It

"Cash Investigation" : on vous résume la polémique autour d'Iqvia et des données personnelles de santé

"Cash Investigation" révèle ce jeudi que dans la moitié des pharmacies françaises, les informations sur les médicaments achetés par les consommateurs sont transmises à la société Iqvia, le plus gros revendeur de données médicales au monde.

franceinfo
France Télévisions

Publié le 20/05/2021 14:09 Mis à jour le 20/05/2021 15:04 Temps de lecture : 4 min.

L'émission "Cash Investigation" révèle que les données médicales sont récoltées dans de nombreuses pharmacies en France, sans que le consommateur en soit toujours informé (photo d'illustration). (ALICE S. / BSIP / AFP)

Because Data Has a Value

IQVIA Holdings (IQV) Stock

IQVIA offers advanced analytics, technology solutions, and clinical research services to the life sciences industry. The company helps to expedite the drug discovery and development process by leveraging artificial intelligence and machine learning.

During the Q3 conference, the company reassured investors that it is not seeing any indications of a slowdown in demand despite macro challenges. IQVIA has an extensive presence in over 100 countries and works with more than 10,000 customers, including top 25 large pharma companies.

IQVIA delivered market-beating Q3 results, with revenue rising 5% to $3.6 billion and adjusted EPS increasing 14.3% to $2.48. While demand remains strong, IQVIA has been facing operational challenges due to increased wages, high levels of attrition, disruptions caused by the Russia-Ukraine war, and lockdowns in China. It expects minor delays in the timing of deliveries due to macro challenges and staff shortages at certain sites. Overall, IQVIA projects full-year revenue growth in the range of 3.2% to 4.0%.

Is IQVIA Stock a Buy?

Barclays analyst Luke Sergott increased the price target for IQVIA stock to $235 from $215 following the Q3 results. Sergott noted that while the Q3 performance was "solid," guidance was "mixed." The analyst feels that IQVIA's earnings could be under pressure next year due to higher interest expenses. Nonetheless, Sergott expects the company to bring down its debt to address higher interest expenses.

Overall, IQVIA scores a Strong Buy consensus rating backed by 10 Buys and one Hold. The average IQV stock price target of $254.73 implies nearly 17% upside potential. Shares have fallen 22.7% year-to-date.

that machine learning can help leveraging

Because Large ML Models May Leak Information About the Training Data

[...] Reuters 5/11 Tim Hortons releases 'Java Daddy' tv ad where actor plays non-binary character called 'Java' and challenges Michael Jackson to an Apple Watch video choosing between the two A man was killed early Monday in a drive-by shooting on his front porch in the Englewood neighborhood on the South Side, police said. The shooting happened about 1:30 a.m. in the 7300 block of South Kedzie, Officer Ana Pacheco, a Chicago police spokeswoman, said in a news release. The victim, who had his back to the gunman when the shooting occurred, was struck in the chest by gunfire, [...]

https://github.com/ftramer/LM_Memorization [Extracting Training Data from Large Language Models, 2021], see more advanced attacks against LLMs in [Universal and Transferable Adversarial Attacks on Aligned Language Models, arXiv 2307.15043, 2023; Scalable Extraction of Training Data from (Production) Language Models, arXiv 2311.17035, 2023]

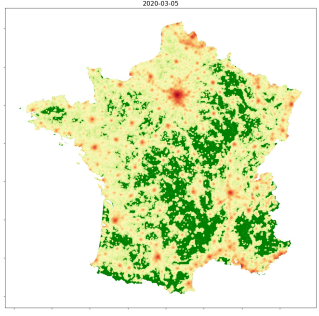
Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Model inversion attack from [Fredrikson et al. '2015]

- Pseudonymization
- Grouping
- Censoring
- Aggregation

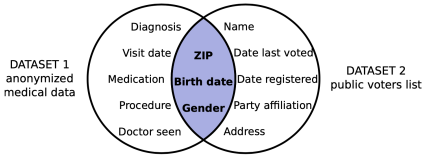
Name	Birth date	Zip code	Gender	Diagnosis	...
	1993-09-15	13741	M	Asthma	...
	1999-11-07	13440	F	Type-1 diabetes	...
	1945-07-31	02110	M	Cancer	...
	1950-03-13	02061	F	Cancer	...

	Quasi identifiers			Sensitive attribute	
Name	Age	Zip code	Gender	Diagnosis	...
	20-30	13***		Asthma	...
	20-30	13***		Type-1 diabetes	...
	70-80	02***		Cancer	...
	70-80	02***		Cancer	...



These Traditional Approaches Can Be Defeated

Linkage Attack [L. Sweeney, 2000] showed that gender, date of birth, and zip code are sufficient to uniquely identify the vast majority of Americans.
$\Longrightarrow$ By linking these attributes in a supposedly anonymized healthcare database to public voter records, she was able to identify the individual health record of the Governor of Massachusetts.



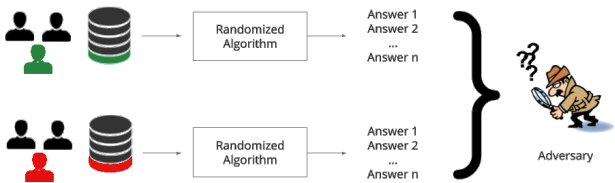
Differencing Attack Imposing queries on many lines is not the solution

- How many people in the database have the sickle cell trait (SCT)?
- How many people, not named Z, in the database have the sickle cell trait?

Attacks Against ML Models Aggregation does not prevent reidentification when # model parameters $\approx$ # data "dimension".

$\Longrightarrow$ Differential Privacy

DP (informally): an attacker should learn (almost) nothing about individual data from the output of the analysis, even if he new the data of others.



Smoker Paradox

If an individual is openly "smoking" but wants privacy on his medical status,

- a *DP* medical study will make it impossible to determine his medical status, even to someone who would have all the remaining information;
- but the same study could reveal a risk associated with smoking (whether he participates or not to the study).

To date there's a number real-world deployments of differential privacy, the most important being:

- 2008 U.S. Census Bureau, for showing commuting patterns.
- 2014 Google's RAPPOR, for telemetry such as learning statistics about unwanted software hijacking users' settings.
- 2015 Google, for sharing historical traffic statistics.
- 2016 Apple announced its intention to use differential privacy in iOS 10 to improve its Intelligent personal assistant technology.
- 2017 Microsoft, for telemetry in Windows.
- 2019 Privitar Lens is an API using differential privacy.
- 2020 LinkedIn, for advertiser queries.
- 2021 The US Census Bureau uses differential privacy to release redistricting data from the 2020 Census.

An Historical Example

RANDOMIZED RESPONSE: A SURVEY TECHNIQUE
FOR ELIMINATING EVASIVE ANSWER BIAS

STANLEY L. WARNER
Claremont Graduate School

For various reasons individuals in a sample survey may prefer not to confide to the interviewer the correct answers to certain questions. In such cases the individuals may elect not to reply at all or to reply with incorrect answers. The resulting evasive answer bias is ordinarily difficult to assess. In this paper it is argued that such bias is potentially removable through allowing the interviewee to maintain privacy through the device of randomizing his response. A randomized response method for estimating a population proportion is presented as an example. Unbiased maximum likelihood estimates are obtained and their mean square errors are compared with the mean square errors of conventional estimates under various assumptions about the underlying population.

1. INTRODUCTION

FOR reasons of modesty, fear of being thought bigoted, or merely a reluctance to confide secrets to strangers, many individuals attempt to evade certain questions put to them by interviewers. In survey vernacular, these people become the "non-cooperative" group [5, pp. 235–72], either refusing outright to be surveyed, or consenting to be surveyed but purposely providing wrong answers to the questions. In the one case there is the problem of refusal bias [1, pp. 355–61], [2, pp. 33–6], [5, pp. 261–9]; in the other case there is the problem of response bias [3, p. 89], [4, pp. 280–325].

Journal of the American Statistical Association, Mar. 1965, Vol.60, No.309, pp. 63-69.

Proposes and analyzes a protocol for carrying out statistical surveys on sensitive issues.

Example, Student Attendance Survey: "do you skip lectures?"

Statistical Model Student's answer $X_i \overset{i.i.d.}{\sim} \text{Bernoulli}(\theta)$, where one wants to estimate θ.

But actual answer $\tilde{X}_i \in \{0, 1\}$, will be different from X_i (particularly when $X_i = 1$) and hence the empirical mean $\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i$ is under-estimating θ.

Randomized Response [Warner'65]

Flip a coin, then:

- $\rightarrow$ if tails, answer according to another coin flip
- $\rightarrow$ if heads, give the right answer

$$\mathbb{P}(\tilde{X}_i = 1 | X_i = x_i) = 1/4 + x_i/2$$

at least $1/4$th of the students will answer $\tilde{X}_i = 1$ just by chance (irrespectively of their attendance).

Privacy Guarantee

$$\frac{\mathbb{P}(\tilde{X}_i = 1 | X_i = 1)}{\mathbb{P}(\tilde{X}_i = 1 | X_i = 0)} = 3$$

which implies that the posterior odds ratio

$$\frac{\mathbb{P}(X_i = 1 | \tilde{X}_i = 1)}{\mathbb{P}(X_i = 0 | \tilde{X}_i = 1)} = \underbrace{\frac{\mathbb{P}(\tilde{X}_i = 1 | X_i = 1)}{\mathbb{P}(\tilde{X}_i = 1 | X_i = 0)}}_{3} \underbrace{\frac{\mathbb{P}(X_i = 1)}{\mathbb{P}(X_i = 0)}}_{\theta/(1-\theta)}$$

is only marginally more significant than the prior odds ratio.

Statistical Accuracy

$$\hat{\theta}_n = \frac{2}{n} \sum_{i=1}^{n} \tilde{X}_i - \frac{1}{2}$$

is and unbiased estimator of θ with variance that is, at least, 4 times larger than that of the truthful estimator $\frac{1}{n} \sum_{i=1}^{n} X_i$ (i.e., requires at least 4 times more data)

Table of Contents

Randomized Mechanisms

Consider

- $x = (x_1, \ldots, x_n)$ a dataset, containing n individual records,
- M a mechanism such that for all x, $M(x)$ is a $\mathcal{Y}$-valued random variable.

Equivalently, one may write $M(x)$ as a randomized function $\varphi(x, U)$, where

- $\varphi : \mathcal{X}^n \times [0,1] \to \mathcal{Y}$ is a deterministic function,
- and U is a uniformly distributed random seed.

The definition of DP relies on controlling the change in the distribution of the output $M(x)$ when only one individual record in x is modified.

Example: Noisy Function Evaluation

Let $f(x)$ denote the function of interest to the analyst.

A typical mechanism consists in

$$M(x) = f(x) + Z \qquad \text{where } \mathbb{E}(Z) = 0$$

Where the distribution of Z has to be chosen as a tradeoff between the two conflicting objectives

- (Privacy) the distribution of $M(x)$ doesn't vary much when a single record in x is modified;
- (Accuracy) $M(x)$ is not too different from $f(x)$.

Note that in this course we will also sometimes consider a slighlty different notion of "statistical" accuracy.

Neighboring Datasets

Datasets x and x' are said to be neighboring, denoted $d(x, x') = 1$, if they match for all records but one (i.e., there exists i such that $x_j = x'_j$ for $j \in \{1, \ldots, n\} \backslash \{i\}$).

ε–Differential Privacy (also called $(\varepsilon, 0)$ or "pure" DP)

A $\mathcal{Y}$-valued mechanism M is ε–DP if, for all neighboring x and x', and any set $\mathcal{S} \subset \mathcal{Y}$,

$$\mathbb{P}(M(x) \in \mathcal{S}) \leq e^{\varepsilon} \mathbb{P}(M(x') \in \mathcal{S})$$

where $\mathbb{P}$ refers to the randomization in M.

It will sometimes be helpful to think of M as the randomized function such that $M(x) = \varphi(x, U)$ and of $\mathbb{P}$ as $\mathbb{P}^U$.

In situations where it is meaningful to consider that X itself is random, we will use the notation $M(X) = \varphi(X, U)$ and $\mathbb{P}$ should be understood as $\mathbb{P}^{U,X}$.

In Picture

ε–DP is equivalent to

- if $\mathcal{Y}$ is discrete,

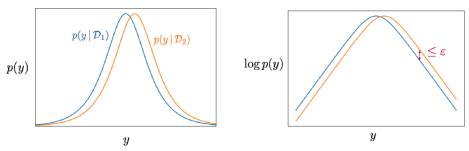
$$-\varepsilon \leq \log \frac{\mathbb{P}\left(M(x) = y\right)}{\mathbb{P}\left(M(x') = y\right)} \leq \varepsilon \quad \text{for all } y \in \mathcal{Y}$$

- if $M(x)$ has density $p(\cdot|x)$,

$$-\varepsilon \leq \log \frac{p(y|x)}{p(y|x')} \leq \varepsilon \quad \text{for all } y \in \mathcal{Y}$$

for all neighboring datasets x and x'.

Visually:



Notice that the tail behavior is important.

DP of the Randomized Response Mechanism

In the previous example of the randomized response survey

$$M(X) = (\tilde{X}_1, \ldots, \tilde{X}_n)$$

is $\log(3)$–DP.

- Note that RR outputs an entire (differentially private) version $M(X)$ of the original dataset X.
- This implies, for instance, that the estimator $\hat{\theta}_n = \frac{2}{n}\sum_{i=1}^{n} \tilde{X}_i - \frac{1}{2}$ is also $\log(3)$-DP (see below), but of course if one needs only and estimator of θ, there are, potentially better, differentially private mechanisms.
- DP is created "locally", at the scale of each individual record X_i, which can be of interest in distributed settings (more about "local differential privacy" to come later)

Variants

- In some references, x' is defined as $x' = (x_1, \ldots, x_{i-1}, x_{i+1}, x_n)$ (i.e. "remove a record" instead of substitute one).
- For structured data, substituting the relevant information may be involve more than just modifying a single record

Time stamp	User id	IP Address	Action
18:02:21	uid11	192.168.0.12	close
18:02:34	uid54	192.168.0.66	close
18:03:03	uid27	192.168.0.32	open
⋮	⋮	⋮	⋮
19:10:21	uid11	192.168.0.66	open
19:11:25	uid27	192.168.0.32	close

Protection Against Statistical Disclosure of Information

Imagine that one knows the nature of the mechanism M and all the data, but x_i, and consider testing for

H_0: $x_i \in \mathcal{B}$

H_1: $x_i \notin \mathcal{B}$

This can be done by designing a test statistic t, with threshold s such that

$$\mathbb{P}_{H_1}\left[t(M(x)) > s\right] \gg \mathbb{P}_{H_0}\left[t(M(x)) > s\right]$$

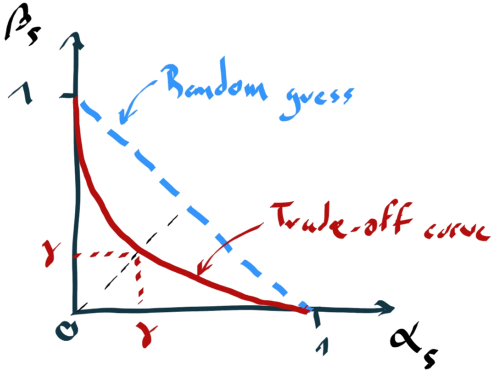
Define

False Positive Rate $\alpha_s = \mathbb{P}_{H_0}\left[t(M(x)) > s\right]$ (a.k.a. "probability of Type I error")

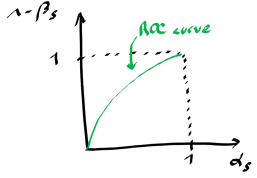
False Negative Rate $\beta_s = 1 - \mathbb{P}_{H_1}\left[t(M(x)) > s\right]$ (a.k.a. "probability of Type II error)

Equal Error Rate $\gamma = \alpha_{s_0} = \beta_{s_0}$, where s_0 is chosen such that equality occurs

$1 - \beta_s$ is also known as the "true positive rate" or "power" of the test.



Also frequently represented as



Property

If M is ε–DP, then, for any test statistic t,

$$\max(\alpha_s, \beta_s) \geq \gamma \geq \frac{1}{1+e^{\varepsilon}}$$

Parameterized Random Response

ν–Randomized Response

Flip a biased coin such that $\mathbb{P}(\text{heads}) = \nu$ and

$\rightarrow$ if tails, answer according to another (balanced) coin flip

$\rightarrow$ if heads, give the right answer

$$\mathbb{P}(\tilde{X}_i = 1 | X_i = x_i) = \frac{1-\nu}{2} + \nu x_i$$

- $M(X)$ is $\log \frac{1+\nu}{1-\nu} \leq 2\frac{\nu}{1-\nu}$ –DP
- The estimator $\hat{\theta}_n = \frac{1}{n\nu}\sum_{i=1}^n \tilde{X}_i - \frac{1-\nu}{2\nu}$ has (L^2) statistical accuracy

$$\mathbb{E}(\hat{\theta}_n - \theta)^2 = \frac{1}{n\nu^2} \underbrace{[(1-\nu)/2 + \nu\theta][(1+\nu)/2 - \nu\theta]}_{\theta(1-\theta) \leq \cdot \leq \frac{1}{4}}$$

- But in some context it may also be more appropriate to consider the (empirical) accuracy

$$\mathbb{E}\left[\left(\hat{\theta}_n - \frac{1}{n}\sum_{i=1}^n X_i\right)^2 \middle| X = x\right] = \frac{1}{n^2\nu^2}\sum_{i=1^n} \underbrace{[(1-\nu)/2 + \nu x_i][(1+\nu)/2 - \nu x_i]}_{0 \leq \cdot \leq \frac{1}{4}}$$

Post-processing

Mechanisms M_1 and M_2 are said to be independent when $M_1(\cdot) = \varphi_1(\cdot, U_1)$, $M_2(\cdot) = \varphi_2(\cdot, U_2)$ and U_1 and U_2 are independent random variables.

Post-Processing

If M_1 is ε–DP, then for any independent mechanism M_2,

$$M_2 \circ M_1 \text{ is also } \varepsilon\text{-DP}$$

The output of an ε–DP mechanism is protected and can be used freely without further degradation of the privacy guarantee.

Group Privacy

Group Privacy

If M is ε–DP and $d(x, x') = k$,

$$\mathbb{P}\left(M(x) \in \mathcal{S}\right) \leq e^{k\varepsilon}\, \mathbb{P}\left(M(x') \in \mathcal{S}\right)$$

The privacy guarantee extends to datasets that differ by k entries.

"Composition" (Multiple Results)

Composition

If M_1 and M_2 are independent mechanisms, respectively, ε_1–DP and ε_2–DP,

$$M : x \mapsto (M_1(x), M_2(x)) \text{ is } (\varepsilon_1 + \varepsilon_2)\text{–DP}$$

When revealing the output of several independent mechanisms applied to the same data, the privacy parameters add up.

This is somewhat pessimistic and can be improved in some cases, for instance in "Parallel composition": if $x = (x^1, x^2)$, revealing $\big(M_1(x^1), M_2(x^2)\big)$ is $\max(\varepsilon_1, \varepsilon_2)$–DP only.

Adaptive Composition

Sometimes, it is needed to consider mechanisms that also depend on previous results obtained from the data.

$\Rightarrow$ As an example, consider applying a recursive algorithm to x, where each iteration is randomized (e.g., in DP-SGD to be covered later).

Adaptive Composition

If $M_1(x) = \varphi_1(x, U_1)$ and $M_2(x, y) = \varphi_2(x, y, U_2)$, with U_1 and U_2 independent and

- $x \mapsto M_1(x)$ is ε_1–DP
- $x \mapsto M_2(x, y)$ is ε_2–DP for all $y \in \mathcal{Y}$

Then

$$M : x \mapsto (M_1(x), M_2(x, M_1(x))) \text{ is } (\varepsilon_1 + \varepsilon_2)\text{–DP}$$

Table of Contents

For discrete $\mathcal{Y}$, a mechanism $M(x)$ can be defined by its distribution, i.e., the probability that it assigns to each possible outcome $y \in \mathcal{Y}$ for any dataset x.

In this part of the course, we will discuss

- a generic construction for discrete-valued mechanisms (the exponential mechanism);
- bounds that quantify the privacy/accuracy tradeoff.

Introducing Example: Median of Binary Observations

Consider a dataset of $n = 2k + 1$ binary observations, $x = (x_1, \ldots, x_n)$, and

$$f(x) = \text{median}(x) = \mathbb{1}\Big\{ \sum_{i=1}^{n} x_i \geq n/2 \Big\}$$

A mechanism $M(x)$ for the median should (intuitively)

- depend only $s = \sum_{i=1}^{n} x_i$
- be defined by $p(s) = \mathbb{P}(M(x) = 1)$, for $s = 0, \ldots, n$
- be symmetric, i.e., $p(n - s) = 1 - p(s)$, with $p(s)$ increasing

ε–DP implies that

$$p(k+1) \leq e^{\varepsilon} p(k) = e^{\varepsilon}(1 - p(k+1)) \implies p(k+1) \leq \frac{1}{1 + e^{-\varepsilon}}$$

and more generally, that for all $s > n/2$, $p(s) \leq \dfrac{1}{1 + e^{-(2s-n)\varepsilon}}$

Mechanism for the Median

$$p(s) = \frac{1}{1 + e^{-(2s-n)\varepsilon/2}} \quad \text{is } \varepsilon\text{–DP}$$

- When $|s - n/2| \geq 3/\varepsilon$, $M(x)$ outputs the sample median with probability $\geq 95\%$
- From a statistical viewpoint, we can assume (with probability at least 95%), that the median in the sample does correspond to the median in the population, only when $|s - n/2| \geq \underbrace{\sqrt{\log(2/0.05)/2}}_{\approx 1.36}\sqrt{n}$
- Thus, if $\varepsilon \geq \sqrt{5/n} \iff n \geq 5/\varepsilon^2$, the fact that $M(x)$ is ε–DP does not significantly degrade the statistical accuracy, when compared to s (which can reveal individual x_i)

Utility and Sensitivity

To construct a mechanism $M(x)$ for $f(x) : x \mapsto y \in \mathcal{Y}$, where $\mathcal{Y}$ is discrete, define the following

Utility

The utility $u(x, y)$ is a real-valued function that measures how much the outcome y matches $f(x)$ for the dataset x;
typically $u(x, y) \leq u(x, f(x))$.

Sensitivity

The sensitivity of u is defined as

$$\Delta_u = \max_{y \in \mathcal{Y}} \max_{x, x' : d(x, x') = 1} \left| u(x, y) - u(x', y) \right|$$

Exponential Mechanism

The utility function u can be used to define a generic $\mathcal{Y}$–valued mechanism.

Exponential Mechanism

The mechanism defined by

$$\mathbb{P}\big(M(x)=y\big)=\frac{\exp\left(\frac{u(x,y)\varepsilon}{2\Delta_u}\right)}{\sum_{y'\in\mathcal{Y}}\exp\left(\frac{u(x,y')\varepsilon}{2\Delta_u}\right)}$$ is ε–DP

In the median example: $u(x,y)=(2y-1)\left(s-\frac{n}{2}\right)=-u(x,1-y)$,

$$\mathbb{P}\big(M(x)=1\big)=\frac{\exp\left(\frac{(s-\frac{n}{2})\varepsilon}{2}\right)}{\exp\left(\frac{(s-\frac{n}{2})\varepsilon}{2}\right)+\exp\left(-\frac{(s-\frac{n}{2})\varepsilon}{2}\right)}=\frac{1}{1+\exp\left(-\left(s-\frac{n}{2}\right)\varepsilon\right)}$$

Proof For every $y \in \mathcal{Y}$ and x, x' such that $d(x, x') = 1$,

$$\frac{\mathbb{P}\big(M(x) = y\big)}{\mathbb{P}\big(M(x') = y\big)} = \frac{\exp\left(\frac{u(x,y)\varepsilon}{2\Delta_u}\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\frac{u(x,y')\varepsilon}{2\Delta_u}\right)} \Bigg/ \frac{\exp\left(\frac{u(x',y)\varepsilon}{2\Delta_u}\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\frac{u(x',y')\varepsilon}{2\Delta_u}\right)}$$

$$= \exp\left(\frac{\big(u(x,y) - u(x',y)\big)\varepsilon}{2\Delta_u}\right) \frac{\sum_{y' \in \mathcal{Y}} \exp\left(\frac{\big(u(x',y') - u(x,y')\big)\varepsilon}{2\Delta_u}\right) \exp\left(\frac{u(x,y')\varepsilon}{2\Delta_u}\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\frac{u(x,y')\varepsilon}{2\Delta_u}\right)}$$

$$\leq \exp\left(\frac{\varepsilon}{2}\right) \frac{\sum_{y' \in \mathcal{Y}} \exp\left(\frac{\varepsilon}{2}\right) \exp\left(\frac{u(x,y')\varepsilon}{2\Delta_u}\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\frac{u(x,y')\varepsilon}{2\Delta_u}\right)} = \exp(\varepsilon)$$

□

Bound on the Resulting Utility

Theorem

For every dataset x, the exponential mechanism satisfies:

$$\mathbb{P}\left(u\big(x, M(x)\big) \leq u\big(x, f(x)\big) - \frac{2\Delta_u}{\varepsilon} \log \frac{|\mathcal{Y}|}{\delta}\right) \leq \delta$$

Equivalently, for every $v > 0$

$$\mathbb{P}\left(u\left(x, M(x)\right) \leq u\left(x, f(x)\right) - v\right) \leq |\mathcal{Y}| e^{-\frac{v\varepsilon}{2\Delta_u}}$$

In the median example:

$$\mathbb{P}\big(M(x) \neq f(x)\big) = \mathbb{P}\Big(u\big(x, M(x)\big) \leq u\big(x, f(x)\big) - 2u\big(x, f(x)\big)\Big) \leq 2\exp\left(-\frac{u\big(x, f(x)\big)\varepsilon}{\Delta_u}\right) = 2e^{-|s-\frac{n}{2}|\varepsilon}$$

Proof For any $y \in \mathcal{Y}$ such that $u\big(x,y\big) \leq u\big(x,f(x)\big) - 2\Delta_u \varepsilon^{-1} \log\big(|\mathcal{Y}|/\delta\big)$,

$$\mathbb{P}\big(M(x)=y\big) \leq \frac{\exp\left(\frac{\left(u(x,f(x)) - \frac{2\Delta_u}{\varepsilon}\log\frac{|\mathcal{Y}|}{\delta}\right)\varepsilon}{2\Delta u}\right)}{\exp\left(\frac{u(x,f(x))\varepsilon}{2\Delta_u}\right)} = \frac{\delta}{|\mathcal{Y}|}$$

and there are at most $|\mathcal{Y}|$ of them

□

A discrete mechanism M is said to be unbiased if for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$\mathbb{P}\big(M(x) = y\big) \leq \mathbb{P}\big(M(x) = f(x)\big)$$

Inverse Sensitivity

The inverse sensitivity of a function $f : x \mapsto y \in \mathcal{Y}$ is defined as

$$\mathcal{D}_f(x, y) = \min \Big\{ k : \exists x' : d(x, x') = k \text{ and } f(x') = y \Big\}$$

Lower Bound

For every unbiased ε–DP mechanism M, $\mathbb{P}\big(M(x) = f(x)\big) \leq \dfrac{1}{\sum_{y \in \mathcal{Y}} e^{-2D_f(x,y)\varepsilon}}$

Proof

Let $y \in \mathcal{Y}$ and x' be such that $d(x,x') = D_f(x,y)$ and $f(x') = y$. DP implies

$$\frac{\mathbb{P}\big(M(x)=y\big)}{\mathbb{P}\big(M(x)=f(x)\big)} = \frac{\mathbb{P}\big(M(x)=y\big)}{\mathbb{P}\big(M(x')=y\big)} \frac{\mathbb{P}\big(M(x')=y=f(x')\big)}{\mathbb{P}\big(M(x')=f(x)\big)} \frac{\mathbb{P}\big(M(x')=f(x)\big)}{\mathbb{P}\big(M(x)=f(x)\big)} \geq e^{-\mathcal{D}_f(x,y)\varepsilon} \times 1 \times e^{-\mathcal{D}_f(x,y)\varepsilon}$$

and hence

$$1 = \sum_{y\in\mathcal{Y}} \frac{\mathbb{P}\big(M(x)=y\big)}{\mathbb{P}\big(M(x)=f(x)\big)} \mathbb{P}\big(M(x)=f(x)\big) \geq \sum_{y\in\mathcal{Y}} e^{-2\mathcal{D}_f(x,y)\varepsilon} \mathbb{P}\big(M(x)=f(x)\big)$$

□

In the median example, $\mathcal{D}_f(x, 1-f(x)) = \left|s - \frac{n}{2}\right| + \frac{1}{2}$ which yields $\mathbb{P}\big(M(x)=f(x)\big) \leq \frac{1}{1+e^{-\left(|2s-n|+1\right)\varepsilon}}$

Hence, the proposed exponential mechanism is almost optimal, as $\mathbb{P}\big(M(x)=f(x)\big) = \frac{1}{1+e^{-\frac{|2s-n|\varepsilon}{2}}}$

Inverse Sensitivity Mechanism

$-\mathcal{D}_f$ is a good candidate utility function for an exponential mechanism (with $\Delta_{\mathcal{D}_f} = 1$)

ε–DP Inverse Sensitivity Mechanism (ISM)

$$\mathbb{P}\big(M(x) = y\big) = \frac{e^{-\mathcal{D}_f(x,y)\varepsilon/2}}{\sum_{y' \in \mathcal{Y}} e^{-\mathcal{D}_f(x,y')\varepsilon/2}}$$

Near-Optimality of the Inverse Sensitivity Mechanism

$1/4$–Optimality of the ISM

The ISM M is "more accurate" than any $\varepsilon/4$-DP unbiased mechanism M' in the sense that

$$\mathbb{P}\big(M'(x) = f(x)\big) \leq \mathbb{P}\big(M(x) = f(x)\big)$$

Proof Since $\mathcal{D}_f(x, f(x)) = 0$, $\quad \mathbb{P}\big(M(x) = f(x)\big) = 1/\sum_{y \in \mathcal{Y}} e^{-\mathcal{D}_f(x,y)\varepsilon/2}$

Recall the lower bound: for every unbiased ε–DP mechanism M', $\quad \mathbb{P}\big(M'(x) = f(x)\big) \leq 1/\sum_{y \in \mathcal{Y}} e^{-2\mathcal{D}_f(x,y)\varepsilon}$

Possible Improvement in the Case where $|\mathcal{Y}| = 2$

Remark

If $\mathcal{Y} = \{0, 1\}$ the denominator 2 is not needed:

$$\mathbb{P}\big(M(x) = 1\big) = \frac{e^{-\mathcal{D}_f(x,1)\varepsilon}}{e^{-\mathcal{D}_f(x,1)\varepsilon} + e^{-\mathcal{D}_f(x,0)\varepsilon}}$$

is ε–DP

In the median example, $\mathcal{D}_f(x, 1 - f(x)) = \left|s - \frac{n}{2}\right| + \frac{1}{2}$, and the ISM $p(s) = \frac{1}{1 + e^{-(s-k)\varepsilon}}$, for $s > k$, is an ε–DP mechanism slightly better than the exponential mechanism discussed previously

It is $\varepsilon/2$-optimal by the lower-bound argument

Table of Contents

Continuous Exponential Mechanism

For a continuous $\mathcal{Y}$, given a bounded-sensitivity utility $u(x, y)$, sampling $M(x)$ with conditional density

$$p(y|x) = \frac{\exp\left(\frac{u(x,y)\varepsilon}{2\Delta_u}\right)}{\int_{\mathcal{Y}} \exp\left(\frac{u(x,y')\varepsilon}{2\Delta_u}\right) dy'}$$

also yields an ε–DP mechanism

But

- it is usually very hard to sample from
- in fact, the discrete case is already computationally challenging when the output space $\mathcal{Y}$ is "big"
- approximate simulations (e.g. MCMC methods) are not guaranteed to preserve DP in general

If $\mathcal{Y} = \mathbb{R}$ and $u(x, y) = -|y - f(x)|$ and $\Delta_u = \Delta_f = \sup_{x,x':d(x,x')=1} |f(x) - f(x')|$ the exponential mechanism

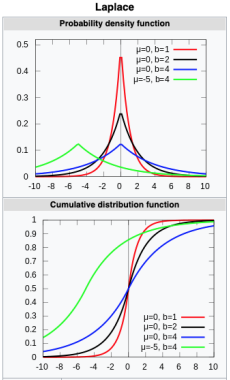
$$p(y|x) = \frac{\exp\left(\frac{u(x,y)\varepsilon}{\Delta_u}\right)}{\int_{\mathbb{R}} \exp\left(\frac{u(x,y')\varepsilon}{\Delta_u}\right) dy'} = \frac{\exp\left(\frac{-|y-f(x)|\varepsilon}{\Delta_f}\right)}{\frac{2\Delta_f}{\varepsilon}}$$

corresponds to adding Laplace (two-sided exponential) noise

$$M(x) = f(x) + Z$$

where $Z \sim \text{Laplace}(0, \Delta_f/\varepsilon) \sim \Delta_f/\varepsilon \, \text{Laplace}(0, 1)$ where $\text{Laplace}(0, 1)$ has density $q_{\text{Lap}(0,1)}(z) = \frac{1}{2}e^{|z|}$ on $\mathbb{R}$.

$\implies M(x)$ is ε–DP



Parameters	μ location (real) $b > 0$ scale (real)
Support	$\mathbb{R}$
PDF	$\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$
CDF	$\begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right) & \text{if } x \le \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \ge \mu \end{cases}$
Quantile	$\begin{cases} \mu + b\ln(2F) & \text{if } F \le \frac{1}{2} \\ \mu - b\ln(2-2F) & \text{if } F \ge \frac{1}{2} \end{cases}$
Mean	μ
Median	μ
Mode	μ
Variance	$2b^2$

Laplace mechanism

More generally, the L^1-sensitivity of a function $f : x \mapsto y \in \mathbb{R}^k$ is defined as

$$\Delta_f = \sup_{x,x' : d(x,x')=1} \|f(x) - f(x')\|_1$$

Laplace Mechanism

The Laplace mechanism for f defined by

$$M(x) = f(x) + (Z_1, \ldots, Z_k), \qquad Z_i \overset{i.i.d.}{\sim} \text{Laplace}\left(0, \frac{\Delta_f}{\varepsilon}\right)$$

is ε–differentially private

Proof

The conditional density of $M(x)$ is defined as

$$p(y|x) = \prod_{j=1}^{k} \frac{\varepsilon}{2\Delta_f} \exp\left(-\frac{\varepsilon|y_j - f(x)_j|}{\Delta_f}\right) = \left(\frac{\varepsilon}{2\Delta_f}\right)^k \exp\left(-\frac{\varepsilon\|y - f(x)\|_1}{\Delta_f}\right)$$

For every (x, x') such that $d(x, x') = 1$ and $y \in \mathbb{R}^k$,

$$\frac{p(y|x)}{p(y|x')} = \frac{\exp\left(-\frac{\varepsilon\|y-f(x)\|_1}{\Delta_f}\right)}{\exp\left(-\frac{\varepsilon\|y-f(x')\|_1}{\Delta_f}\right)} = \exp\left(-\frac{\varepsilon\left(\|y - f(x)\|_1 - \|y - f(x')\|_1\right)}{\Delta_f}\right)$$

$$\leq \exp\left(\frac{\varepsilon\|f(x) - f(x')\|_1}{\Delta_f}\right) \leq \exp(\varepsilon)$$

since by definition $\|f(x) - f(x')\|_1 \leq \Delta_f$ □

Estimating the Mean

When $x = (x_1, \ldots, x_n)$, where $x_i \in \mathbb{R}$ and $|x_i - x_j| \leq \Delta$, the mean $f(x) = \overline{x}_n$ has sensitivity Δ/n.

Laplace Mechanism for the Mean

The Laplace mechanism $M(x) = \overline{x}_n + Z$ where $Z \sim \text{Laplace}(0, \frac{\Delta}{n\varepsilon})$

- is ε–DP
- has (empirical) accuray $\mathbb{E}(M(x) - \overline{x}_n)^2 = \frac{2\Delta^2}{n^2\varepsilon^2}$
- if $X_i \overset{i.i.d.}{\sim} \mathbb{P}_X$ with expectation μ and variance σ^2, $M(X)$ has statistical accuracy

$$\mathbb{E}^{U,X}(M(X) - \mu)^2 = \frac{2\Delta^2}{n^2\varepsilon^2} + \frac{\sigma^2}{n} \leq \frac{\Delta^2}{4n}\left(1 + \frac{8}{n\varepsilon^2}\right)$$

which suggests that one can safely take ε to be of order of $1/\sqrt{n}$

A significant pratical concern occurs when the variations of x_i are not bounded or are unknown a priori. The usual solution consists in truncating the x_i's prior to applying the Laplace mechanism. The resulting mechanism is still ε–DP but its accuracy is usually degraded due to the presence of a bias (note that using a data-dependent truncation mechanism does require privacy accounting)

The former analysis directly extends to case of arbitrary empirical averages where $f(x) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$

Wen releasing multiples empirical averages for functions $\phi_1, \ldots, \phi_k$ (think of, e.g., wavelets or Fourier coefficients), one has to rely on the composition lemma and the Laplace noise has to be inflated by a factor k to ensure that the overall mechanism stays ε–DP

Histogram

The use of the general composition result may be pessimistic for specific sets of functions

Differentially Private Histogram

```
# DP Histogram
def dp_histogram(x, epsilon, bins=10):
    counts, bin_hedges = np.histogram(x, bins=bins, range=(0,1), density=False)
    Z = np.random.laplace(loc=0, scale=2/epsilon, size=bins)
    dp_counts = counts + Z
    dp_counts[dp_counts < 0] = 0    # Post processing
    Y = dp_counts/sum(dp_counts)
    return Y
```

is ε–DP (assuming $x_i \in [0, 1]$)

Report Noisy Max and Argmax

More generally, it is important to favor mechanisms (and privacy analyses) that release just the right amount of information

$\longrightarrow$ For example, neither $\max x$, or $\arg\max x$ require to release complete differentially private copies of x

- $\max x + Z$ where $Z \sim \text{Laplace}(0, \frac{1}{\varepsilon})$ is ε–DP (assuming $x_i \in [0, 1]$)
- $\arg\max\{x_i + Z_i\}_{1 \leq i \leq n}$ where $i \sim \text{Laplace}(0, \frac{1}{\varepsilon})$ (which is called "Report Noisy Max") is ε–DP [Claim 3.9 of Dwork & Roth, 2014]

Above Threshold: Sequentially Testing for a Statistic Above a Threshold

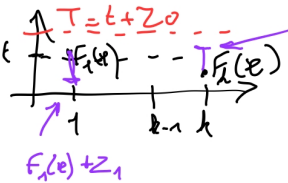
Consider sequentially comparing $f_1(x), f_2(x), \ldots$ (where $\Delta_{f_i} = 1$) to a threshold t, until one index k such that $f_k(x) > t$ is found.

$T = t + Z_0,\ Z_0 \sim \text{Laplace}\left(0, \frac{2}{\varepsilon}\right)$
for $i = 1, \ldots$ **do**
  $Z_i \sim \text{Laplace}\left(0, \frac{4}{\varepsilon}\right)$
  if $f_i(x) + Z_i > T$ **then**
    return i

```
# Above Threshold Algorithm
def above_threshold(x, stats, t, epsilon):
    T = t + np.random.laplace(loc=0, scale=2/epsilon)
    for i, f in enumerate(stats):
        Z_i = np.random.laplace(loc=0, scale=4/epsilon)
        if f(x) + Z_i > T:
            return i
    return None
```

Proposition

Above Threshold is ε–DP



Comparing $\mathbb{P}(A_k(x)=1)$ to $\mathbb{P}(A_k(x')=1)$

define
- A_k stopping indicator
- $\mathcal{F}_{k-1} = \sigma(x_1, \dots x_{k-1})$
- $g(x) = \max_{1 \le i \le k-1} (F_i(x) + Z_i) - t$

$$\mathbb{P}(A_k(x)=1 \mid \mathcal{F}_{k-1}) = \mathbb{P}(Z_0 \ge g(x),\ Z_k + F(x_k) \ge t + Z_0)$$
$$= \mathbb{P}(g(x) \le Z_0 \le Z_k + F(x_k) - t)$$

Change of prob: argument

$p(\tilde{z}_0) = p(z_0 + g(x') - g(x))$
$= \frac{\varepsilon}{4} \exp(-|z_0 + g(x') - g(x)|\ \varepsilon/2)$ ($g(x') - g(x) \le 1$)
$\le p(z_0) \exp(\varepsilon/2)$
$\Rightarrow \mathbb{P}(\tilde{Z}_0 \in S) \le e^{\varepsilon/2} \mathbb{P}(Z_0 \in S)$

$$= \mathbb{P}(g(x') \le z_0 + g(x') - g(x) \le z_k + \underbrace{g(x') - g(x)}_{\le 1} + \underbrace{f(x_k) - f(x'_k)}_{\le 1} + f(x'_k) - t)$$
$$\le e^{\varepsilon/2} \times \mathbb{P}(g(x') \le z_0 \le z_k + f(x'_k) - t) \times e^{\varepsilon/4} \times e^{\varepsilon/4} = e^{\varepsilon}$$
$$= \mathbb{P}(A_k(x')=1 \mid \mathcal{F}_{k-1})$$

Consider a target function $f : x \mapsto y \in \mathbb{R}$. Let

$$\Delta_f(x,k) = \sup_{x' : d(x,x') \leq k} |f(x) - f(x')| \qquad \text{and } \Delta_f(x) = \Delta_f(x,1)$$

Lower bound

If M is unbiased, in the sense that $\mathbb{E}[M(x)] = f(x)$ for every x. Then,

$$\mathbb{E}\left[\big(M(x) - f(x)\big)^2\right] \geq \sup_k \frac{\Delta_f(x,k)^2}{4(1+e^{2k\varepsilon})}$$

and, in particular if $\Delta_f(x,k) = k\Delta_f(x)$, then for $\varepsilon \leq 1/2$

$$\mathbb{E}\left[\big(M(x) - f(x)\big)^2\right] \geq \frac{\Delta_f(x)^2}{68\varepsilon^2}$$

For the mean $f(x) = \overline{x}_n$, $\Delta_f(x,k) = k\Delta_f(x) = k/n\Delta_x$ and $\mathbb{E}\left[\big(M(x) - f(x)\big)^2\right] \geq \frac{\Delta_x^2}{68n^2\varepsilon^2}$

Proof

Let $k \geq 1$ and let x' be such that $d(x, x') = k$ and $|f(x) - f(x')| = \Delta_f(x, k)$. By definition,

$$\mathbb{E}\left[\left(M(x') - t\right)^2\right] = \int_{\mathcal{Y}} (y-t)^2 p(y|x')dy \leq \int_{\mathcal{Y}} (y-t)^2 e^{k\varepsilon} p(y|x)dy = e^{k\varepsilon}\, \mathbb{E}\left[\left(M(x) - t\right)^2\right]$$

and hence

$$\frac{\mathbb{E}\left[\left(M(x') - f(x')\right)^2\right]}{\mathbb{E}\left[\left(M(x) - f(x)\right)^2\right]} = \frac{\mathbb{E}\left[\left(M(x') - f(x')\right)^2\right]}{\mathbb{E}\left[\left(M(x') - f(x)\right)^2\right]} \frac{\mathbb{E}\left[\left(M(x') - f(x)\right)^2\right]}{\mathbb{E}\left[\left(M(x) - f(x)\right)^2\right]} \leq 1 \times e^{k\varepsilon}$$

as M is unbiased. Therefore, by the Bienaymé-Chebishev inequality

$$\mathbb{P}\left(\left|M(x) - f(x)\right| \geq \frac{\Delta_f(x,k)}{2}\right) \leq \frac{4\, \mathbb{E}\left[\left(M(x) - f(x)\right)^2\right]}{\Delta_f(x,k)^2} \text{ and}$$

$$\mathbb{P}\left(\left|M(x) - f(x')\right| \geq \frac{\Delta_f(x,k)}{2}\right) \leq e^{k\varepsilon}\mathbb{P}\left(\left|M(x') - f(x')\right| \geq \frac{\Delta_f(x,k)}{2}\right) \leq \frac{4e^{2k\varepsilon}\, \mathbb{E}\left[\left(M(x) - f(x)\right)^2\right]}{\Delta_f(x,k)^2}$$

But since $|f(x') - f(x)| = \Delta_f(x,k)$,

$$1 \leq \mathbb{P}\left(\left|M(x) - f(x)\right| \geq \frac{\Delta_f(x,k)}{2}\right) + \mathbb{P}\left(\left|M(x) - f(x')\right| \geq \frac{\Delta_f(x,k)}{2}\right) \leq \frac{4\left(1 + e^{2k\varepsilon}\right) \mathbb{E}\left[\left(M(x) - f(x)\right)^2\right]}{\Delta_f(x,k)^2}$$

The second statement is obtained by the choice $k = \lceil 1/(2\varepsilon) \rceil$, noting that $1/(2\varepsilon) \leq k \leq 1/\varepsilon$ □

Table of Contents

Motivations

Our previous notion of DP ("pure DP") may be impossible to apply, in particular in continuous settings where $p(y|x)/p(y|x')$ can be unbounded.

A typical example is the Gaussian Mechanism

$$Y = f(x) + Z \quad \text{where } Z \sim \mathcal{N}(0, \sigma^2)$$

which cannot be ε–DP as

$$\log \frac{q_{\mathcal{N}(\mu,\sigma^2)}(y)}{q_{\mathcal{N}(\mu',\sigma^2)}(y)} = \frac{1}{\sigma^2}(\mu - \mu')\left(y - \frac{\mu + \mu'}{2}\right)$$

tends to $\pm\infty$ when $|y| \to \infty$.

But intuitively, the Gaussian mechanism is expected to provide privacy guarantees when $\sigma \gg |f(x) - f(x')|$.

Definition

(ε, δ)–Differential Privacy [Dwork *et al.*, 2006]

A $\mathcal{Y}$-valued mechanism M is (ε, δ)–DP if, for all neighboring x and x', and any set $\mathcal{S} \subset \mathcal{Y}$,

$$\mathbb{P}(M(x) \in \mathcal{S}) \leq e^{\varepsilon}\, \mathbb{P}(M(x') \in \mathcal{S}) + \delta$$

where $\mathbb{P}$ refers to the randomization in M.

ε has the same interpretation as in pure DP and δ is a slack variable that is homogeneous to a probability:

- The definition is not exactly symmetric in x and x' anymore
- The equal error rate γ of statistical tests for the value of x_i is now bounded as $\gamma \geq \frac{1-\delta}{1+e^{\varepsilon}}$

Properties

Post-Processing [Proposition 2.1 of Dwork & Roth, 2014]

If M_1 is (ε, δ)–DP, then for any independent mechanism M_2,

$$M_2 \circ M_1 \text{ is also } (\varepsilon, \delta)\text{-DP}$$

Composition [Theorem 3.16 of Dwork & Roth, 2014]

If M_1 and M_2 are independent mechanisms such that M_1 is $(\varepsilon_1, \delta_1)$–DP and M_2 is $(\varepsilon_2, \delta_2)$–DP,

$$M : x \mapsto (M_1(x), M_2(x)) \text{ is } (\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)\text{–DP}$$

Also holds for adaptive composition.

Why Should $\delta \ll 1/n$?

⚠one may satisfy (ε, δ)–DP without any guarantee on events of probability up to δ

- For instance, the Subsampling Mechanism which reports a fraction of the original n records of $x = (x_1, \ldots, x_n)$, using any randomized subsampling procedure such that the probability of reporting any individual x_i is equal to δ, is $(0, \delta)$–DP

- However, it is nonetheless a bad mechanism that will completely expose, on average, δn records
- Note that if $\delta n \ll 1$, it will also be useless as it will most often return an empty sample

To avoid such pathologic behaviors, we are typically interested in values of δ that are smaller than $1/n$.

The Privacy Curve

More fundamentally, this new definition associates a family of (ε, δ) values to a given mechanism.

Privacy Curve (or Privacy Profile)

A mechanism M is $(\varepsilon, \delta(\varepsilon))$–DP for

$$\delta(\varepsilon) = \sup_{x,x':d(x,x')=1} \sup_{\mathcal{S}\subset\mathcal{Y}} \left(\mathbb{P}[M(x) \in \mathcal{S}] - e^{\varepsilon}\mathbb{P}[M(x') \in \mathcal{S}]\right)$$

Note that $\delta(0)$ is the maximal value of the total variation norm between the distributions of $M(x)$ and $M(x')$.

The fact that there isn't a single value of ε and δ, together with the added difficulty of determining $\delta(\varepsilon)$, generates a (sometimes confusing) multiplicity of results.

Alternative Expression of the Privacy Curve

Introducing the privacy loss random variable $\ell_{x,x'}[M(x)]$ where $\ell_{x,x'}$ is the log likelihood ratio corresponding to M for datasets x and x':

$$\ell_{x,x'}(y) = \log \frac{p(y|x)}{p(y|x')}$$

Yields an alternative expression of the privacy curve:

$$\delta(\varepsilon) = \sup_{x,x':d(x,x')=1} \left\{ \mathbb{P}\left(\ell_{x,x'}[M(x)] \geq \varepsilon\right) - e^{\varepsilon}\mathbb{P}\left(\ell_{x',x}[M(x')] \leq -\varepsilon\right) \right\}$$

The Gaussian Mechanism

Let $f : x \mapsto f(x) \in \mathbb{R}^k$ denote a $\mathbb{R}^k$–valued function with ℓ^2 sensitivity

$$\Delta_{f,2} = \max_{x,x' : d(x,x')=1} \|f(x) - f(x')\|_2$$

where $\|f(x) - f(x')\|_2 = \sqrt{\sum_{i=1}^{k} (f(x)_i - f(x')_i)^2}$

Gaussian Mechanism

$$Y = f(x) + Z \quad \text{where } Z \sim \mathcal{N}(0, \sigma^2 I_k)$$

Consists in adding independent Gaussian (instead of Laplace) noises to the coordinates.

Privacy Curve of the Gaussian Mechanism [Theorem 8 of Balle & Wang, 2018]

$$\delta(\varepsilon) = \Phi\left(\frac{\Delta_{f,2}}{2\sigma} - \frac{\varepsilon\sigma}{\Delta_{f,2}}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta_{f,2}}{2\sigma} - \frac{\varepsilon\sigma}{\Delta_{f,2}}\right)$$

where Φ denotes the Gaussian CDF
i.e., $\Phi(x) = \mathbb{P}(X \leq x)$ when $X \sim \mathcal{N}(0,1)$.

- There is no simple relationship between δ, ε and σ.
- δ and ε depend only on the "signal-to-noise" ratio $\Delta_{f,2}/\sigma$ (which means that $\sigma \propto \Delta_{f,2}$).

Closed-Form Results

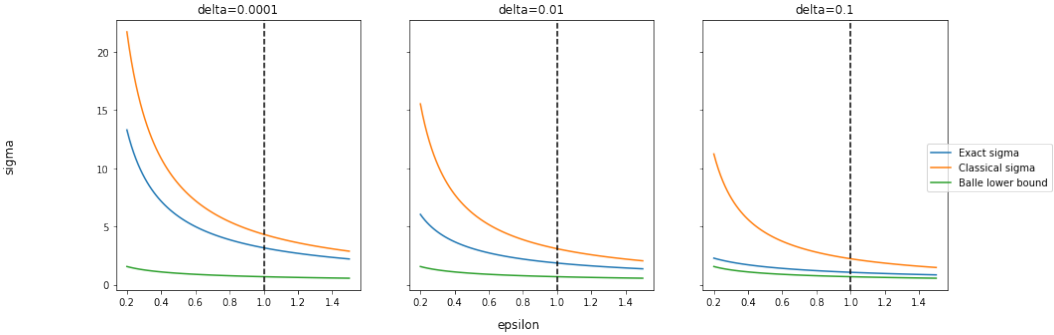
Except with numerical methods, the previous formula cannot be exploited, hence the need for closed-form (albeit suboptimal) results.

Property [Theorem 3.22 of Dwork & Roth, 2014]

For $\varepsilon < 1$, the Gaussian mechanism is (ε, δ)–DP as soon as

$$\sigma \geq \frac{\sqrt{2\log(\frac{1.25}{\delta})}\Delta_{f,2}}{\varepsilon}$$

$\rightarrow$ Prove this result when $k = 1$, with slighlty worse constant 2.65.


delta=0.0001
delta=0.01
delta=0.1
sigma
epsilon
Exact sigma
Classical sigma
Balle lower bound

The Basic Composition Result Is Not Tight for the Gaussian Mechanism

Consider k scalar functions $f_1, \dots, f_k$ with sensitivity bounded by Δ and the Gaussian mechanism M_k

$$M_k(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_k(x) \end{pmatrix} + Z \quad \text{where } Z \sim \mathcal{N}(0, \sigma^2 I_k)$$

$M_k(x)$ is (ε, δ)–DP for all k, as soon as

$$\sigma \propto \sqrt{k}\Delta$$

- One does not need to inflate σ proportionally to k (as the basic composition sugests or as we would need for $(\varepsilon, 0)$–DP with the Laplace mechanism).
- This gives rise to (many) so-called "advanced composition" results, as well as to alternative definitions of DP that handle composition more tightly.

Privacy Curve of the Laplace Mechanism

The Laplace mechanism itself can also be analyzed in term of (ε, δ)–DP

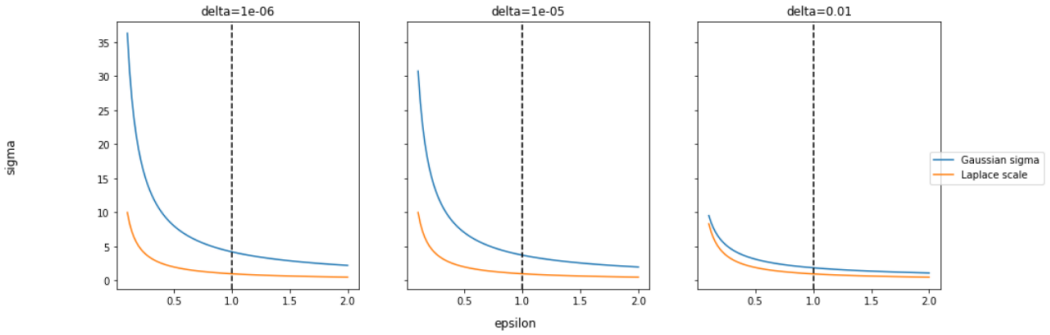
Privacy Curve of the Laplace Mechanism [Theorem 3 of Balle *et al.*, 2018]

If f is a scalar function with sensitivity Δ, the mechanism $M(x) = f(x) + Z$, where $Z \sim \mathrm{Laplace}(0, b)$ has privacy curve

$$\delta(\varepsilon) = \left[1 - \exp\left(\frac{\varepsilon - \Delta/b}{2}\right)\right]_+$$

(when $\varepsilon \geq \Delta/b$, $\delta(\varepsilon) = 0$).

- This implies that the Laplace mechanism is (ε, δ)–DP when $b \geq \frac{\Delta}{\varepsilon + 2\log\left(\frac{1}{1-\delta}\right)}$
- There also exist advanced composition results for the Laplace mechanism


delta=1e-06
delta=1e-05
delta=0.01
sigma
epsilon
Gaussian sigma
Laplace scale
0
5
10
15
20
25
30
35
0.5
1.0
1.5
2.0

Advanced Composition Result for ε–DP Mechanisms

Advanced Composition Lemma for ε–DP Mechanisms [Corollary 1 of Mironov 2017]

If $M_1, \ldots, M_k$ are k independent ε–DP mechanisms and δ is such that $\log(1/\delta) \geq \varepsilon^2 k$, then their composition $(M_1(x), \ldots, M_k(x))$ is (ε', δ)–DP, where

$$\varepsilon' = 4\varepsilon\sqrt{2k\log(1/\delta)}$$

Can be used for analyzing the composition of Laplace mechanisms in terms of (ε, δ)–DP.

Alternatives

Many other variants of the definition of DP have been proposed —see [Desfontaines & Pejó, 2020], among which

1. Rényi Differential Privacy [Mironov, 2017]
2. Gaussian Differential Privacy [Dong *et al.*, 2019]

Rényi Divergence

Rényi Divergence

For two probability densities p and q and $\alpha > 1$ the Rényi divergence of order α is

$$D_\alpha(p\|q) = \frac{1}{\alpha-1}\log \mathbb{E}_{Y\sim q}\left(\frac{p(Y)}{q(Y)}\right)^\alpha = \frac{1}{\alpha-1}\log \mathbb{E}_{Y\sim p}\left(\frac{p(Y)}{q(Y)}\right)^{\alpha-1}$$

where

$$D_{1+}(p\|q) = \mathbb{E}_{Y\sim p}\log\frac{p(Y)}{q(Y)}$$

is the Kullback-Leibler divergence and

$$D_\infty(p\|q) = \sup_y \log\frac{p(y)}{q(y)}$$

is the L^∞ norm of the log likelihood ratio (used in the definition of $(\varepsilon, 0)$–DP).

Properties of Rényi Divergence

1. $D_\alpha(p\|q) \geq 0$ (positivity)
2. If p' and q' are marginals of p and q, $D_\alpha(p'\|q') \leq D_\alpha(p\|q)$ (data-processing inequality)
3. $D_\alpha(wp_1 + (1-w)p_2 \| wq_1 + (1-w)q_2) \leq \max\{D_\alpha(p_1\|q_1), D_\alpha(p_2\|q_2)\}$ (quasi-convexity)
4. For $\alpha < \alpha'$, $D_\alpha(p\|q) \leq D_{\alpha'}(p\|q)$ (monotonicity)
5. $D_\alpha\left[\mathcal{N}(\mu,\sigma^2)\|\mathcal{N}(\mu',\sigma^2)\right] = \alpha\frac{(\mu-\mu')^2}{2\sigma^2}$

[Proposition 10 of Mironov, 2017]

For all $\mathcal{S}$ and α,

$$\mathbb{P}_p(\mathcal{S}) \leq \exp\left[D_\alpha(p\|q)\right]^{\frac{\alpha-1}{\alpha}} \mathbb{P}_q(\mathcal{S})^{\frac{\alpha-1}{\alpha}}$$

Rényi Differential Privacy [Mironov, 2017]

A mechanism $M(x)$ with conditional density $p(y|x)$ is (α, ε)–Rényi Differentially Private (RDP) if, for all neighboring x and x',

$$D_\alpha \left(p(\cdot|x) \| p(\cdot|x') \right) \leq \varepsilon$$

Properties

- Post-processing (as a consequence of data-processing inequality)
- Composition: If M_1 and M_2 are independent mechanisms that are, respectively, (α, ε_1) and (α, ε_2)–RDP, $(M_1(x), M_2(x))$ is $(\alpha, \varepsilon_1 + \varepsilon_2)$–RDP.
- Gaussian mechanism: $M(x) = f(x) + Z$, where $Z \sim \mathcal{N}(0, \sigma^2)$ and f has sensitivity Δ is $(\alpha, \alpha\Delta^2/(2\sigma^2))$–RDP (which implies that the composition of k Gaussian mechanisms has the correct $\sigma/\sqrt{k}$ noise scaling).

If $M(x)$ is (α, ε)–RDP, it is $(\varepsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$–DP [Proposition 3 of Mironov, 2017]

The notion of Gaussian DP (and its f–DP generalization) is specific in that it is directly formulated in terms of the control of the probabilities or error of any statistical test aiming to distinguish neighboring datasets x and x'.

Trade-off Function Let M denote a mechanism, x and x' datasets, φ a $\{0,1\}$–valued a decision rule and define

- $\alpha_\varphi = \mathbb{E}(\varphi(M(x)))$
- $\beta_\varphi = 1 - \mathbb{E}(\varphi(M(x')))$

be, respectively, the type I and type II probabilities of error.

The trade-off function $t_{x,x'}$ corresponding to M is the optimal tradeoff-curve for x, x' :

$$t_{x,x'}(\alpha) = \inf_\varphi \{\beta_\varphi : \alpha_\varphi \leq \alpha\}$$

(probability of error of the most powerful test of level α)

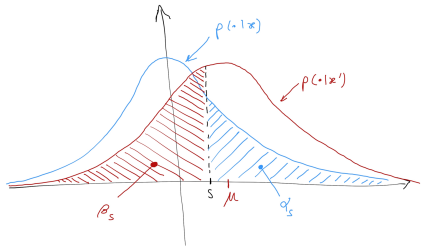
Gaussian Example

Assume that $M(x) \sim \mathcal{N}(0,1)$ while $M(x') \sim \mathcal{N}(\mu,1)$ and that one uses a decision rule on Y, the output of the mechanism, to test if the dataset is x or x'.

Neyman-Pearson Lemma

Optimal decision rules (called UMP—Uniformly Most Powerful) are of the form

$$\varphi_s(Y) = \begin{cases} 1 & \text{if } Y > s \\ 0 & \text{otherwise} \end{cases}$$



$$\alpha_s = 1 - \Phi(s)$$
$$\beta_s = \Phi(s - \mu)$$

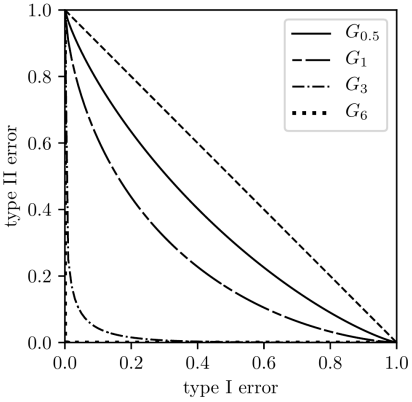
that is, $t(\alpha_s) = \beta_s = \Phi\left(\Phi^{-1}(1-\alpha_s) - \mu\right)$

μ–Gaussian Differential Privacy [Dong *et al.*, 2019]

M is μ–Gaussian Differentially Private (μ–GDP) if for all neighboring x and x' and $\alpha \in [0, 1]$

$$t_{x,x'}(\alpha) \geq G_\mu(\alpha)$$

where $G_\mu(\alpha) = \Phi(\Phi^{-1}(1-\alpha) - \mu)$ is the trade-off curve corresponding to the test of $\mathcal{N}(\mu, 1)$ versus $\mathcal{N}(0, 1)$.



The scalar Gaussian mechanism $M(x) = f(x) + Z$ with $Z \sim \mathcal{N}(0, \Delta^2/\mu^2)$, where Δ is the sensitivity of f, is G_μ–GDP.