Trustworthy Machine Learning From an Optimization Lens

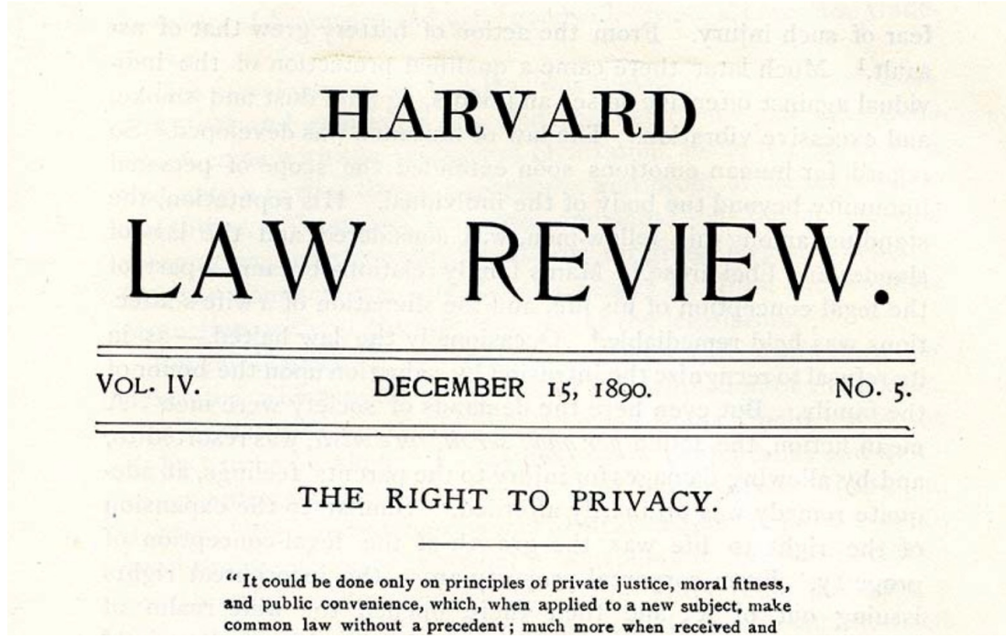
Meisam Razaviyayn

Lecture 8: Privacy Background and Differential Privacy

razaviya@usc.edu

Recap: What is privacy?

- "The right to be let alone." (Warren and Brandeis, 1890).
  - Instant photography + yellow journalism
  - Privacy protects dignity and emotional well-being, not just secrecy or property



HARVARD

LAW REVIEW.

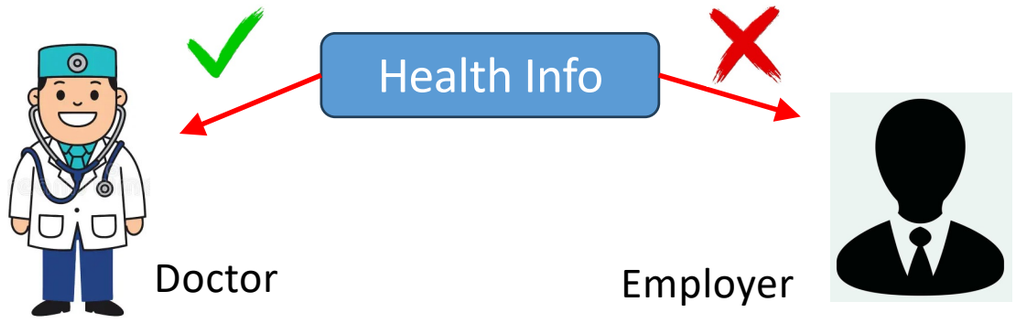
VOL. IV. DECEMBER 15, 1890. NO. 5.

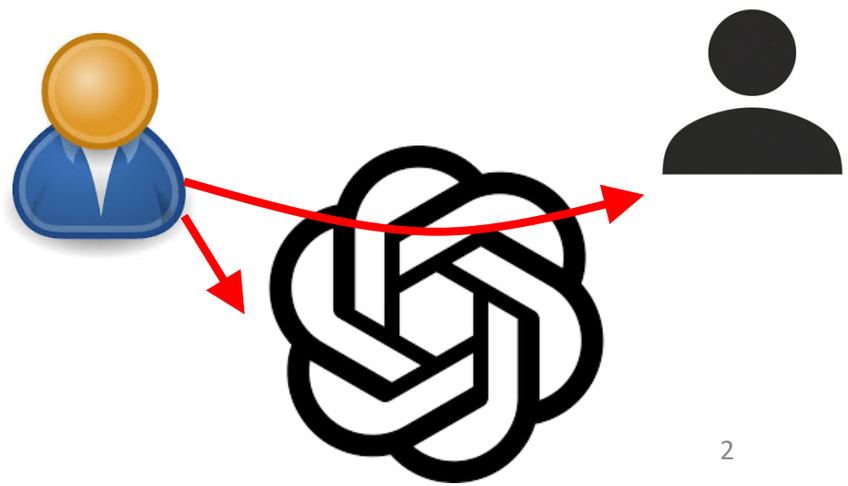
THE RIGHT TO PRIVACY.

"It could be done only on principles of private justice, moral fitness, and public convenience, which, when applied to a new subject, make common law without a precedent; much more when received and

- Modern view → appropriate use, not just secrecy

- Contextual Integrity (Helen Nissenbaum): **privacy ≈ appropriate info flow**





- AI often disrupts these norms.

Recap: Regulations on Privacy

- European Union's General Data Protection Regulation (GDPR)
- California Consumer Privacy Act of 2018
- AI Bill of Rights

- Key Principles:
  - Privacy by Design
  - Data minimization
- Right to be forgotten

THE WHITE HOUSE

DATA PRIVACY

YOU SHOULD BE PROTECTED FROM ABUSIVE DATA PRACTICES VIA BUILT-IN PROTECTIONS AND YOU SHOULD HAVE AGENCY OVER HOW DATA ABOUT YOU IS USED

GDPR.EU

Home Checklist FAQ GDPR News & Upda

Scope, penalties, and key definitions

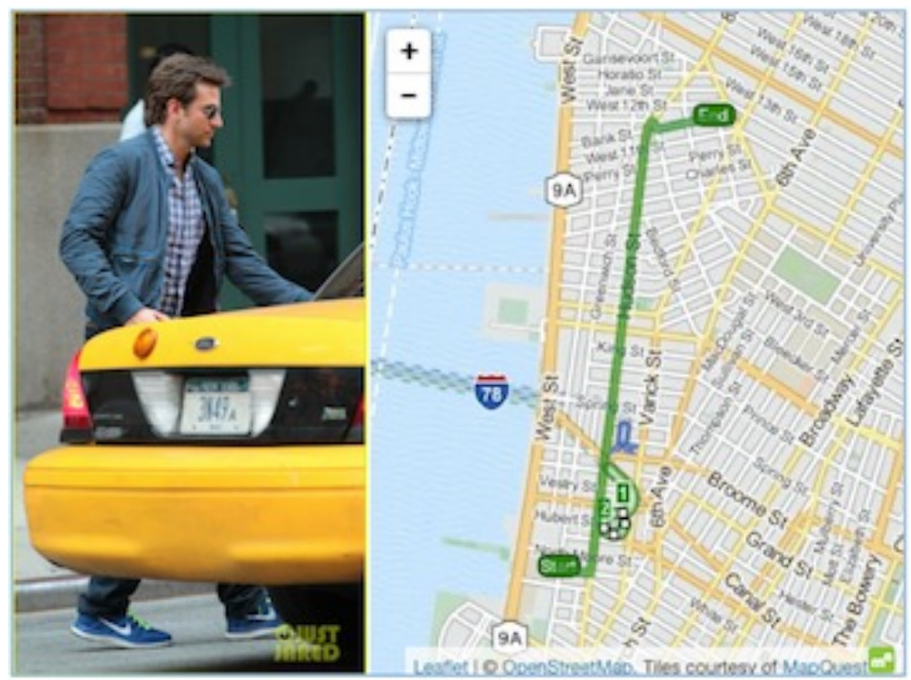
First, if you process the personal data of EU citizens or residents, or you offer goods or services to such people, then **the GDPR applies to you even if you're not in the EU**. We talk more about this in another article.

Second, the **fines for violating the GDPR are very high**. There are two tiers of penalties, which max out at €20 million or 4% of global revenue (whichever is higher), plus data subjects have the right to seek compensation for damages. We also talk **more about GDPR fines**.

The GDPR defines an array of legal terms at length. Below are some of the most important ones that we refer to in this article:

Recap: The Illusion of Anonymization (PII Removal)

- **Naive Approach**: Simply removing obvious Personally Identifiable Information (PII) like names or SSNs
- **Quasi-Identifiers (QIs)**: Attributes that don't uniquely identify on their own but can re-identify individuals when combined (e.g., Zipcode, Age, Gender).
- **Real-World Example 1**: NYC Taxi Data
  - Originally anonymized by removing driver names and medallion numbers
  - Researchers combined timestamps, pickup/dropoff locations (QIs) to re-identify drivers or trips
  - Reddit discussion: https://www.reddit.com/r/bigquery/comments/28ialf/comment/cicr3n2/
  - Chris Whong Analysis: https://chriswhong.com/open-data/foil_nyc_taxi/



Anonymization ≠ Privacy

Recap: (Pure) differential privacy

Let $\epsilon > 0$ and X be the set of possible datasets. A randomized algorithm $M(\cdot): X \to O$ is said to be $\epsilon -$ differentially private if

$$\Pr(M(D) \in \Omega) \leq e^{\epsilon} \Pr(M(D') \in \Omega)$$ for **all** $\Omega \subseteq O$ and **all** neighboring datasets $D, D' \in X$

- Proposed by Dwork, McSherry, Nissim, and. Smith [2017 Godel Prize]
- Property of the algorithm and not a particular output
- $M(\cdot)$ can even be public; only the randomness of the algorithm should be private
- Smaller ϵ means more privacy
- It hold even if the adversary has arbitrary auxiliary information
- Hypothesis testing viewpoint

Example: Randomized Response

- Randomized response is probably the oldest DP mechanism (Warner, 1965)
- For asking sensitive a question
  - Example: Do you frequently waste time at work?
  - Protocol:
    - 1. Flip a coin.
    - 2. If Heads (50%): Answer truthfully.
    - 3. If Tails (50%): Answer randomly (Yes/No with 50% chance each).
- Intuition: Provides "plausible deniability."
- Is this mechanism DP? What is the ϵ level?

But the task we are interested are not necessarily this simple…

Laplace Mechanism

Definition: Let $f: X \to R^k$. The $\ell_1 -$ sensitivity of f is defined as

$$\Delta_1 = \sup_{D,D'} \| f(D) - f(D') \|_1$$

where the supremum is taken over all neighboring datasets D and D'

Recall: Zero-mean Laplace distribution with parameter b: $p(x) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$

Theorem: Let $f: X \to R^k$ have the $\ell_1 -$ sensitivity Δ_1 and

$$M(D) = f(D) + (Z_1, .., Z_k)$$

Where $Z_1, \ldots, Z_k$ are iid Laplace random variables with parameter $\frac{\Delta_1}{\epsilon}$. Then $M(\cdot)$ is $\epsilon -$ DP.

Proof (look at only one output)

Example: Counting queries, histogram queries

Laplace Mechanism: Utility Guarantees

How much perturbations do we have in the Laplace mechanism?

Theorem: For the defined $\epsilon -$ DP Laplace mechanism $M(D) = f(D) + (Z_1, .., Z_k)$, we have

$$E[\| M(D) - f(D) \|_1] = K\frac{\Delta_1}{\epsilon}$$

and

$$\Pr\left(\| M(D) - f(D) \|_\infty > \frac{\Delta_1}{\epsilon}\ln\left(\frac{K}{\beta}\right)\right) < \beta$$

Properties of (pure) DP

Post-Processing: Let $M: X \to O$ be $\epsilon -$ DP and let $G: O \to T$ be an arbitrary (potentially randomized) mapping. Then, $G(M(\cdot))$ is also $\epsilon -$ DP.

Group Privacy: Let $M: X \to O$ be $\epsilon -$ DP and let D and D' be two datasets that differ in k entries. Then, for any $\Omega \in O$, we have $\Pr(M(D) \in \Omega) \leq \mathrm{e}^{k\epsilon} \Pr(M(D') \in \Omega)$.

Basic Composition: Let $M_1, .., M_T$ be a sequence of DP mechanisms. Assume, for each i, M_i is $\epsilon_i -$ DP. Define the (possibly adaptively) composition mechanism $M(D)$ that runs $M_1, .., M_T$ in order and each mechanism may take the previous output(s) as input. Then, $M(\cdot)$ is $\sum_{i=1}^{T} \epsilon_i -$ DP.

Exponential Mechanism: Motivation

- So far, we have learned Laplace mechanisms to make a mapping private

- But these mechanisms have inherent assumptions:
  - The outcome of data processing is a numerical value
    - Not all decisions are numeric: Should we choose vaccine A or vaccine B?
  - The utility is a continuous and well-behaved function of the output of the algorithm
    - Example: Assume we want to decide on the price of a good

Buyer	Price willing to pay
A	$2
B	$4

- If we set the price to $2.1, we will make $2.1
- If we set the price to $3.9, we will make $3.9
- If we set the price to $4.1, we will make $0

- Is Laplacian mechanism reasonable for such scenarios?

Exponential Mechanism: Preliminaries

- Assume we have a utility function for the outcome of our decision, e.g.

u(D, Vaccine A) = 10, u(D, Vaccine B) = 15

u(D, price = 2.1) = 2.1, u(D, price = 3.9) = 3.9, u(D, price = 4.1) = 0

Definition: The sensitivity of the utility function is defined as

$$\Delta_{\mathrm{u}} = \sup_{D,D',o} |u(D,o) - u(D',o)|$$

i.e., it is the worst-case change of the output when one entry is changed

Exponential Mechanism

Definition: Given the utility function $u(\cdot)$ with sensitivity Δ_u, we randomly choose an output o with probability $P(o) = \frac{\exp\left(u(D,o)\frac{\epsilon}{2\,\Delta}\right)}{\sum_{o'} \exp\left(u(D,o')\frac{\epsilon}{2\Delta}\right)}$. This mechanism is called *exponential mechanism*.

Theorem: The above exponential mechanism is $\epsilon -$ DP.

Proof?

- We can also obtain utility guarantees (see the book of Dwork and Roth)
- It is not always easy to implement the exponential mechanism in practice efficiently

(Approximate) differential privacy

- **Motivation**: Pure DP (δ=0) is often too restrictive

- **Example:** Pure DP does not allow simply using additive Gaussian noise (why?)
  - Gaussian noise has infinite tail
  - To satisfy Pure DP, the probability ratio must be bounded **everywhere**

$$\frac{\Pr(f(D)+z=x)}{\Pr(f(D')+z=x)}=?$$

- $(\epsilon,\delta)-DP$ (Approximate DP) is a relaxation of $\epsilon-DP$ (Pure DP)

Definition: Let $\epsilon,\delta>0$ and X be the set of possible datasets. A randomized algorithm $M(\cdot): X \to O$ is said to be $(\epsilon,\delta)-$ differentially private if

$\Pr(M(D)\in\Omega) \le e^{\epsilon}\Pr(M(D')\in\Omega)+\delta$ for all $\Omega \subseteq O$ and all neighboring datasets $D, D' \in X$

(Approximate) differential privacy

Definition: Let $\epsilon, \delta > 0$ and X be the set of possible datasets. A randomized algorithm $M(\cdot): X \to O$ is said to be $(\epsilon, \delta) -$ differentially private if

$\Pr(M(D) \in \Omega) \leq e^{\epsilon} \Pr(M(D') \in \Omega) + \delta$ for all $\Omega \subseteq O$ and all neighboring datasets $D, D' \in X$

Connection to pure DP

- **Definition**: Consider two fixed datasets D, D', and a randomized mechanism M. The privacy loss random variable draws an outcome o from $M(D)$ and outputs $\ln\left(\frac{P(M(D)=o)}{P(M(D')=o)}\right)$. In other words, the random variable takes the value of $\ln\left(\frac{P(M(D)=o)}{P(M(D')=o)}\right)$ with probability $P(M(D) = o)$.

- **Theorem:** If privacy loss is bounded by ϵ with probability $\geq 1 - \delta$, then the algorithm is $(\epsilon, \delta) -$ DP

Converse? Proof

Connection to pure DP

- **Theorem:** If privacy loss is bounded by ϵ with probability $\geq 1-\delta$, then the algorithm is $(\epsilon, \delta) -$ DP

- Proof: Let us denote $M(D)$ random variable with Y, and $M(D')$ with Z. Define $G = \{t \mid \ln\left(\frac{P(Y=t)}{P(Z=t)}\right) \leq \epsilon\}$.

$$P(Y \in \Omega) = \int_{\Omega} P(Y=t)dt = \int_{\Omega \cap G} P(Y=t)dt + \int_{\Omega \cap G^c} P(Y=t)dt$$

$$\leq \int_{\Omega \cap G} e^{\epsilon} P(Z=t)dt + P(Y \in \Omega \cap G^c) \leq e^{\epsilon} P(Z \in \Omega) + P(Y \in G^c) \leq e^{\epsilon} P(Z \in \Omega) + \delta$$

- Converse? Consider the following algorithm

$$P(M(D)=0) = \frac{1}{2} \qquad P(M(D)=1) = \frac{1}{2}$$

$$P(M(D')=0) = \frac{1}{3} \qquad P(M(D')=1) = \frac{2}{3}$$

- The algorithm is $\left(\epsilon = 0, \delta = \frac{1}{6}\right)$ –DP. Is the privacy loss bounded by 0 w.p 5/6?

More discussions on (approximate) DP

$\Pr(M(D) \in \Omega) \leq e^{\epsilon} \Pr(M(D') \in \Omega) + \delta$ for all $\Omega \subseteq O$ and all neighboring datasets $D, D' \in X$

- (ϵ, δ) value is not necessarily unique (by increasing one we may be able to decrease the other
- Unlike pure DP case, it is **NOT** sufficient to show the definition hold for every outcome

Example:

- Consider an algorithm that outputs the entire dataset and a random number in $\{1,2, \dots, m\}$
- For any possible outcome o, we have $P(M(D) = o) = 0 \; or \; \frac{1}{m}$
- Therefore, $P(M(D) = o) \leq P(M(D') = o) + \frac{1}{m}, \forall \, D, D'$
- But is this algorithm $\left(0, \frac{1}{m}\right)$ –private? Does it satisfy the condition $P(M(D) \in \Omega) \leq P(M(D') \in \Omega) + \frac{1}{m}$

More discussions on (approximate) DP

$\Pr(M(D) \in \Omega) \leq \mathrm{e}^{\epsilon} \Pr(M(D') \in \Omega) + \delta$ for all $\Omega \subseteq O$ and all neighboring datasets $D, D' \in X$

- Should be careful about the choice of δ
  - An algorithm that does nothing or returns the entire dataset with probability δ is $(0, \delta) -$ DP
  - We typically choose $\delta \ll \frac{1}{n}$
  - Worst-case scenarios are less likely to happen in practice (algorithm is DP for various (ϵ, δ) values)
- **Hypothesis testing viewpoint**: Let $o = M(\cdot)$ be the algorithm output. Consider the hypothesis testing:

$$\begin{cases} H_0: & o \text{ came from } D_0 \\ H_1: & o \text{ came from } D_1 \end{cases}$$

- Then, $P_{MD} + e^{\epsilon} P_{FA} \geq 1 - \delta$ and $P_{FA} + e^{\epsilon} P_{MD} \geq 1 - \delta$.

Proof: Let S be the rejection region (declaring H_1). Then,

$$1 - P_{FA} = P(H_0|H_0) = P(M(D_0) \in S^c) \leq e^{\epsilon} P(M(D_1) \in S^c) + \delta = e^{\epsilon} P(H_0|H_1) + \delta = e^{\epsilon} P_{MD} + \delta$$

Similarly, prove the other one by changing the role of D_0 and D_1

Gaussian Mechanism

Definition: Let $f: X \to R^k$. The $\ell_2 -$ sensitivity of f is defined as

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

where the supremum is taken over all neighboring datasets D and D'

Recall: Zero-mean Gaussian distribution with variance σ^2: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$

Theorem: Let $f: X \to R^k$ have the $\ell_2 -$ sensitivity Δ_2 and

$$M(D) = f(D) + (Z_1, .., Z_k)$$

Where $Z_1, \ldots, Z_k$ are iid Gaussian with variance $\sigma = \frac{\Delta_2 \sqrt{2 \ln \frac{1.25}{\delta}}}{\epsilon}$. Then $M(\cdot)$ is $(\epsilon, \delta) -$ DP.

Example: *Mean estimation*: If data is bounded ($\left|\left|x_i\right|\right|_2 \leq c$), the sensitivity of the mean (assuming neighboring datasets defined by replacement) is $\Delta_2 = 2c/n$

Gaussian Mechanism: Proof of Privacy

Theorem: Let $f: X \to R^k$ have the ℓ_2 − sensitivity Δ_2 and $M(D) = f(D) + (Z_1, .., Z_k)$

Where $Z_1, \ldots, Z_k$ are iid Gaussian with variance $\sigma = \frac{\Delta_2 \sqrt{2 \ln \frac{1.25}{\delta}}}{\epsilon}$. Then $M(\cdot)$ is (ϵ, δ) − DP.

- Proof sketch (for $k = 1$ for simplicity):
  - Recall the connection between pure and approximate DP
  - We need to compute the probability of the event $\log\left(\frac{P(M(D)=x)}{P(M(D')=x)}\right) \leq \epsilon$
  - $\log\left(\frac{P(M(D)=x)}{P(M(D')=x)}\right) = -\frac{1}{2\sigma^2}\left(\left(x - f(D)\right)^2 - \left(x - f(D')\right)^2\right) \leq \frac{1}{2\sigma^2}(2\Delta_2 Z + \Delta_2^2)$
  - To have (ϵ, δ) −DP, it suffices to have $Z \leq \frac{\sigma^2 \epsilon}{\Delta_2} - \frac{\Delta_2}{2}$ with probability 1- δ
  - Using the tail bound on the CDF of Gaussian, it suffices to have $\sigma = \frac{\Delta_2 \sqrt{2 \ln \frac{1.25}{\delta}}}{\epsilon}$

Comments on Gaussian Mechanism

Theorem: Let $f: X \to R^k$ have the $\ell_2 -$ sensitivity Δ_2 and $M(D) = f(D) + (Z_1, .., Z_k)$

Where $Z_1, \ldots, Z_k$ are iid Gaussian with variance $\sigma = \frac{\Delta_2 \sqrt{2 \ln \frac{1.25}{\delta}}}{\epsilon}$. Then $M(\cdot)$ is $(\epsilon, \delta) -$ DP.

- Dependence on $1/\delta$ is logarithmic. Remember δ needs to be chosen very small
- It is not possible to achieve $\delta = 0$. For pure DP, we need Laplace mechanism
- A given noise with a fixed variance will be approximate DP for different values of (ϵ, δ).
- We can obtain high probability bounds on the deviations:

Theorem: For the defined $(\epsilon, \delta) -$ DP Gaussian mechanism $M(D) = f(D) + (Z_1, .., Z_k)$, we have

$$\Pr\left(\| M(D) - f(D) \|_\infty > \frac{\Delta_2}{\epsilon} \sqrt{2 \ln(1.25/\delta) \ln\left(\frac{K}{\beta}\right)} \right) < \beta$$

Properties of approximate DP

Post-Processing: Let $M: X \to O$ be $(\epsilon, \delta) -$ DP and let $G: O \to T$ be an arbitrary (potentially randomized) mapping. Then, $G(M(\cdot))$ is also $(\epsilon, \delta) -$ DP.

Use cases: Integer optimization/decisions, projections involved, etc.

Group Privacy: Let $M: X \to O$ be $(\epsilon, \delta) -$ DP and let D and D' be two datasets that differ in k entries. Then, for any $\Omega \in O$, we have $\Pr(M(D) \in \Omega) \leq \mathrm{e}^{k\epsilon} \Pr(M(D') \in \Omega) + \delta \frac{e^{k\epsilon}-1}{e^{\epsilon}-1}$.

Proof?