

Trustworthy Machine Learning

Differential Privacy 1

Sangdon Park

POSTECH

Contents from

Foundations and Trends® in
Theoretical Computer Science
Vol. 9, Nos. 3–4 (2014) 211–407
© 2014 C. Dwork and A. Roth
DOI: 10.1561/0400000042

now
the essence of knowledge

The Algorithmic Foundations of Differential Privacy

Cynthia Dwork
Microsoft Research, USA
dwork@microsoft.com

Aaron Roth
University of Pennsylvania, USA
aaroath@cis.upenn.edu

- and contents partially from Gautam Kamath at University of Waterloo and Roger Grosse at University of Toronto.

Why Privacy Guarantees in Learning?

- Not anonymized dataset for learning – privacy leak

Why Privacy Guarantees in Learning?

- Not anonymized dataset for learning – privacy leak
- Anonymized dataset for learning – looks okay but possible to leak private information

Why Privacy Guarantees in Learning?

Anonymized Dataset

Name	Age	Gender	Zip Code	Smoker	Diagnosis
*	60-70	Male	191**	Y	Heart disease
*	60-70	Female	191**	N	Arthritis
*	60-70	Male	191**	Y	Lung cancer
*	60-70	Female	191**	N	Crohn's disease
*	60-70	Male	191**	Y	Lung cancer
*	<i>50-60</i>	<i>Female</i>	191**	N	HIV
*	50-60	Male	191**	Y	Lyme disease
*	50-60	Male	191**	Y	Seasonal allergies
*	<i>50-60</i>	<i>Female</i>	191**	N	Ulcerative colitis

Figure: An example from Kearns & Roth, The Ethical Algorithm

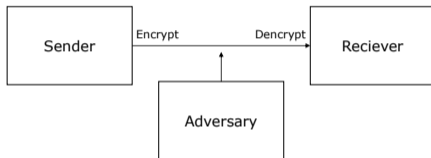
- anonymized dataset – looks okay but still privacy leak
 - ▶ If we know Rebecca is 55 years old and in this database, then we know she has 1 of 2 diseases.

Why Not Use Cryptosystems?

- Just encrypt the entire data.

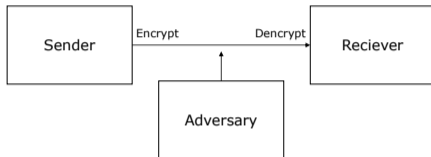
Why Not Use Cryptosystems?

- Just encrypt the entire data.
- Entities in encryption: Sender, Receiver, and Adversary



Why Not Use Cryptosystems?

- Just encrypt the entire data.
- Entities in encryption: Sender, Receiver, and Adversary

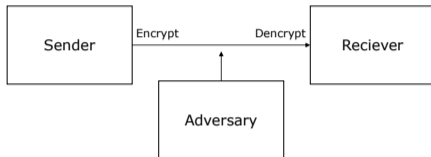


- Entities in learning, Sender (a.k.a. Curator) and Receiver = Adversary (a.k.a. Analyst)



Why Not Use Cryptosystems?

- Just encrypt the entire data.
- Entities in encryption: Sender, Receiver, and Adversary



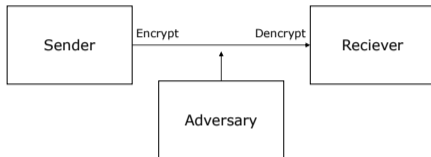
- Entities in learning, Sender (a.k.a. Curator) and Receiver = Adversary (a.k.a. Analyst)



- ▶ e.g., a learning algorithm (=curator) releases a model for some “benefits”

Why Not Use Cryptosystems?

- Just encrypt the entire data.
- Entities in encryption: Sender, Receiver, and Adversary



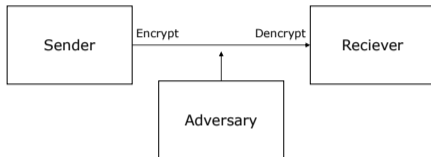
- Entities in learning, Sender (a.k.a. Curator) and Receiver = Adversary (a.k.a. Analyst)



- ▶ e.g., a learning algorithm (=curator) releases a model for some “benefits”
- ▶ But, the model should not reveal private information.

Why Not Use Cryptosystems?

- Just encrypt the entire data.
- Entities in encryption: Sender, Receiver, and Adversary



- Entities in learning, Sender (a.k.a. Curator) and Receiver = Adversary (a.k.a. Analyst)



- ▶ e.g., a learning algorithm (=curator) releases a model for some “benefits”
 - ▶ But, the model should not reveal private information.
- Note that homomorphic encryption could be alternatives but slow (yet)

Goal for Privacy In Learning

Goal

Learning nothing about an individual while learning useful information about a population.

- Here, an algorithm can transform a dataset into another dataset
- Sender: an algorithm that releases a model
- Receiver: A model user

Goal for Privacy In Learning

Goal

Learning nothing about an individual while learning useful information about a population.

- Here, an algorithm can transform a dataset into another dataset
- Sender: an algorithm that releases a model
- Receiver: A model user
- How to achieve this goal? Add noise!

Randomized Response

An Example

- Goal of a survey: estimate a statistic on illegal behaviors of participants
 - ▶ Curator: a participant
 - ▶ Analyst: a researcher

Randomized Response

An Example

- Goal of a survey: estimate a statistic on illegal behaviors of participants
 - ▶ Curator: a participant
 - ▶ Analyst: a researcher
- Each participant follows the following survey process:
 - ① Flip a coin
 - ② If “tails”, then respond truthfully.
 - ③ If “heads”, then flip a second coin and respond “Yes” if “heads” and “No” if “tails”.

Randomized Response

General Description

$$Y_i = \begin{cases} X_i & \text{with probability } \frac{1}{2} + \gamma \\ 1 - X_i & \text{with probability } \frac{1}{2} - \gamma, \end{cases}$$

- $X_i \in \{0, 1\}$: the truthful response
- $Y_i \in \{0, 1\}$: a randomized response
- $\gamma = 0$: a uniformly random strategy
 - ✓ private
 - ✗ not informative
- $\gamma = 1/2$: an honest strategy
 - ✗ no privacy
 - ✓ informative
- $\gamma = 1/4$: the previous example.
 - ✓ private \rightarrow no learning on an individual response
 - ✓ informative \rightarrow learning on a population statistic

Randomized Response

How Informative?

$$Y_i = \begin{cases} X_i & \text{with probability } \frac{1}{2} + \gamma \\ 1 - X_i & \text{with probability } \frac{1}{2} - \gamma, \end{cases}$$

- How to estimate $p = \mathbb{E}\{X_i\}$?

Randomized Response

How Informative?

$$Y_i = \begin{cases} X_i & \text{with probability } \frac{1}{2} + \gamma \\ 1 - X_i & \text{with probability } \frac{1}{2} - \gamma, \end{cases}$$

- How to estimate $p = \mathbb{E}\{X_i\}$?
- Observe that

$$\mathbb{E}\{Y_i\} = X_i \left(\frac{1}{2} + \gamma\right) + (1 - X_i) \left(\frac{1}{2} - \gamma\right) = 2\gamma X_i + \frac{1}{2} - \gamma \implies X_i = \mathbb{E} \left\{ \frac{1}{2\gamma} \left(Y_i - \frac{1}{2} + \gamma \right) \right\}$$

Randomized Response

How Informative?

$$Y_i = \begin{cases} X_i & \text{with probability } \frac{1}{2} + \gamma \\ 1 - X_i & \text{with probability } \frac{1}{2} - \gamma, \end{cases}$$

- How to estimate $p = \mathbb{E}\{X_i\}$?
- Observe that

$$\mathbb{E}\{Y_i\} = X_i \left(\frac{1}{2} + \gamma\right) + (1 - X_i) \left(\frac{1}{2} - \gamma\right) = 2\gamma X_i + \frac{1}{2} - \gamma \implies X_i = \mathbb{E} \left\{ \frac{1}{2\gamma} \left(Y_i - \frac{1}{2} + \gamma \right) \right\}$$

- Consider the following estimator:

$$\hat{p} := \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2\gamma} \left(Y_i - \frac{1}{2} + \gamma \right) \right)$$

Randomized Response

How Informative?

$$Y_i = \begin{cases} X_i & \text{with probability } \frac{1}{2} + \gamma \\ 1 - X_i & \text{with probability } \frac{1}{2} - \gamma, \end{cases}$$

- How to estimate $p = \mathbb{E}\{X_i\}$?
- Observe that

$$\mathbb{E}\{Y_i\} = X_i \left(\frac{1}{2} + \gamma\right) + (1 - X_i) \left(\frac{1}{2} - \gamma\right) = 2\gamma X_i + \frac{1}{2} - \gamma \implies X_i = \mathbb{E} \left\{ \frac{1}{2\gamma} \left(Y_i - \frac{1}{2} + \gamma \right) \right\}$$

- Consider the following estimator:

$$\hat{p} := \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2\gamma} \left(Y_i - \frac{1}{2} + \gamma \right) \right)$$

- We have

$$\mathbb{E}\{\hat{p}\} = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2\gamma} \left(\mathbb{E}\{Y_i\} - \frac{1}{2} + \gamma \right) \right) = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2\gamma} \left(\left(2\gamma \mathbb{E}\{X_i\} + \frac{1}{2} - \gamma \right) - \frac{1}{2} + \gamma \right) \right) = \mathbb{E}\{X_i\}.$$

Randomized Response

How Informative?

$$Y_i = \begin{cases} X_i & \text{with probability } \frac{1}{2} + \gamma \\ 1 - X_i & \text{with probability } \frac{1}{2} - \gamma, \end{cases}$$

- How to estimate $p = \mathbb{E}\{X_i\}$?
- Observe that

$$\mathbb{E}\{Y_i\} = X_i \left(\frac{1}{2} + \gamma\right) + (1 - X_i) \left(\frac{1}{2} - \gamma\right) = 2\gamma X_i + \frac{1}{2} - \gamma \implies X_i = \mathbb{E} \left\{ \frac{1}{2\gamma} \left(Y_i - \frac{1}{2} + \gamma \right) \right\}$$

- Consider the following estimator:

$$\hat{p} := \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2\gamma} \left(Y_i - \frac{1}{2} + \gamma \right) \right)$$

- We have

$$\mathbb{E}\{\hat{p}\} = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2\gamma} \left(\mathbb{E}\{Y_i\} - \frac{1}{2} + \gamma \right) \right) = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2\gamma} \left(\left(2\gamma \mathbb{E}\{X_i\} + \frac{1}{2} - \gamma \right) - \frac{1}{2} + \gamma \right) \right) = \mathbb{E}\{X_i\}.$$

- The randomized response looks “working”! How to rigorously measure whether this method does not leak privacy?

Differential Privacy

Definition

A randomized algorithm \mathcal{M} is (ϵ, δ) -differentially private if for any $\mathcal{S} \in \text{Range}(\mathcal{M})$ and for any two “neighboring” datasets \mathcal{D}_1 and \mathcal{D}_2 ,

$$\mathbb{P} \{ \mathcal{M}(\mathcal{D}_1) \in \mathcal{S} \} \leq \exp(\epsilon) \mathbb{P} \{ \mathcal{M}(\mathcal{D}_2) \in \mathcal{S} \} + \delta,$$

where the probability is taken over the randomness of \mathcal{M} .

- If \mathcal{M} behaves nearly identically for \mathcal{D}_1 and \mathcal{D}_2 , an attacker cannot tell whether \mathcal{D}_1 or \mathcal{D}_2 was used (so cannot learn an individual information)

Randomized Response is DP

Theorem

The randomized response is $(\ln 3, 0)$ -differentially private.

Randomized Response is DP

Theorem

The randomized response is $(\ln 3, 0)$ -differentially private.

Proof sketch.

- \mathcal{M} : a randomized response
 - ▶ $\mathcal{M}(X_1, \dots, X_n) = (Y_1, \dots, Y_n)$
- Let $\gamma = \frac{1}{4}$
- Consider any realization $\mathcal{S} \in \{0, 1\}^n$ of (Y_1, \dots, Y_n) .
- Consider X and X' which differ only in coordinate j .
- Then, we have

$$\frac{\mathbb{P}\{\mathcal{M}(X) = a\}}{\mathbb{P}\{\mathcal{M}(X') = a\}} = \frac{\prod_{i=1}^n \mathbb{P}\{\mathcal{M}(X_i) = a_i\}}{\prod_{i=1}^n \mathbb{P}\{\mathcal{M}(X'_i) = a_i\}} = \frac{\mathbb{P}\{\mathcal{M}(X_j) = a_j\}}{\mathbb{P}\{\mathcal{M}(X'_j) = a_j\}} = \frac{\mathbb{P}\{Y_j = a_j\}}{\mathbb{P}\{Y'_j = a_j\}} \leq \frac{1/2 + \gamma}{1/2 - \gamma} = e^{\ln 3}.$$

- ▶ The inequality holds when $X_j = a_j$ and $X'_j \neq a_j$.

Laplace Mechanism

Definition

Given any function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$, the Laplace mechanism is defined as:

$$\mathcal{M}_L(x, f, \varepsilon) := f(x) + (Y_1, \dots, Y_k),$$

where Y_i are i.i.d. random variables drawn from $\text{Lap}\left(f(x)_i | \frac{\Delta f}{\varepsilon}\right)$.

- $\text{Lap}(x|b) = \text{Lap}(b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$
- The ℓ_1 -sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ is

$$\Delta f := \max_{x, y \in \mathbb{N}^{|\mathcal{X}|}, \|x-y\|_1=1} \|f(x) - f(y)\|_1.$$

- e.g., x is a dataset and f is a post-processor.

Laplace Mechanism is DP

Theorem

The Laplace mechanism preserves $(\epsilon, 0)$ -differential privacy.

Laplace Mechanism is DP

Proof Sketch

- Let $x \in \mathbb{N}^{|\mathcal{X}|}$ and $y \in \mathbb{N}^{|\mathcal{X}|}$ be such that $\|x - y\|_1 \leq 1$
- p_x : the PDF of $\mathcal{M}_L(x, f, \varepsilon)$
- p_y : the PDF of $\mathcal{M}_L(y, f, \varepsilon)$
- For any $z \in \mathbb{R}^k$, we have

$$\begin{aligned}\frac{p_x(z)}{p_y(z)} &= \prod_{i=1}^k \left(\exp\left(-\frac{\varepsilon|f(x)_i - z_i|}{\Delta f}\right) / \exp\left(-\frac{\varepsilon|f(y)_i - z_i|}{\Delta f}\right) \right) \\ &= \prod_{i=1}^k \exp\left(\frac{\varepsilon(|f(y)_i - z_i| - |f(x)_i - z_i|)}{\Delta f}\right) \\ &\leq \prod_{i=1}^k \exp\left(\frac{\varepsilon|f(x)_i - f(y)_i|}{\Delta f}\right) \\ &= \exp\left(\frac{\varepsilon\|f(x) - f(y)\|_1}{\Delta f}\right) \\ &\leq \exp(\varepsilon).\end{aligned}$$

- $\frac{p_x(z)}{p_y(z)} \geq \exp(-\varepsilon)$ follows by symmetry.

Gaussian Mechanism is DP

Definition

Let $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^d$ be an arbitrary d -dimensional function, and define its ℓ_2 sensitivity to be $\Delta_2 f = \max_{\text{adjacent } x, y} \|f(x) - f(y)\|_2$. The Gaussian Mechanism with parameter σ adds noise scaled to $\mathcal{N}(0, \sigma^2)$ to each of the d components of the output.

Theorem

Let $\varepsilon \in (0, 1)$ be arbitrary. For $c^2 > 2 \ln \left(\frac{1.25}{\delta} \right)$, the Gaussian Mechanism with parameter $\sigma \geq \frac{c \Delta_2 f}{\varepsilon}$ is (ε, δ) -differentially private.

How Can It be Connected to Learning?

Journal of Machine Learning Research 12 (2011) 1069-1109

Submitted 6/10; Revised 2/11; Published 3/11

Differentially Private Empirical Risk Minimization

Kamalika Chaudhuri

*Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093, USA*

KCHAUDHURI@UCSD.EDU

Claire Monteleoni

*Center for Computational Learning Systems
Columbia University
New York, NY 10115, USA*

CMONTEL@CCLS.COLUMBIA.EDU

Anand D. Sarwate

*Information Theory and Applications Center
University of California, San Diego
La Jolla, CA 92093-0447, USA*

ASARWATE@UCSD.EDU

Editor: Nicolas Vayatis

Empirical Risk Minimization (ERM)

Setup

- \mathcal{X} : an example space
 - ▶ Assume that $\|x\|_2 \leq 1$ for $x \in \mathcal{X}$
- \mathcal{Y} : a label space
- $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$: a training set
- $f : \mathcal{X} \rightarrow \mathcal{Y}$: a predictor
- $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$: a loss function
- Regularized empirical risk minimization:

$$J(f, \mathcal{D}) := \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i) + \Lambda N(f),$$

where $N(f)$ is a regularizer.

Assumptions

Definition

A function $H(f)$ over $f \in \mathbb{R}^d$ is said to be *strictly convex* if for all $\alpha \in (0, 1)$, f , and $g (\neq f)$,

$$H(\alpha f + (1 - \alpha)g) < \alpha H(f) + (1 - \alpha)H(g).$$

It is said to be λ -*strongly convex* if for all $\alpha \in (0, 1)$, f , and $g (\neq f)$,

$$H(\alpha f + (1 - \alpha)g) \leq \alpha H(f) + (1 - \alpha)H(g) - \frac{1}{2}\lambda\alpha(1 - \alpha)\|f - g\|_2^2.$$

- A strictly convex function has a unique minimum.
- (strongly convex) \implies (strictly convex)
- The regularizer $N(\cdot)$ and loss $\ell(\cdot, \cdot)$ are differentiable.
 - ▶ No ℓ_1 -norm regularizer
 - ▶ No hinge loss

Privacy Model

Goal: Produce a classifier which preserves the privacy of individual entities of the training set \mathcal{D} .

Definition (ϵ -differential privacy)

An algorithm \mathcal{A} provides ϵ -differential privacy if for any two data sets \mathcal{D} and \mathcal{D}' that differ in a single entry and for any \mathcal{S}

$$e^{-\epsilon} \leq \frac{\mathbb{P}\{\mathcal{A}(\mathcal{D}) \in \mathcal{S}\}}{\mathbb{P}\{\mathcal{A}(\mathcal{D}') \in \mathcal{S}\}} \leq e^{\epsilon}.$$

- $\mathcal{A}(\mathcal{D})$: a randomized algorithm that produces a classifier from a training set \mathcal{D} .
- \mathcal{D}' and \mathcal{D} have $n - 1$ samples (x_i, y_i) in common; the differing sample contains private values.

Is ERM differentially private?

- Given \mathcal{D} and \mathcal{D}' , let

$$f_{\mathcal{D}}^* = \arg \min_f J(f, \mathcal{D}) \quad \text{and} \quad f_{\mathcal{D}'}^* = \arg \min_f J(f, \mathcal{D}')$$

- Letting $\mathcal{S} = \{f_{\mathcal{D}}^*\}$,

$$\mathbb{P}\{f_{\mathcal{D}}^* \in \mathcal{S}\} = 1 \neq \mathbb{P}\{f_{\mathcal{D}'}^* \in \mathcal{S}\} = 0$$

- Not differentially private!

Algorithm 1: Output Perturbation

output perturbation

$$f_{\text{priv}} = \arg \min_f J(f, \mathcal{D}) + \mathbf{b}$$

- \mathbf{b} is random noise with density

$$v(\mathbf{b}) \propto e^{-\beta \|\mathbf{b}\|}$$

with $\beta = \frac{n\Delta\epsilon}{2}$.

Algorithm 2: Objective Perturbation

output perturbation

$$f_{\text{priv}} = \arg \min_f J(f, \mathcal{D}) + \frac{1}{n} \mathbf{b}^T f$$

- \mathbf{b} is random noise with density

$$v(\mathbf{b}) \propto e^{-\beta \|\mathbf{b}\|}$$

with $\beta = \frac{\varepsilon - \log\left(1 + \frac{2c}{n\Lambda} + \frac{c^2}{n^2\Lambda^2}\right)}{2}$ (assuming ε is chosen to be $\beta > 0$).

Privacy Guarantee

Theorem

If $N(\cdot)$ is differentiable and 1-strongly convex, and ℓ is convex and differentiable with $|\ell'(z)| \leq 1$ for all z , then Algorithm 1 provides ε -differential privacy.

Theorem

If $N(\cdot)$ is **doubly** differentiable and 1-strongly convex, and ℓ is convex and **doubly** differentiable with $|\ell'(z)| \leq 1$ and $|\ell''(z)| \leq c$ for all z , then Algorithm 2 provides ε -differential privacy.

- Algorithm 2 requires stronger assumptions.
- What's the benefit of Algorithm 2?

Correctness Guarantee

Lemma

Suppose $N(\cdot)$ is doubly differentiable with $\|\nabla N(f)\|_2 \leq \eta$ for all f , ℓ is differentiable and has continuous c -Lipschitz derivatives. Then, letting f_{priv} be the output of Algorithm 1, we have

$$\mathbb{P}_{\mathbf{b}} \left\{ J(f_{priv}, \mathcal{D}) \leq J(f^*, \mathcal{D}) + \frac{2d^2 \left(\frac{c}{\Lambda} + \eta\right) \log^2 \frac{d}{\delta}}{\Lambda n^2 \varepsilon^2} \right\} \geq 1 - \delta.$$

Lemma

Suppose $N(\cdot)$ is 1-strongly convex and globally differentiable, and ℓ is convex and differentiable with $|\ell'(z)| \leq 1$ for all z . Then, letting f_{priv} be the output of Algorithm 2, we have

$$\mathbb{P}_{\mathbf{b}} \left\{ J(f_{priv}, \mathcal{D}) \leq J(f^*, \mathcal{D}) + \frac{4d^2 \log^2 \frac{d}{\delta}}{\Lambda n^2 \varepsilon^2} \right\} \geq 1 - \delta.$$

- If $\frac{c}{\Lambda} + \eta > 2$, Algorithm 2 is better.
- why not use Algorithm 1 in Certified Removal?

Conclusion

- Differential privacy in learning
 - ▶ Hide “local” information → satisfying the privacy guarantee
 - ▶ Learn “global” information → satisfying the correctness guarantee
- Two goals are conflicting each other.