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

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
aaroth@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

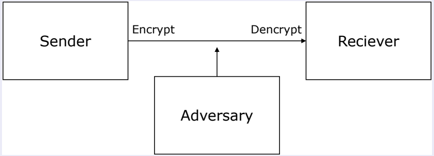
Name	Age	Gender	Zip Code	Smoker	Diagnosis
*	60–70	Male	191**	Y	Heart disease
*	60–70	Female	191**	N	Arthritis
*	60–70	Male	191**	Y	Lung cancer
*	60–70	Female	191**	N	Crohn's disease
*	60–70	Male	191**	Y	Lung cancer
*	***50–60***	***Female***	191**	N	HIV
*	50–60	Male	191**	Y	Lyme disease
*	50–60	Male	191**	Y	Seasonal allergies
*	***50–60***	***Female***	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?

Set-up for Encryption



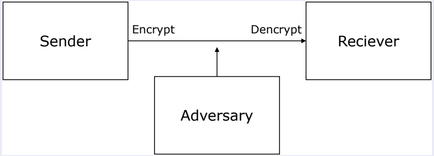
- Entities in encryption: Sender, Receiver, and Adversary

Set-up for Private Learning

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

Why Not Use Cryptosystems?

Set-up for Encryption



- Entities in encryption: Sender, Receiver, and Adversary

Set-up for Private Learning

- 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.

- How to achieve this goal? Add noise!

Goal for Privacy In Learning

Goal

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

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

Randomized Response

An Example

Goal of A Survey

Estimate a statistic on illegal behaviors of participants, where

- Curator: a participant
- Analyst: a researcher

Randomized Response

An Example

Goal of A Survey

Estimate a statistic on illegal behaviors of participants, where

- 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

Randomized Response

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

- $X_i \in \{0, 1\}$: 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?

Randomized Response

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

- How to estimate $p = \mathbb{E}_{\mathcal{P}}\{X_i\}$? – Let $X_i \sim \mathcal{P}$ and $\mathcal{Q}$ is a distribution over "unfair coin flips".

Randomized Response

How Informative?

Randomized Response

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

- How to estimate $p = \mathbb{E}_{\mathcal{P}}\{X_i\}$? – Let $X_i \sim \mathcal{P}$ and $\mathcal{Q}$ is a distribution over "unfair coin flips".
- Observe the following expectation over "unfair coin flips":

$$\mathbb{E}_{\mathcal{Q}}\{Y_i\} = X_i\left(\frac{1}{2} + \gamma\right) + (1 - X_i)\left(\frac{1}{2} - \gamma\right) = 2\gamma X_i + \frac{1}{2} - \gamma \quad \Longrightarrow \quad X_i = \mathbb{E}_{\mathcal{Q}}\left\{\frac{1}{2\gamma}\left(Y_i - \frac{1}{2} + \gamma\right)\right\}$$

Randomized Response

How Informative?

Randomized Response

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

- How to estimate $p = \mathbb{E}_{\mathcal{P}}\{X_i\}$? – Let $X_i \sim \mathcal{P}$ and $\mathcal{Q}$ is a distribution over "unfair coin flips".
- Observe the following expectation over "unfair coin flips":

$$\mathbb{E}_{\mathcal{Q}}\{Y_i\} = X_i\left(\frac{1}{2}+\gamma\right) + (1-X_i)\left(\frac{1}{2}-\gamma\right) = 2\gamma X_i + \frac{1}{2} - \gamma \quad \Longrightarrow \quad X_i = \mathbb{E}_{\mathcal{Q}}\left\{\frac{1}{2\gamma}\left(Y_i - \frac{1}{2} + \gamma\right)\right\}$$

- 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?

Randomized Response

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

- How to estimate $p = \mathbb{E}_{\mathcal{P}}\{X_i\}$? – Let $X_i \sim \mathcal{P}$ and $\mathcal{Q}$ is a distribution over "unfair coin flips".
- Observe the following expectation over "unfair coin flips":

$$\mathbb{E}_{\mathcal{Q}}\{Y_i\} = X_i \left(\frac{1}{2} + \gamma\right) + (1 - X_i)\left(\frac{1}{2} - \gamma\right) = 2\gamma X_i + \frac{1}{2} - \gamma \quad \Longrightarrow \quad X_i = \mathbb{E}_{\mathcal{Q}}\left\{\frac{1}{2\gamma}\left(Y_i - \frac{1}{2} + \gamma\right)\right\}$$

- Consider the following estimator:

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

- Unbiased? We have

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

Randomized Response

How Informative?

Randomized Response

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

- How to estimate $p = \mathbb{E}_{\mathcal{P}}\{X_i\}$? – Let $X_i \sim \mathcal{P}$ and $\mathcal{Q}$ is a distribution over "unfair coin flips".
- Observe the following expectation over "unfair coin flips":

$$\mathbb{E}_{\mathcal{Q}}\{Y_i\} = X_i\left(\frac{1}{2} + \gamma\right) + (1 - X_i)\left(\frac{1}{2} - \gamma\right) = 2\gamma X_i + \frac{1}{2} - \gamma \quad \Longrightarrow \quad X_i = \mathbb{E}_{\mathcal{Q}}\left\{\frac{1}{2\gamma}\left(Y_i - \frac{1}{2} + \gamma\right)\right\}$$

- Consider the following estimator:

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

- Unbiased? We have

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

- The randomized response looks "working"! How can we prove that this "algorithm" does not leak privacy?

A Goodness Metric in Differential Privacy (DP)

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 $\mathcal{D}_2$,

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

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

- Consider the following special case (*i.e.*, $\delta = 0$ and $\varepsilon \to 0$):

$$1 \approx \frac{1}{\exp(\varepsilon)} \leq \frac{\mathbb{P}\left\{\mathcal{M}(\mathcal{D}_1) \in \mathcal{S}\right\}}{\mathbb{P}\left\{\mathcal{M}(\mathcal{D}_2) \in \mathcal{S}\right\}} \leq \exp(\varepsilon) \approx 1$$

  - After applying differentially private $\mathcal{M}$, *i.e.*, $\mathcal{S} = \mathcal{M}(\mathcal{D}_1)$, an attacker cannot tell whether $\mathcal{S}$ is from $\mathcal{D}_1$ or $\mathcal{D}_2$ so cannot extract information from the difference between $\mathcal{D}_1$ and $\mathcal{D}_2$.
  - *e.g.*, $\mathcal{D}_1 = \{X_1 = 0, X_2 = 1\}$, $\mathcal{D}_2 = \{X_1 = 0\}$, $\mathcal{S} = \{1\}$, $\mathcal{M} =$ "contain 1?"

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
  - $\mathcal{M}(X_1, \ldots, X_n) = (Y_1, \ldots, Y_n)$
- Let $\gamma = \frac{1}{4}$
- Consider any realization $\mathcal{S} \in \{0,1\}^n$ of $(Y_1, \ldots, Y_n)$.
- Consider $X := (X_1, \ldots, X_n)$ and $X' := (X'_1, \ldots, X'_n)$ which differ only in coordinate j.
- Then, we have

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

  - Note that S_j is fixed but the left-hand side of the inequality maximizes if $\mathcal{S}_j = 1$.

Laplace Mechanism

Definition

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

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

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

- Lap$(x|b)$ = Lap$(b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$
- The ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$ 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 $(\varepsilon, 0)$*-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 $\mathcal{M}_L(x, f, \varepsilon)$, *i.e.*, $p_x(z) \coloneqq \mathbb{P}\{\mathcal{M}_L(x, f, \varepsilon) = z\}$
- p_y: the PDF of $\mathcal{M}_L(y, f, \varepsilon)$
- For any $z \in \mathbb{R}^k$, we have

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

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

Gaussian Mechanism is DP

Definition

Let $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^d$ 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 Gaussan 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 KCHAUDHURI@UCSD.EDU
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093, USA

Claire Monteleoni CMONTEL@CCLS.COLUMBIA.EDU
Center for Computational Learning Systems
Columbia University
New York, NY 10115, USA

Anand D. Sarwate ASARWATE@UCSD.EDU
Information Theory and Applications Center
University of California, San Diego
La Jolla, CA 92093-0447, USA

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
- $f : \mathcal{X} \to \mathcal{Y}$: a predictor
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$: a loss function
- Regularized empirical risk minimization (ERM):

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

where $N(f)$ is a regularizer.

Assumptions

Definition

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

$$f\left(\alpha x + (1-\alpha)y\right) < \alpha f(x) + (1-\alpha) f(y).$$

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

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

- A strictly convex function has a unique minimum.
- (strongly convex) $\Longrightarrow$ (strictly convex)
- The regularizer $N(\cdot)$ and loss $\ell(\cdot,\cdot)$ are differentiable.
  - No ℓ_1-norm regularizer
  - No hinge loss
- Note that these assumptions are for handy analyses (and could be relaxed).

Privacy Model

Goal: Learn a classifier which preserves the privacy of individual entities of a 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 $\mathcal{S}$

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

- $\mathcal{A}(\mathcal{D})$: a randomized algorithm that returns a classifier from a training set $\mathcal{D}$.
- $\mathcal{D}'$ and $\mathcal{D}$ have $n-1$ samples (x_i, y_i) in common; the different 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 $\mathcal{S} := \{f^*_{\mathcal{D}}\}$,

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

  - Note that our ERM is deterministic.
- Thus, ERM is not differentially private!

Algorithm 1: Output Perturbation

Output Perturbation

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

- $\mathbf{b}$ is random noise with density
$$v(\mathbf{b}) \propto e^{-\beta\|\mathbf{b}\|}$$
with $\beta = \frac{n\Lambda\varepsilon}{2}$.
- This algorithm is randomized.

Algorithm 2: Objective Perturbation

Objective Perturbation

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

- $\mathbf{b}$ 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 $\beta > 0$).
- This algorithm is randomized.

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. Given $\mathcal{D}$, let $f^* := \arg\max_f J(\mathcal{D}, f)$ let f_{priv} be the output of Algorithm 1. Then, we have

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

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. Given $\mathcal{D}$, let $f^* := \arg\max_f J(\mathcal{D}, f)$ and let f_{priv} be the output of Algorithm 2. Then, we have

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

- If $\frac{c}{\Lambda} + \eta > 2$, Algorithm 2 is better.
- Intuition: if perturbations are considered in learning, the algorithm finds a better classifier.

Conclusion

- Differential privacy in learning:
    - Hide “local” information $\rightarrow$ satisfying the privacy guarantee
    - Learn “global” information $\rightarrow$ satisfying the correctness guarantee
- Two goals are conflicting each other and balancing two is critical.