Differential Privacy without Sensitivity

Kentaro Minami	(The University of Tokyo)
Hiromi Arai	(RIKEN AIP)
Issei Sato	(The University of Tokyo)
Hiroshi Nakagawa	(The University of Tokyo)

Overview

- Differential Privacy (DP)
  - A privacy protection measure [Dwork+ 06]
- Differentially Private Learning
  - Statistical learning under differential privacy constraints
  - **Requires some randomization of estimators**
- Gibbs Posterior
  - A natural class of distributions of randomized estimators

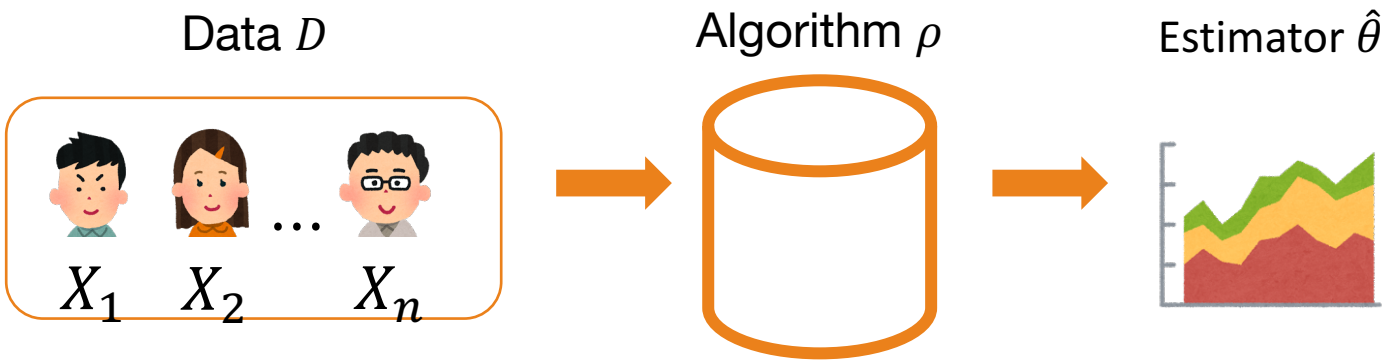
Contribution

- A new proof for (ε, δ)-DP (tbd.) of Gibbs posteriors
- Applicable for **Lipschitz convex losses & log-concave priors**

Differential Privacy [Dwork+ 06]

Adversarial Formulation of Privacy:

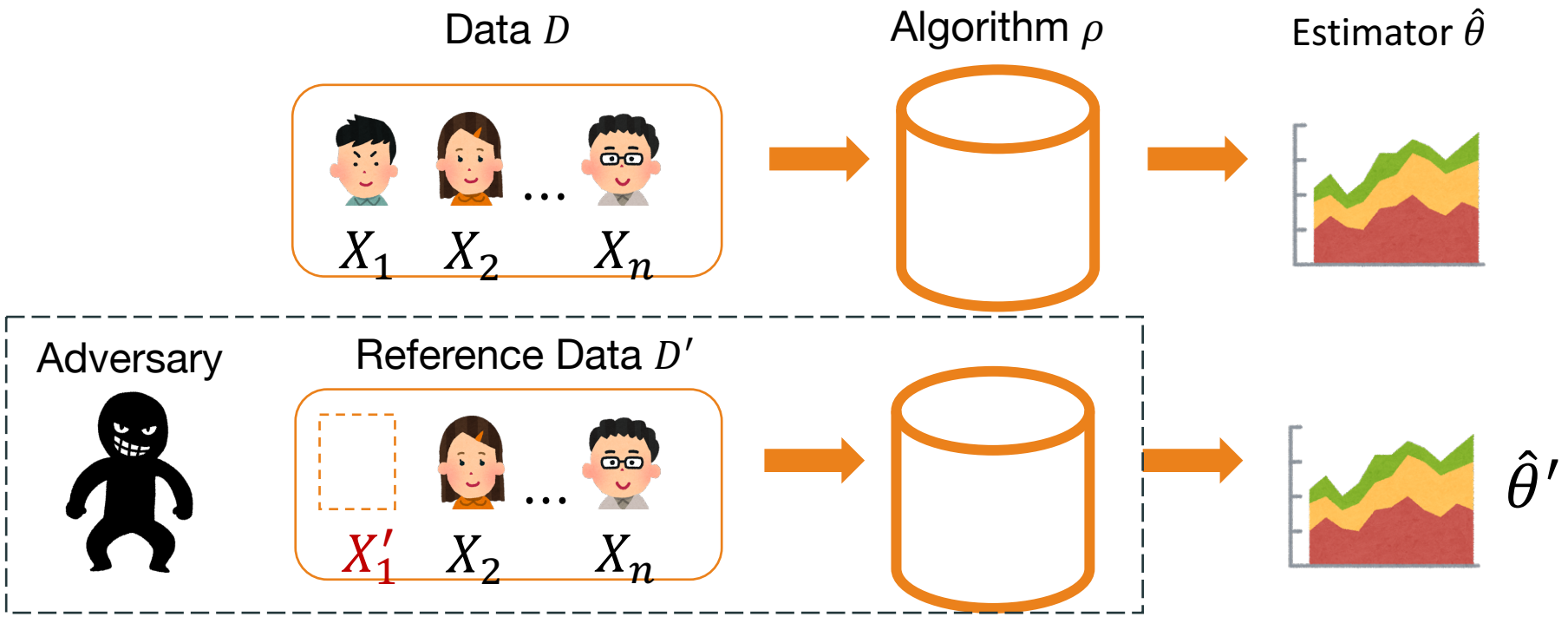
1. Algorithm ρ releases an estimator using private data $\mathcal{D}$



Differential Privacy [Dwork+ 06]

Adversarial Formulation of Privacy:

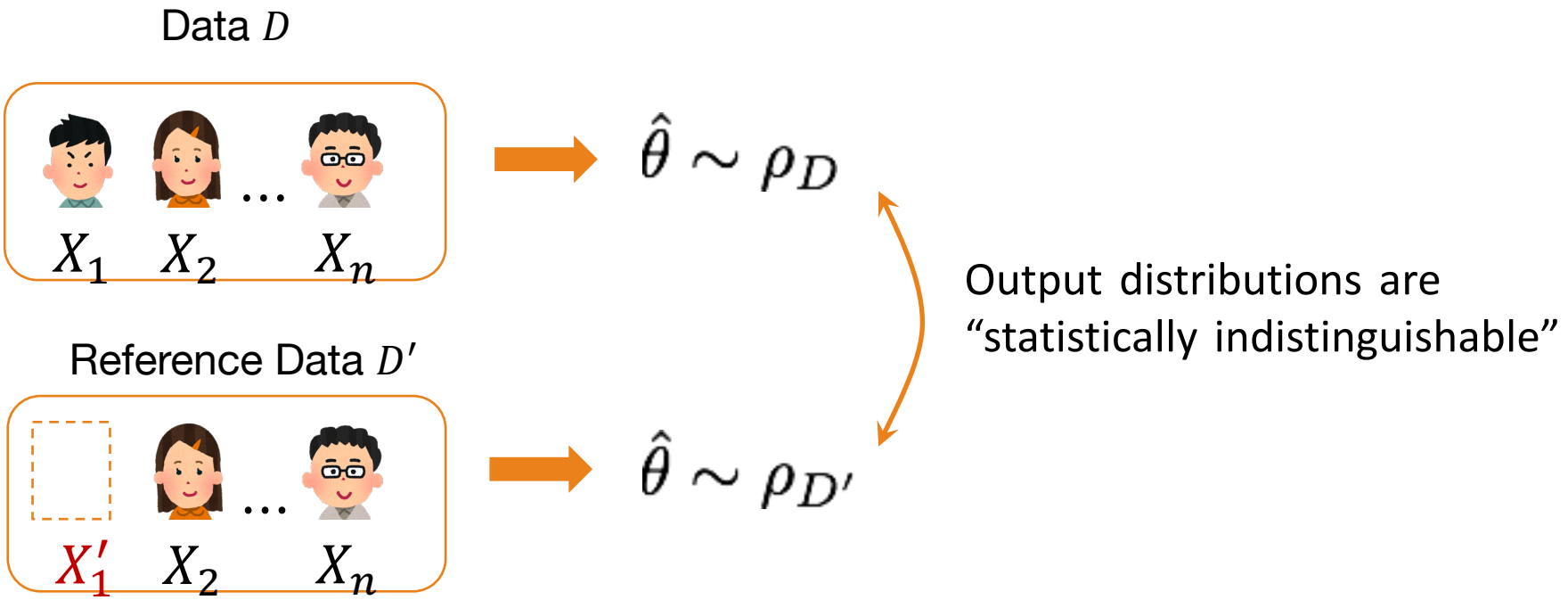
2. Adversary $\mathcal{A}$ has a reference data $\mathcal{D}'$ that differs from $\mathcal{D}$ in a single individual (say X_1).
   Adversary $\mathcal{A}$ also knows how Algorithm ρ works.



Differential Privacy [Dwork+ 06]

We want to make $\hat{\theta}$ and $\hat{\theta}'$ indistinguishable for any reference data D'.

Idea: Release a sufficiently randomized version of $\hat{\theta}$



Differential Privacy [Dwork+ 06]

Definition:

- $\varepsilon > 0, 0 \leq \delta < 1$ privacy parameters

- Output distribution ρ_D satisfies (ε, δ)-differential privacy if
  1. For any pair of adjacent datasets (D, D'), and
  2. For any measurable set $A \subset \Theta$ in the parameter space,

$$\Pr_{\theta \sim \rho_D}[\theta \in A] \leq e^{\varepsilon} \Pr_{\theta \sim \rho_{D'}}[\theta \in A] + \delta$$

Differentially Private Learning

(Classical) statistical learning:

- Minimize the empirical risk $L(\theta, D) = \sum_{x \in D} \ell(\theta, x)$ or its expectation

Differentially private learning:

- Outputs of DP algorithms are random.
- Minimize the "expected" risk under a DP constraint:

$$\begin{aligned} \min_{\rho} \quad & \mathbb{E}_{D \sim P^n} \mathbb{E}_{\theta \sim \rho_D} [L(\theta, D)] \\ \text{s.t.} \quad & \rho_D \text{ satisfies } (\varepsilon, \delta)\text{-DP} \end{aligned}$$

Gibbs posterior

In statistics and ML, a typical data-dependent distribution is the Bayesian posterior.

$$\frac{\prod_{i=1}^{n} p(x_i \mid \theta)\pi(\theta)}{\int \prod_{i=1}^{n} p(x_i \mid \theta)\pi(\theta)\mathrm{d}\theta}$$

Gibbs posterior

In statistics and ML, a typical data-dependent distribution is the Bayesian posterior.

$$\frac{\prod_{i=1}^{n} p(x_i \mid \theta)\pi(\theta)}{\int \prod_{i=1}^{n} p(x_i \mid \theta)\pi(\theta)\mathrm{d}\theta}$$

Generalization: Introduce a scale parameter $\beta > 0$

$$\frac{\prod_{i=1}^{n} p(x_i \mid \theta)^{\beta}\pi(\theta)}{\int \prod_{i=1}^{n} p(x_i \mid \theta)^{\beta}\pi(\theta)\mathrm{d}\theta}$$

Gibbs posterior

We consider a class of posterior distribution with a scale.

Definition:

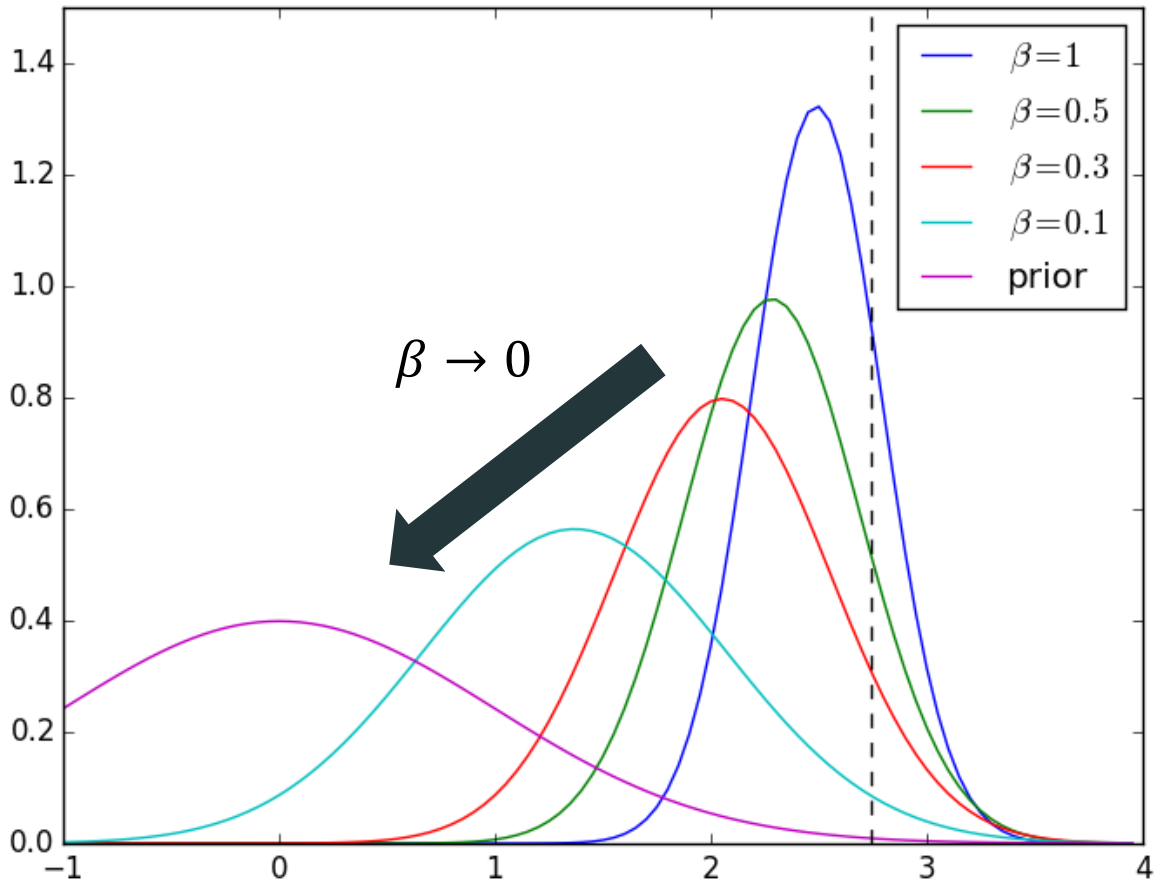
- $\ell: \Theta \times \mathcal{X} \to \mathbb{R}$ Loss function
- $\pi(\theta)$ Prior distribution on Θ
- $\beta > 0$ Scale parameter

We define the density of Gibbs posterior $G_{\beta,D}$ as

$$G_\beta(\theta \mid D) = \frac{\exp(-\beta \sum_{i=1}^n \ell(\theta, x_i))\pi(\theta)}{\int_\Theta \exp(-\beta \sum_{i=1}^n \ell(\theta, x_i))\pi(\theta)\mathrm{d}\theta}$$

Gibbs posterior

If $\beta \to 0$, the Gibbs posterior becomes flat, and converges to the prior.



A Method for $(\varepsilon, 0)$-DP

For any function $\mathcal{L}$: $\Theta \times \mathcal{X}^n \to \mathbb{R}$, define the sensitivity as

$$\Delta_{\mathcal{L}} := \sup_{\substack{D, D': \\ \text{adjacent}}} \sup_{\theta \in \Theta} |\mathcal{L}(\theta, D) - \mathcal{L}(\theta, D')|$$

Exponential mechanism [McSherry & Talwar 07]

If $\Delta_{\mathcal{L}} < \infty$, an algorithm that draws θ from a distribution

$$\frac{\exp(-\beta \mathcal{L}(\theta, D))\pi(\theta)}{\int_{\Theta} \exp(-\beta \mathcal{L}(\theta, D))\pi(\theta) \mathrm{d}\theta}$$

satisfies $(\varepsilon, 0)$-DP, where $\beta > 0$ is taken as

$$\beta \leq \frac{\varepsilon}{2\Delta_{\mathcal{L}}}$$

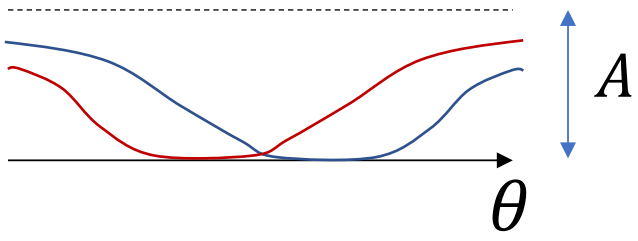
A Method for $(\varepsilon, 0)$-DP

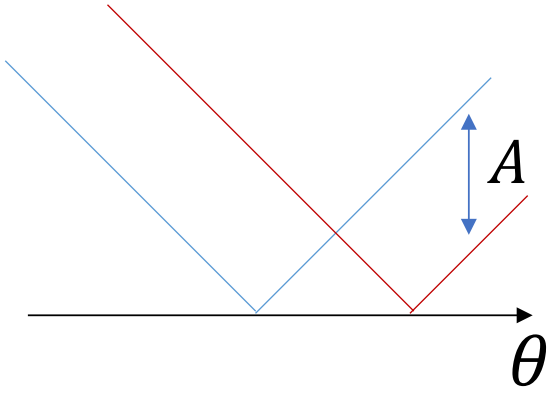
The Gibbs posterior can be seen as the exponential mechanism if it has a bounded sensitivity.

Question: When is the sensitivity bounded?

Theorem [Wang+ 15]

The Gibbs posterior can satisfy $(\varepsilon, 0)$-DP if:
(A) the loss function ℓ is bounded for all $x \in \mathcal{X}$, or
(B) the difference $\ell(\cdot, x) - \ell(\cdot, x')$ is bounded for all $x, x' \in \mathcal{X}$





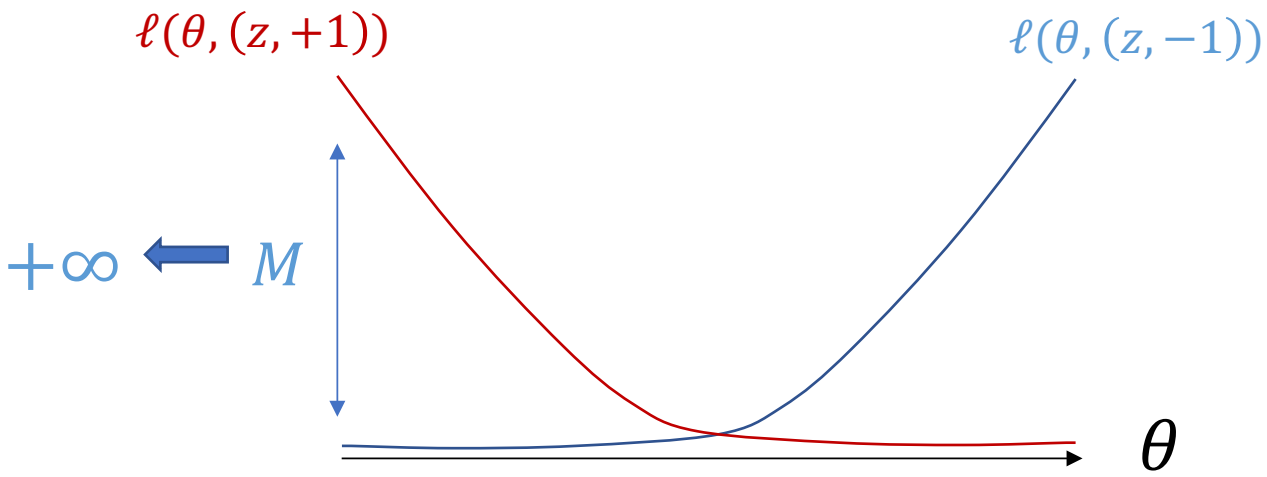
Unbounded Loss Differences

- For $(\varepsilon, 0)$-DP, the loss functions need to be bounded.
- However, many loss functions in statistical learning theory is unbounded.

Example: The logistic regression loss

$$\ell(\theta, (z, y)) = \log(1 + \exp(-y\langle\theta, z\rangle))$$

has unbounded sensitivity.



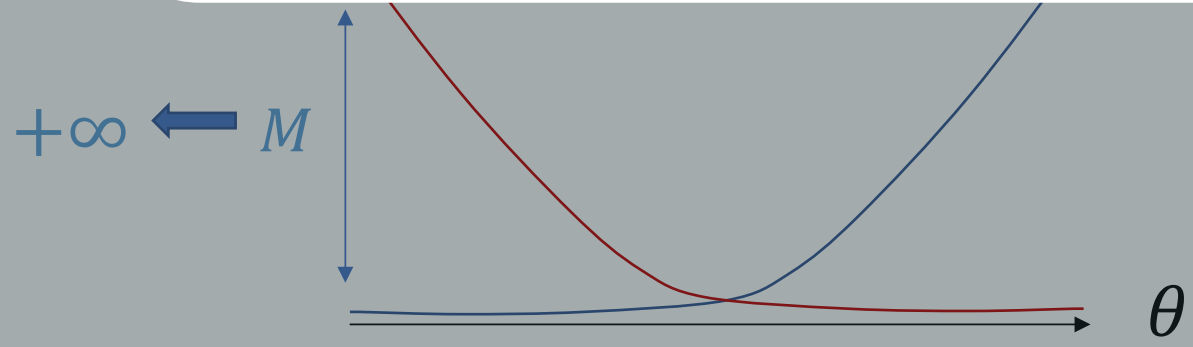
Unbounded Loss Differences

- For $(\varepsilon, 0)$-DP, the loss functions need to be bounded.
- However, many loss functions in statistical learning theory is unbo

Exampl

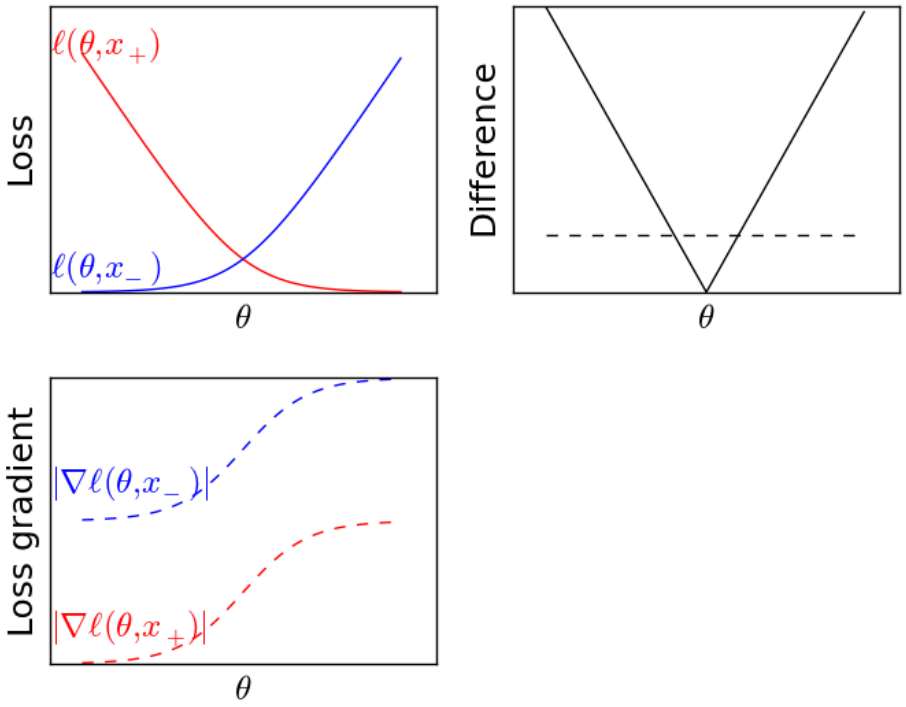
has unb

Our Goal:
Prove DP of the Gibbs posterior when the loss is unbounded!



Key Idea: From Bounded to Lipschitz

- In the example of logistic loss, the first derivative is bounded
- The Lipschitz constant L is not influenced by the size of parameter space



Main Theorem

Theorem

1. For all $x \in \mathcal{X}$, $\ell(\cdot, x)$ is L-Lipschitz and convex
2. $-\log \pi(\cdot)$ is m_π-strongly convex
3. $\Theta = \mathbb{R}^d$

The Gibbs posterior $G_{\beta,D}$ satisfies (ε, δ)-DP if

$$\beta \leq \frac{\varepsilon}{2L} \sqrt{\frac{m_\pi}{1 + 2\log(1/\delta)}}$$

The upper bound does not require boundedness of ℓ!

Lipschitz & Convex Losses

1. Classification
   - Logistic regression
   - Hinge loss (SVM)

2. Regression
   - Quantile regression
   - Huber loss for robust regression

3. Parametric statistics
   - Negative log-likelihood of binomial / multinomial distributions w.r.t. the natural parametrization of the exponential family

Utility Analysis

A DP guarantee for Gibbs posteriors implies a trade-off between privacy and learning efficiency.

Examples:

1. Statistical efficiency
2. PAC-Bayesian bound

Statistical Efficiency

- In parametric statistics, $\ell(\theta, x) = -\log p(x \mid \theta)$
- The Gibbs posterior is β^{-1} times less efficient than the Bayesian posterior.

Proposition (Misspecified Bernstein—von Mises)

Under some regularity conditions, $G_{\beta,D}$ converges to a normal distribution with covariance matrix $(n\beta)^{-1} I_{\theta_0}^{-1}$:

$$\sup_{B \subset \Theta} \left| G_{\beta, D_n} \left\{ \sqrt{n}(\theta - \theta_0) \in B \right\} - N_{0, \beta^{-1} I_{\theta_0}^{-1}}(B) \right| \xrightarrow{p} 0$$

PAC-Bayesian Bounds

- The Gibbs posterior has a fundamental role in PAC-Bayes theory [McAllester 98] [Catoni 07] [Alquier 15]

- Tightness of the empirical bound is controlled by β^{-1}.

Proposition (PAC-Bayesian bound) [Alquier 15]

Under some regularity conditions, $G_{\beta,D}$ satisfies the following risk bound

$$\mathbb{E}_{\theta\sim G_{\beta,D}}[R(\theta)] \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)} \mathbb{E}_{\theta\sim\rho}[R(\theta)] + 2\frac{f(\beta,n) + D_{\mathrm{KL}}(\rho,\pi) + \log(\eta^{-1})}{\beta}$$

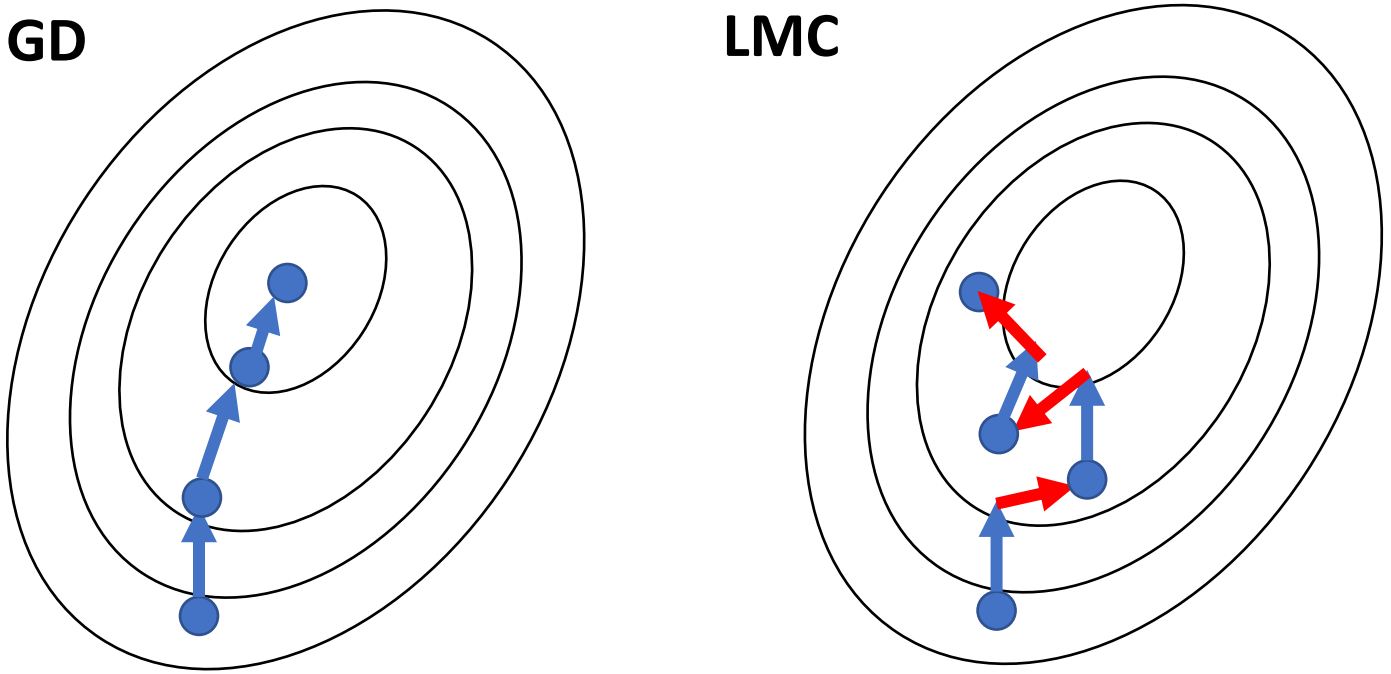
with probability at least $1-2\eta$.

Approximation of Posteriors

In practice, sampling from the Gibbs posterior can be a computationally hard problem.

Langevin Monte Carlo (LMC)

A variant of MCMC like a “noisy gradient descent”.



Privacy-Preserving Approximate Posterior (PPAP)

If $G_{\beta,D}$ satisfies (ε, δ)-DP, we can also prove (ε, δ')-DP of approximate Gibbs posterior by LMC.

Proposition:

- Assume that ℓ and π satisfy the assumption in Main Theorem. We also assume that $\ell(\cdot, x)$ is M-smooth for every $x \in \mathcal{X}$
- After $O\left(\left(\frac{n}{\gamma}\right)^2 \left(\log\frac{n}{\gamma}\right)^2\right)$ iterations, the output of the LMC satisfies $(\varepsilon, \delta + (e^{\varepsilon} + 1)\gamma)$-DP.

Conclusion

- A new proof of (ε, δ)-DP for Gibbs posteriors
  - Removed the boundedness assumption
  - Lipschitz convex losses & (strongly) log-concave priors
- Properties
  - Easily connected to statistical efficiency / PAC-Bayesian bounds
  - Finite-time DP guarantee of Langevin Monte-Carlo

Thank you!