Differentially Private Empirical Risk Minimization with Non-convex Loss Functions

Di Wang, Changyou Chen and Jinhui Xu

State University of New York at Buffalo

International Conference on Machine Learning 2019

Outline

Outline

Empirical Risk Minimization (ERM)

- **Given:** A dataset $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, where each $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} \sim \mathcal{P}$.
- **Regularization** $r(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$, we use ℓ_2 regularization with $r(w) = \frac{\lambda}{2}\|w\|_2^2$.
- For a loss function ℓ, the (regularized) Empirical Risk:

$$\hat{L}^r(w; D) = \frac{1}{n}\sum_{i=1}^{n} \ell(w; x_i, y_i) + r(w).$$

the (regularized) Population Risk:

$$L^r_{\mathcal{P}}(w) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(w; x, y)] + r(w).$$

Goal: Find w so as to minimize the empirical or population risk.

(ϵ, δ)- Differential Privacy (DP)

Differential Privacy (DP) [Dwork et al,. 2006]

We say that two datasets, D and D', are neighbors if they differ by only one entry, denoted as $D \sim D'$.

A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for all neighboring datasets D, D', and for all events S in the output space of $\mathcal{A}$, we have

$$\text{Pr}(\mathcal{A}(D) \in S) \leq e^{\epsilon} \text{Pr}\mathcal{A}(D') \in S) + \delta.$$

DP-ERM

DP-ERM

Determine a sample complexity $n = n(1/\epsilon, 1/\delta, p, 1/\alpha)$ such that there is an (ϵ, δ)-DP algorithm whose output w^{priv} achieves an α-error in the **expected excess empirical risk**:

$$\text{Err}_D^r(w^{\text{priv}}) = \mathbb{E}\hat{L}(w^{\text{LDP}}; D) - \min_{w \in \mathbb{R}^d} \hat{L}(w; D) \leq \alpha.$$

or in the **expected excess empirical risk**:

$$\text{Err}_{\mathcal{P}}^r(w^{\text{priv}}) = \mathbb{E}[L_{\mathcal{P}}^r(w^{\text{priv}})] - \min_{w \in \mathbb{R}^d} L_{\mathcal{P}}^r(w) \leq \alpha.$$

Motivation

- Previous work on DP-ERM mainly focuses on **convex loss** functions.

Motivation

- Previous work on DP-ERM mainly focuses on **convex loss** functions.
- For non-convex loss functions, [Zhang et al, 2017] and [Wang and Xu 2019] studied the problem and used, as error measurement, the ℓ_2 gradient norm of a private estimator, *i.e.*,

$$\|\nabla \hat{L}_D^r(w^{\text{priv}})\|_2 \text{ and } \mathbb{E}_{\mathcal{P}}\|\nabla \ell(w^{\text{priv}}; x, y)\|_2$$

Motivation

- Previous work on DP-ERM mainly focuses on **convex loss** functions.
- For non-convex loss functions, [Zhang et al, 2017] and [Wang and Xu 2019] studied the problem and used, as error measurement, the ℓ_2 gradient norm of a private estimator, *i.e.*,

$$\|\nabla \hat{L}^r_D(w^{\text{priv}})\|_2 \text{ and } \mathbb{E}_{\mathcal{P}}\|\nabla \ell(w^{\text{priv}}; x, y)\|_2$$

- **Main Question:** Can the excess empirical (population) risk be used to measure the error of non-convex loss functions in the differential privacy model?

Outline

Result 1

Theorem 1

If the loss function is L-Lipschitz, twice differentiable and M-smooth, by using the private version of Gradient Langevin Dynamics (DP-GLD) we show that the excess empirical (or population) risk is upper bounded by $\tilde{O}(\frac{d\log(1/\delta)}{\log n\epsilon^2})$.

Result 1

Theorem 1

If the loss function is L-Lipschitz, twice differentiable and M-smooth, by using the private version of Gradient Langevin Dynamics (DP-GLD) we show that the excess empirical (or population) risk is upper bounded by $\tilde{O}(\frac{d \log(1/\delta)}{\log n\epsilon^2})$.

- The proof is based on some recent developments in Bayesian learning and analysis of GLD. By using a finer analysis of the time-average error of some SDE, we show the following

Result 1

Theorem 1

If the loss function is L-Lipschitz, twice differentiable and M-smooth, by using the private version of Gradient Langevin Dynamics (DP-GLD) we show that the excess empirical (or population) risk is upper bounded by $\tilde{O}(\frac{d \log(1/\delta)}{\log n\epsilon^2})$.

- The proof is based on some recent developments in Bayesian learning and analysis of GLD. By using a finer analysis of the time-average error of some SDE, we show the following

Theorem 2

For the excessed empirical risk, there is an (ϵ, δ)-DP algorithm which satisfies

$$\lim_{T \to \infty} \text{Err}_D^r(w_T) \leq \tilde{O}\left(\frac{C_0(d) \log(1/\delta)}{n^\tau \epsilon^\tau}\right),$$

where $C_0(d)$ is a function of d and $0 < \tau < 1$ is some cosntant.

Outline

Result 2

- **Are these bounds tight?**

Result 2

- **Are these bounds tight?**
  Based on the exponential mechanism, we have

Empirical Risk

For any $\beta < 1$, there is an ϵ-differentially private algorithm whose output w^{priv} induces an excess empirical risk $\text{Err}_D^r(w^{\text{priv}}) \leq \tilde{O}(\frac{d}{n\epsilon})$ with probability at least $1 - \beta$.

Result 2

- **Are these bounds tight?**
  Based on the exponential mechanism, we have

Empirical Risk

For any $\beta < 1$, there is an ϵ-differentially private algorithm whose output w^{priv} induces an excess empirical risk $\text{Err}_D^r(w^{\text{priv}}) \leq \tilde{O}(\frac{d}{n\epsilon})$ with probability at least $1 - \beta$.

Population Risk

For **Generalized Linear model** and **Robust Regressions** (whose loss function is $\ell(w; x, y) = (\sigma(\langle w, x \rangle) - y)^2$ and $\ell(w; x, y) = \Phi(\langle w, x \rangle - y)$, respectively), under some reasonable assumptions, there is an (ϵ, δ)-DP algorithm whose excess population risk is upper bounded by

$$\text{Err}_{\mathcal{P}}(w^{\text{priv}}) \leq O(\frac{\sqrt[4]{d \ln \frac{1}{\delta}}}{\sqrt{n\epsilon}}).$$

Outline

1. Introduction
   - Problem Description
   - Result 1
   - Result 2
   - Result 3

Finding Approximate Local Minimum Privately

- **Finding global minimum of non-convex function is challenging!**

Finding Approximate Local Minimum Privately

- **Finding global minimum of non-convex function is challenging!**
- Recent research on Deep Learning and other non-convex problems show that **local minima**, but not critical points, are sufficient.

Finding Approximate Local Minimum Privately

- **Finding global minimum of non-convex function is challenging!**
- Recent research on Deep Learning and other non-convex problems show that **local minima**, but not critical points, are sufficient.
- **But**, finding local minima is still NP-hard.

Finding Approximate Local Minimum Privately

- **Finding global minimum of non-convex function is challenging!**
- Recent research on Deep Learning and other non-convex problems show that **local minima**, but not critical points, are sufficient.
- **But**, finding local minima is still NP-hard.
- Fortunately, many non-convex functions are strict saddle. Thus, it is sufficient to find the **second order stationary point** (or approximate local minimum).

Definition

w is an α-second-order stationary point (α-SOSP), if

$$\|\nabla F(w)\|_2 \leq \alpha \text{ and } \lambda_{\min}(\nabla^2 F(w)) \geq -\sqrt{\rho\alpha}. \tag{1}$$

Finding Approximate Local Minimum Privately

- **Finding global minimum of non-convex function is challenging!**
- Recent research on Deep Learning and other non-convex problems show that **local minima**, but not critical points, are sufficient.
- **But**, finding local minima is still NP-hard.
- Fortunately, many non-convex functions are strict saddle. Thus, it is sufficient to find the **second order stationary point** (or approximate local minimum).

Definition

w is an α-second-order stationary point (α-SOSP), if

$$\|\nabla F(w)\|_2 \leq \alpha \text{ and } \lambda_{\min}(\nabla^2 F(w)) \geq -\sqrt{\rho\alpha}. \tag{1}$$

- **Can we find some approximate local minimum which escapes saddle points and still keeps the algorithm (ϵ, δ)-differentially private?**

Result 3

- On one hand, (Ge et al. 2015) proposed an algorithm, **noisy Stochastic Gradient Descent**, to find approximate local minima.

Result 3

- On one hand, (Ge et al. 2015) proposed an algorithm, **noisy Stochastic Gradient Descent**, to find approximate local minima.
- On the other hand, in DP community, one popular method for ERM is called **DP-SGD**, which adds some Gaussian noise in each iteration.

Result 3

- On one hand, (Ge et al. 2015) proposed an algorithm, **noisy Stochastic Gradient Descent**, to find approximate local minima.
- On the other hand, in DP community, one popular method for ERM is called **DP-SGD**, which adds some Gaussian noise in each iteration.
- Using DP-GD, we can show

Theorem 4

If the data size n is large enough such that

$$n \geq \tilde{\Omega}\left(\frac{\sqrt{\log \frac{1}{\delta} d} \log \frac{1}{\xi}}{\epsilon \alpha^2}\right), \tag{2}$$

then with probability $1 - \zeta$, one of the outputs is an α-SOSP of the empirical risk $\hat{L}(\cdot, D)$.

Thank you!