Minimisation privée du risque empirique par descente par coordonnées

Paul Mangold[1], Aurélien Bellet[1], Joseph Salmon[2], Marc Tommasi[1]

[1]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille,
[2]IMAG, Univ. Montpellier, CNRS Montpellier, France

CAp 2021

June 15th, 2021

1 Introduction

2 Background: Private ERM

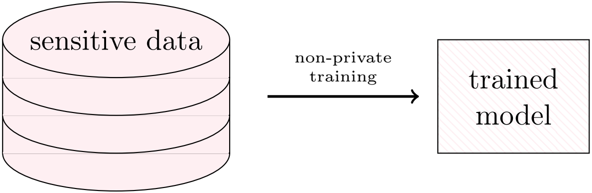
3 Our Algorithm: Private Coordinate Descent

4 Experiments: Linear Regression

5 Conclusion and Perspectives

Machine Learning Uses Data

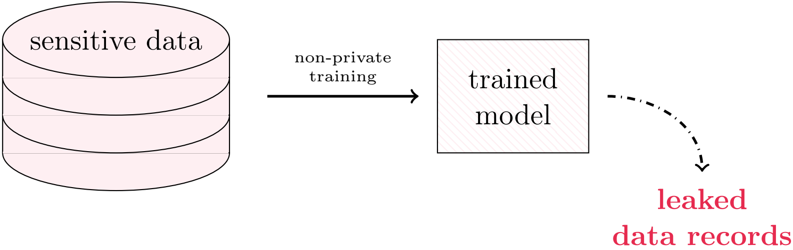
- ML models are trained on **sensitive data**.
- In classical training procedures:



– R. Shokri et al., *"Membership Inference Attacks against Machine Learning Models"*, 2017.

Machine Learning Uses Data

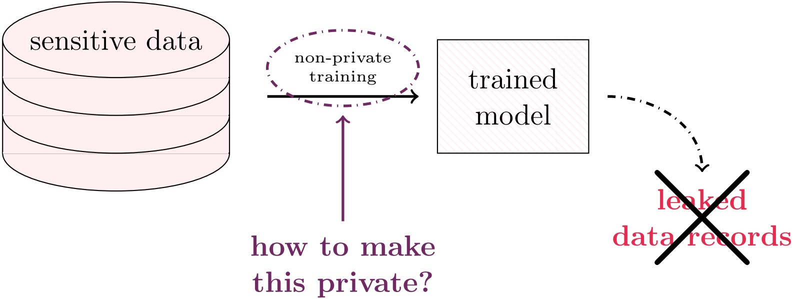
- ML models are trained on **sensitive data**.
- In classical training procedures:



– R. Shokri et al., *"Membership Inference Attacks against Machine Learning Models"*, 2017.

Machine Learning Uses Data

- ML models are trained on **sensitive data**.
- In classical training procedures:



– R. Shokri et al., *"Membership Inference Attacks against Machine Learning Models"*, 2017.

What Private **Formally** Means?

Definition (Differential Privacy)

An algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{M}$ is **(ϵ, δ)-differentially private** if for all $S \subseteq \mathcal{M}$ and for all $D, D' \in \mathcal{D}$ that **differ on at most one element**

$$P(\mathcal{A}(D) \in S) \leq \exp(\epsilon) P(\mathcal{A}(D') \in S) + \delta, \tag{1}$$

where the probability is taken over the coin flips of $\mathcal{A}$.

– C. Dwork, *"Differential Privacy"*, 2006.

1 Introduction

2 Background: Private ERM

3 Our Algorithm: Private Coordinate Descent

4 Experiments: Linear Regression

5 Conclusion and Perspectives

Background: Private ERM

Private Empirical Risk Minimization

Let

- $d_1, \ldots, d_n \in \mathcal{X} \times \mathcal{Y}$: data points.
- $h_w : \mathcal{X} \to \mathcal{Y}$: hypothesis function parameterized by $w \in \mathbb{R}^p$.
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$: loss function.

Goal: find a **(ϵ, δ)-DP approximation** of

$$w^* = \arg\min_{w \in \mathbb{R}^p} \left\{ f(w) := \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(x_i); y_i) \right\}.$$

– K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, *"Differentially Private Empirical Risk Minimization"*, 2011.

DP-SGD for DP-ERM: The Algorithm

When f is **convex**: DP-SGD works.

Algorithm DP-SGD (essentially).

Input: noise scale $\sigma > 0$; initial point $w^0 \in \mathbb{R}^p$; $T > 0$; data d.

1: **for** $t = 0, \ldots, T-1$ **do**

2: $w^{t+1} = w^t - \eta_t(g^t + \boldsymbol{b^t})$ with $\begin{cases} \mathbb{E}[g^t] = \nabla f(w^t; d), \\ \boldsymbol{b^t \sim \mathcal{N}(0, \sigma^2)}. \end{cases}$

3: **return** $w^{priv} = w^T$.

(and it works faster when f is **smooth**.)

– R. Bassily, A. Smith, and A. Thakurta, *"Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds"*, 2014.

Background: Private ERM

DP-SGD for DP-ERM: Calibrating noise

- Gradient **sensitivity**: for all d, d':

$$\left\| \nabla \ell(\cdot, d) - \nabla \ell(\cdot, d') \right\|_2 \leq \Delta_2(\nabla \ell).$$

DP-SGD for DP-ERM: Calibrating noise

- Gradient **sensitivity**: for all d, d':

$$\left\|\nabla \ell(\cdot, d) - \nabla \ell(\cdot, d')\right\|_2 \leq \Delta_2(\nabla \ell).$$

Theorem (Privacy Guarantees)

For $T > 0$, $\sigma^2 = \frac{8\Delta_2(\nabla \ell)^2 T \log(1/\delta)}{n^2 \epsilon^2}$.

DP-SGD is (ϵ, δ)*-differentially-private.*

- R. Bassily, A. Smith, and A. Thakurta, *"Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds"*, 2014.
- D. Wang, M. Ye, and J. Xu, *"Differentially Private Empirical Risk Minimization Revisited: Faster and More General"*, 2018.

DP-SGD for DP-ERM: Calibrating noise

In practice, $\Delta_2(\nabla\ell)$ can be **big** or even **unknown**: clip it!

$$\mathrm{clip}(\nabla\ell, C) = \begin{cases} \nabla\ell(w) & \text{if } \|\nabla\ell(w)\| \leq C, \\ \frac{C}{\|\nabla\ell(w)\|_2}\nabla\ell(w) & \text{otherwise.} \end{cases}$$

Consequently: $\boldsymbol{\Delta_2(\nabla\ell) \leq 2C}$.

– M. Abadi et al., *"Deep Learning with Differential Privacy"*, 2016.

DP-SGD for DP-ERM: Convergence?

Measure utility as $\mathbb{E}[f(w_{priv}) - f(w^*)]$, for which we know:

- A **lower bound**: it can not be arbitrarily small.
- An **upper bound**: DP-SGD is (nearly) optimal.

– R. Bassily, A. Smith, and A. Thakurta, *"Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds"*, 2014.

If DP-SGD is optimal, why look further?

Drawbacks of DP-SGD

If DP-SGD is optimal, why look further?

Well, in DP-SGD:

- **Unique** learning rate for all coordinates.
- **Global** sensitivity.

Drawbacks of DP-SGD

If DP-SGD is optimal, why look further?

Well, in DP-SGD:

- **Unique** learning rate for all coordinates.
- **Global** sensitivity.

→ We hope for better utility with **coordinate methods**.

1 Introduction

2 Background: Private ERM

3 Our Algorithm: Private Coordinate Descent

4 Experiments: Linear Regression

5 Conclusion and Perspectives

Algorithm DP-CD.

Input: noise scales $\sigma_1, \dots, \sigma_p > 0$; learning rates $\eta_1, \dots, \eta_p > 0$; initial point $\bar{w}^0 = w^0 \in \mathbb{R}^p$; $T, K > 0$

for $t = 0, \dots, T-1$ **do**

Set $\theta^0 = \bar{w}^t$

for $k = 0, \dots, K-1$ **do**

Pick j from $\{1, \dots, p\}$ uniformly at random and update:

$$\theta^{k+1} = \begin{cases} \theta^k_{j'} & \text{for } j' \neq j, \\ \theta^k_j - \eta_j(\nabla_j f(\theta^k) + \boldsymbol{b^t}) & \text{with } \boldsymbol{b_j \sim \mathcal{N}(0, \sigma_j^2)} \end{cases}$$

Average $\bar{w}_{t+1} = \frac{1}{K} \sum_{k=1}^{K} \theta^k$.

return $w_{priv} = \bar{w}_T$

Our Algorithm: Private Coordinate Descent

More queries, lower sensitivity

- Coordinate gradient **sensitivity**: for all d, d' and j,

$$\left|\nabla_j \ell(\cdot, d) - \nabla_j \ell(\cdot, d')\right| \le \Delta_2(\nabla_j \ell).$$

Theorem (Privacy Guarantees)

For $T > 0$, $\sigma_j^2 = \frac{8\boldsymbol{\Delta_2(\nabla_j \ell)^2} TK \log(1/\delta)}{n^2\epsilon^2}$,

$$\textit{DP-CD is } (\epsilon, \delta)\textit{-differentially-private.}$$

- $\Delta_2(\nabla_j \ell)$ can be **much smaller** than $\Delta_2(\nabla \ell)$.

Regularity Assumptions

For DP-SGD, smoothness was useful:

- ***β*-smoothness:** for $w, v \in \mathbb{R}^p$.

$$f(w) \leq f(v) + \langle \nabla f(v), w - v \rangle + \frac{\beta}{2} \left\| w - v \right\|_2^2 ,$$

Regularity Assumptions

But a finer, **coordinate-wise** measure is:

- ***M*-component-smoothness:** for $w, v \in \mathbb{R}^p$.

$$f(w) \leq f(v) + \langle \nabla f(v), w - v \rangle + \frac{1}{2} \|w - v\|_{\boldsymbol{M}}^2 ,$$

where M_j are **coordinate-wise** smoothness constants,

and $\|w\|_M^2 = \sum_{j=1}^{p} M_j w_j^2$.

(Similarly, measure strong convexity w.r.t. $\|\cdot\|_M$.)

Our Algorithm: Private Coordinate Descent

Utility: comparison with DP-SGD

Bounds on $\mathbb{E}[f(w_{priv}) - f(w^*)]$ are:

f is...	Convex	Strongly-convex
DP-CD	$\widetilde{O}\left(\frac{\sqrt{p\log(1/\delta)}}{n\epsilon}\Delta_{M^{-1}}(\nabla\ell)R_M\right)$	$\widetilde{O}\left(\frac{p\log(1/\delta)}{n^2\epsilon^2}\frac{\Delta_{M^{-1}}(\nabla\ell)^2}{\mu_M}\right)$
DP-SGD DP-SVRG	$\widetilde{O}\left(\frac{\sqrt{p\log(1/\delta)}}{n\epsilon}\Delta_2(\nabla\ell)R_2\right)$	$\widetilde{O}\left(\frac{p\log(1/\delta)}{n^2\epsilon^2}\frac{\Delta_2(\nabla\ell)^2}{\mu_2}\right)$

Where:

- $\Delta_{M^{-1}}(\nabla\ell)^2 = \sum_{j=1}^{p}\frac{1}{M_j}\Delta_2(\nabla_j\ell)^2$.
- $R_M = \left\|w^0 - w^*\right\|_M$, $R_2 = \left\|w^0 - w^*\right\|_2$.
- μ_2 (resp. μ_M) strong convexity parameters w.r.t. $\|\cdot\|_2$ (resp. $\|\cdot\|_M$).

Bounds on $\mathbb{E}[f(w_{priv}) - f(w^*)]$ are:

f is...	Convex
DP-CD	$\widetilde{O}\left(\frac{\sqrt{p\log(1/\delta)}}{n\epsilon}\Delta_{M^{-1}}(\nabla\ell)R_M\right)$
DP-SGD DP-SVRG	$\widetilde{O}\left(\frac{\sqrt{p\log(1/\delta)}}{n\epsilon}\Delta_2(\nabla\ell)R_2\right)$

Where:

- $\Delta_{M^{-1}}(\nabla\ell)^2 = \sum_{j=1}^{p} \frac{1}{M_j}\Delta_2(\nabla_j\ell)^2$.
- $R_M = \left\|w^0 - w^*\right\|_M$, $R_2 = \left\|w^0 - w^*\right\|_2$.

Our Algorithm: Private Coordinate Descent

Utility: comparison with DP-SGD

So we compare $\Delta_{M^{-1}}(\nabla \ell) R_M$ with $\Delta_2(\nabla \ell) R_2$

- If M_j's are equal:

$$1 \leq \frac{\Delta_{M^{-1}}(\nabla \ell) R_M}{\Delta_2(\nabla \ell) R_2} \leq p.$$

$\rightarrow$ DP-CD **is up to p times worse** than DP-SGD.

Utility: comparison with DP-SGD

So we compare $\Delta_{M^{-1}}(\nabla\ell)R_M$ with $\Delta_2(\nabla\ell)R_2$

- If M_j dominates $M_{j\neq 1}$ and $|w_1^0 - w_1^*| \leq |w_j^0 - w_j^*|$:

$$\frac{\Delta_{M^{-1}}(\nabla\ell)R_M}{\Delta_2(\nabla\ell)R_2} \leq \frac{1}{p}.$$

$\rightarrow$ DP-CD **is up to p times better** than DP-SGD.

1 Introduction

2 Background: Private ERM

3 Our Algorithm: Private Coordinate Descent

4 Experiments: Linear Regression

5 Conclusion and Perspectives

Experiments: Linear Regression

plots/reglin_lognormal_none.pdf plots/reglin_lognormal_none_spe

Uniform clipping: $C_j \propto \frac{1}{\sqrt{p}}$, Lipschitz Clipping: $C_j \propto \sqrt{\frac{M_j}{\sum_{j=1}^{p} M_j}}$.

($n = 1000$ samples, $p = 10$ features, $\epsilon = 1$, $\delta = 1/n^2$.)

Experiments: Linear Regression

plots/reglin_balanced_none.pdf

plots/reglin_balanced_none_spec

Uniform clipping: $C_j \propto \frac{1}{\sqrt{p}}$, Lipschitz Clipping: $C_j \propto \sqrt{\frac{M_j}{\sum_{j=1}^{p} M_j}}$.

($n = 1000$ samples, $p = 10$ features, $\epsilon = 1$, $\delta = 1/n^2$.)

Experiments: Linear Regression

plots/reglin_firstbig_none.pdf

plots/reglin_firstbig_none_spec

Uniform clipping: $C_j \propto \frac{1}{\sqrt{p}}$, Lipschitz Clipping: $C_j \propto \sqrt{\frac{M_j}{\sum_{j=1}^{p} M_j}}$.

($n = 1000$ samples, $p = 10$ features, $\epsilon = 1$, $\delta = 1/n^2$.)

1 Introduction

2 Background: Private ERM

3 Our Algorithm: Private Coordinate Descent

4 Experiments: Linear Regression

5 Conclusion and Perspectives

Conclusion and Perspectives

DP-CD:

- More queries to the data than DP-SGD.
- Lower sensitivities and larger learning rates.
- Correct clipping appears crucial.

Conclusion and Perspectives

DP-CD:

- More queries to the data than DP-SGD.
- Lower sensitivities and larger learning rates.
- Correct clipping appears crucial.

Pespectives include:

- **Composite** (non smooth) functions.
- **Adaptive** clipping thresholds.
- **Non-uniform** coordinates sampling.