# Differentially Private Coordinate Descent Methods

Paul Mangold (École Polytechnique)

Joint work with: Aurélien Bellet, Joseph Salmon and Marc Tommasi

ML-MTP Seminar
February 29th, 2024

| Record | Age $x_1$ | Pain $x_2$ | ... ... | Drug $x_p$ | Sick $y$ |
|---|---|---|---|---|---|
| #1 | 27 | 1 | ··· | 1 | 1 |
| #2 | 47 | 0 | ··· | 1 | 0 |
| #3 | 52 | 0 | ··· | 0 | 0 |
| #4 | 81 | 1 | ··· | 0 | 1 |
| ... | ... | ... | ... | ... | ... |
| #n | 13 | 1 | ··· | 0 | 1 |

How to study influence of possibly many features $x_i$'s on an outcome $y$?

| Record | Age $x_1$ | Pain $x_2$ | ... ... | Drug $x_p$ | Sick $y$ |
|---|---|---|---|---|---|
| #1 | 27 | 1 | ··· | 1 | 1 |
| #2 | 47 | 0 | ··· | 1 | 0 |
| #3 | 52 | 0 | ··· | 0 | 0 |
| #4 | 81 | 1 | ··· | 0 | 1 |
| ··· | ··· | ··· | ··· | ··· | ··· |
| #n | 13 | 1 | ··· | 0 | 1 |

How to study influence of possibly many features $x_i$'s on an outcome $y$?

One way: model $\log\left(\frac{\mathbb{P}(\text{sick})}{\mathbb{P}(\text{not sick})}\right)$ as

$$h_{w^*}(x) = w_0^* + w_1^* \cdot x_1 + \cdots + w_p^* \cdot x_p$$

| Record | Age $x_1$ | Pain $x_2$ | ... ... | Drug $x_p$ | Sick $y$ |
|---|---|---|---|---|---|
| #1 | 27 | 1 | ··· | 1 | 1 |
| #2 | 47 | 0 | ··· | 1 | 0 |
| #3 | 52 | 0 | ··· | 0 | 0 |
| #4 | 81 | 1 | ··· | 0 | 1 |
| ··· | ··· | ··· | ··· | ··· | ··· |
| #n | 13 | 1 | ··· | 0 | 1 |

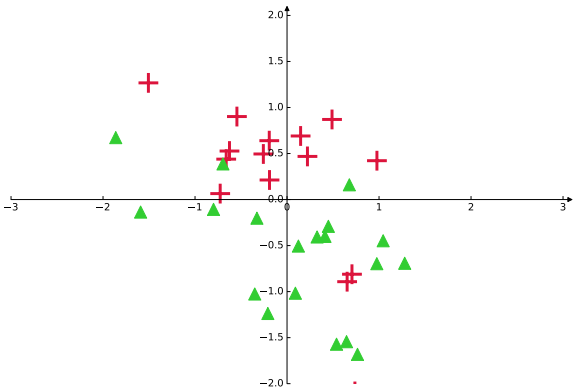
How to study influence of possibly many features $x_i$'s on an outcome $y$?

One way: model $\log\left(\frac{\mathbb{P}(\text{sick})}{\mathbb{P}(\text{not sick})}\right)$ as

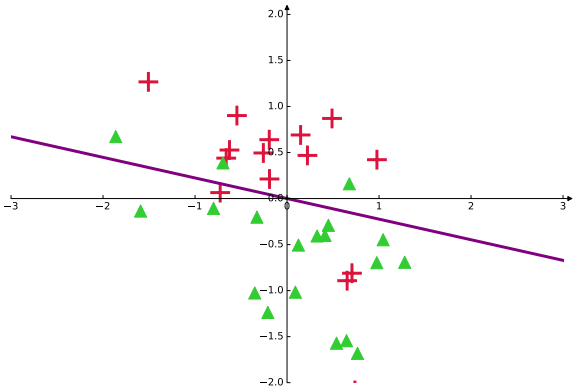
$$h_{w^*}(x) = w_0^* + w_1^* \cdot x_1 + \cdots + w_p^* \cdot x_p$$

Crucial fact: $w^*$ is **computed from the data!**

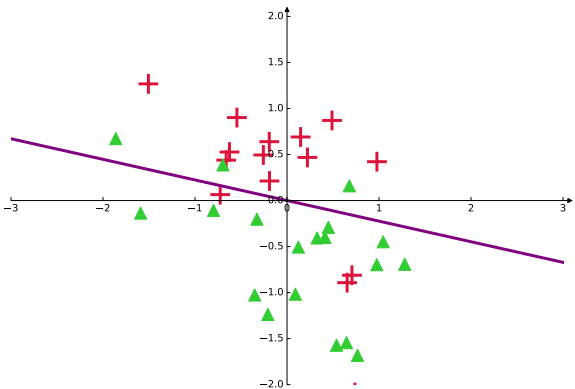
# $\Rightarrow$ Trained Classification Model



# $\Rightarrow$ Trained Classification Model



# ⇒ Trained Classification Model




The resulting model:

- is (quite) accurate
- contains info on data

# Privacy Issues

Membership inference*:

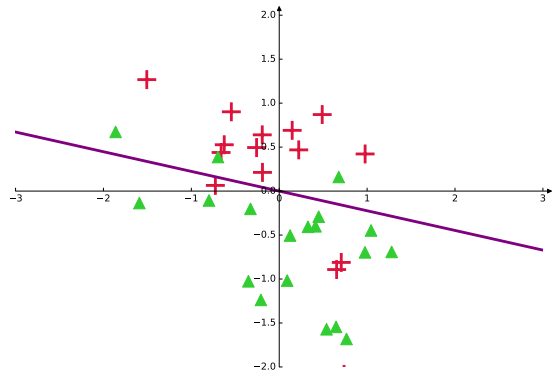
> " *determine whether a given record was part of a model's training dataset* "

*R. Shokri et al. "Membership Inference Attacks Against Machine Learning Models". 2017.

# Privacy Issues

Membership inference*:

> *“ determine whether a given record was part of a model’s training dataset ”*

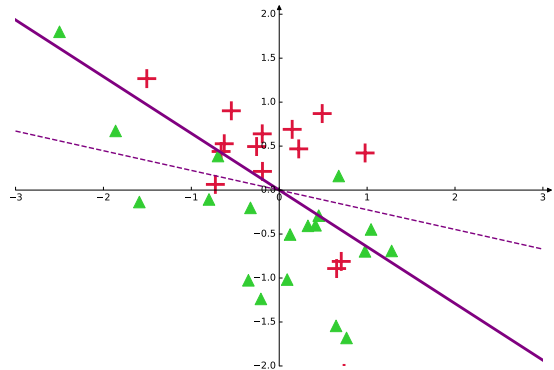



*R. Shokri et al. “Membership Inference Attacks Against Machine Learning Models”. 2017.

# Privacy Issues

Membership inference*:

> *“ determine whether a given record was part of a model’s training dataset ”*




*R. Shokri et al. “Membership Inference Attacks Against Machine Learning Models”. 2017.

# Guaranteeing Privacy

Perturb the predictor:

$$h_{w^*}(x) = w_0^* + w_1^* \cdot x_1 + \cdots + w_p^* \cdot x_p$$

# Guaranteeing Privacy

Perturb the predictor:

$$h_{w^*+\eta}(x) = (w_0^* + \eta_0) + (w_1^* + \eta_1) \cdot x_1 + \cdots + (w_p^* + \eta_p) \cdot x_p$$

# Guaranteeing Privacy

Perturb the predictor:

$$h_{w^*+\eta}(x) = (w_0^* + \eta_0) + (w_1^* + \eta_1) \cdot x_1 + \cdots + (w_p^* + \eta_p) \cdot x_p$$

✓ noise gives *plausible deniability* $\rightarrow$ better privacy

✗ noisy predictions $\rightarrow$ lower accuracy

# Guaranteeing Privacy

Perturb the predictor:

$$h_{w^*+\eta}(x) = (w_0^* + \eta_0) + (w_1^* + \eta_1) \cdot x_1 + \cdots + (w_p^* + \eta_p) \cdot x_p$$

✓ noise gives *plausible deniability* $\rightarrow$ better privacy

✗ noisy predictions $\rightarrow$ lower accuracy

$\Rightarrow$ **tension between privacy and utility**

# How Strong is the Protection?

$\mathcal{A} : D \mapsto w$ is $(\epsilon, \delta)$-*Differentially Private*[*]

[*] C. Dwork. "Differential Privacy". 2006.

# How Strong is the Protection?

$\mathcal{A} : D \mapsto w$ is $(\epsilon, \delta)$-*Differentially Private*[*]

$$\mathbb{P}(\mathcal{A}(D) \in \mathcal{S}) \leq \exp(\boldsymbol{\epsilon}) \cdot \mathbb{P}(\mathcal{A}(D') \in \mathcal{S}) + \boldsymbol{\delta}$$

for *all* $D, D'$ that differ on one element ($D \sim D'$)

---

[*] C. Dwork. "Differential Privacy". 2006.

# How Strong is the Protection?

$\mathcal{A} : D \mapsto w$ is $(\epsilon, \delta)$-*Differentially Private*[*]

$$\mathbb{P}(\mathcal{A}(D) \in \mathcal{S}) \leq \exp(\boldsymbol{\epsilon}) \cdot \mathbb{P}(\mathcal{A}(D') \in \mathcal{S}) + \boldsymbol{\delta}$$

for *all* $D, D'$ that differ on one element ($D \sim D'$)

Rule of thumb: $\boldsymbol{\epsilon} \leq 1$, $\boldsymbol{\delta} = o(1/|D|)$

---

[*] C. Dwork. "Differential Privacy". 2006.

## Ingredients for building $(\epsilon, \delta)$-DP algorithms

1. Gaussian mechanism: $\mathcal{G}_{f,\sigma^2}(D) = f(D) + \mathcal{N}(0; \sigma^2)$
2. Composition of DP algorithms
3. Amplification by sampling

## Ingredients for building $(\epsilon, \delta)$-DP algorithms

1. Gaussian mechanism: $\mathcal{G}_{f,\sigma^2}(D) = f(D) + \boldsymbol{\mathcal{N}(0; \sigma^2)}$

   To guarantee $(\epsilon, \delta)$-DP:

   * Compute sensitivity $\Delta f = \sup_{D \sim D'} \|f(D) - f(D')\|_2$
   * Scale noise variance $\sigma^2 \propto \frac{(\Delta f)^2 \log(1/\delta)}{\epsilon^2}$

2. Composition of DP algorithms

3. Amplification by sampling

## Ingredients for building $(\epsilon, \delta)$-DP algorithms

1. Gaussian mechanism: $\mathcal{G}_{f,\sigma^2}(D) = f(D) + \mathcal{N}(0; \sigma^2)$
2. Composition of DP algorithms

   Release a $(\epsilon, \delta)$-DP value $T$ times over same data

   $\rightarrow$ privacy guarantees **decrease** to $\left(O(\sqrt{T}\epsilon), O(T\delta)\right)$-DP
3. Amplification by sampling

## Ingredients for building $(\epsilon, \delta)$-DP algorithms

1. Gaussian mechanism: $\mathcal{G}_{f,\sigma^2}(D) = f(D) + \mathcal{N}(0; \sigma^2)$
2. Composition of DP algorithms
3. Amplification by sampling

   Sample a fraction $q$ of $D$'s elements and use the Gaussian mechanism

   $\rightarrow$ privacy guarantees **increase** to $\big(O(\boldsymbol{q}\epsilon), O(\boldsymbol{q}\delta)\big)$-DP

# Outline

I. Classical private optimization

II. Private Stochastic CD for imbalanced problems

III. Private Greedy CD for sparse problems

# Empirical Risk Minimization

Note: Most results also hold for composite ERM with Proximal algorithms

$$w^* \in \arg\min_{w \in \mathcal{W}} \left\{ f(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; d_i) \right\}$$

# Empirical Risk Minimization

Note: Most results also hold for composite ERM with Proximal algorithms

$$w^* \in \arg\min_{w \in \mathcal{W}} \left\{ f(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; d_i) \right\}$$

Where $\mathcal{W} \subseteq \mathbb{R}^p$, has diameter $\|\mathcal{W}\|_2$, and $\ell$ is

* strongly convex: $\ell(w; d) \geq \ell(w'; d) + \langle \nabla \ell(w'; d), w - w' \rangle + \frac{\mu_2}{2} \|w - w'\|^2$
* smooth: $\|\nabla \ell(w; d) - \nabla \ell(w'; d)\| \leq M \|w - w'\|$
* Lipschitz: $|\ell(w; d) - \ell(w'; d)| \leq \Lambda \|w - w'\|$

# Empirical Risk Minimization

Note: Most results also hold for composite ERM with Proximal algorithms

$$w^* \in \arg\min_{w \in \mathcal{W}} \left\{ f(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; d_i) \right\}$$

Where $\mathcal{W} \subseteq \mathbb{R}^p$, has diameter $\|\mathcal{W}\|_2$, and $\ell$ is

* strongly convex: $\ell(w; d) \geq \ell(w'; d) + \langle \nabla \ell(w'; d), w - w' \rangle + \frac{\mu_2}{2} \|w - w'\|^2$
* smooth: $\|\nabla \ell(w; d) - \nabla \ell(w'; d)\| \leq M \|w - w'\|$
* Lipschitz: $|\ell(w; d) - \ell(w'; d)| \leq \Lambda \|w - w'\|$

# Empirical Risk Minimization

Note: Most results also hold for composite ERM with Proximal algorithms

$$w^* \in \arg\min_{w \in \mathcal{W}} \left\{ f(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; d_i) \right\}$$

Where $\mathcal{W} \subseteq \mathbb{R}^p$, has diameter $\|\mathcal{W}\|_2$, and $\ell$ is

* strongly convex: $\ell(w; d) \geq \ell(w'; d) + \langle \nabla \ell(w'; d), w - w' \rangle + \frac{\mu_2}{2} \|w - w'\|^2$
* smooth: $\|\nabla \ell(w; d) - \nabla \ell(w'; d)\| \leq M \|w - w'\|$
* Lipschitz: $|\ell(w; d) - \ell(w'; d)| \leq \Lambda \|w - w'\|$

# Empirical Risk Minimization

Note: Most results also hold for composite ERM with Proximal algorithms

$$w^* \in \arg\min_{w \in \mathcal{W}} \left\{ f(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; d_i) \right\}$$

Where $\mathcal{W} \subseteq \mathbb{R}^p$, has diameter $\|\mathcal{W}\|_2$, and $\ell$ is

* strongly convex: $\ell(w; d) \geq \ell(w'; d) + \langle \nabla \ell(w'; d), w - w' \rangle + \frac{\mu_2}{2} \|w - w'\|^2$
* smooth: $\|\nabla \ell(w; d) - \nabla \ell(w'; d)\| \leq M \|w - w'\|$
* Lipschitz: $|\ell(w; d) - \ell(w'; d)| \leq \Lambda \|w - w'\|$

# Empirical Risk Minimization

Note: Most results also hold for composite ERM with Proximal algorithms

$$w^* \in \arg\min_{w \in \mathcal{W}} \left\{ f(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; d_i) \right\}$$

Where $\mathcal{W} \subseteq \mathbb{R}^p$, has diameter $\|\mathcal{W}\|_2$, and $\ell$ is

* strongly convex: $\ell(w; d) \geq \ell(w'; d) + \langle \nabla \ell(w'; d), w - w' \rangle + \frac{\mu_2}{2} \|w - w'\|^2$
* smooth: $\|\nabla \ell(w; d) - \nabla \ell(w'; d)\| \leq M \|w - w'\|$
* Lipschitz: $\|\nabla \ell(w; d)\| \leq \Lambda$

# Empirical Risk Minimization

Note: Most results also hold for composite ERM with Proximal algorithms

How to solve ERM privately?

* smooth: $\|\nabla \ell(w; d) - \nabla \ell(w'; d)\| \leq M\|w - w'\|$
* Lipschitz: $\|\nabla \ell(w; d)\| \leq \Lambda$

# DP-SGD*, †

## Differentially Private Stochastic Gradient Descent

For $t = 0$ to $T - 1$:

* Choose a data record $d_i$
* Draw noise $\boldsymbol{\eta^t \sim \mathcal{N}(0; \sigma^2 \mathbb{I}_p)}$
* Update $w^{t+1} = w^t - \gamma^t \left(\nabla \ell(w^t; d_i) + \eta^t\right)$

Return $w^T$

---

*S. Song et al. "Stochastic Gradient Descent with Differentially Private Updates". 2013.

†R. Bassily et al. "Private ERM: Efficient Algorithms and Tight Error Bounds". 2014.

# Privacy of DP-SGD*, †

For $(\epsilon, \delta)$-differential privacy we need

$$\sigma^2 = O\left(\frac{\Lambda T}{n^2 \epsilon^2}\right) \quad , \quad \text{where } \|\nabla \ell\| \leq \Lambda$$

* Noise increases with number of iterations
* Sampling amplifies privacy

*S. Song et al. "Stochastic Gradient Descent with Differentially Private Updates". 2013.

†R. Bassily et al. "Private ERM: Efficient Algorithms and Tight Error Bounds". 2014.

# Utility of DP-SGD*

When $f$ is $\mu$-strongly-convex w.r.t. the norm $\|\cdot\|_2$,

$$\mathbb{E}(f(w^{SGD}) - f(w^*)) = O\left(\frac{\Lambda^2 \log(T)}{\mu T} + \frac{p\Lambda^2 \log(1/\delta)}{n^2\epsilon^2}\right)$$

optimization error

privacy error

*R. Bassily et al. "Private ERM: Efficient Algorithms and Tight Error Bounds". 2014.

# Utility of DP-SGD*

When $f$ is $\mu$-strongly-convex w.r.t. the norm $\|\cdot\|_2$,

$$\mathbb{E}(f(w^{SGD}) - f(w^*)) = O\left(\frac{p\Lambda^2 \log(T)\log(1/\delta)}{\mu n^2 \epsilon^2}\right)$$

choose $T$ to balance the two terms

*R. Bassily et al. "Private ERM: Efficient Algorithms and Tight Error Bounds". 2014.

# Utility of DP-SGD*

When $f$ is $\mu$-strongly-convex w.r.t. the norm $\|\cdot\|_2$,

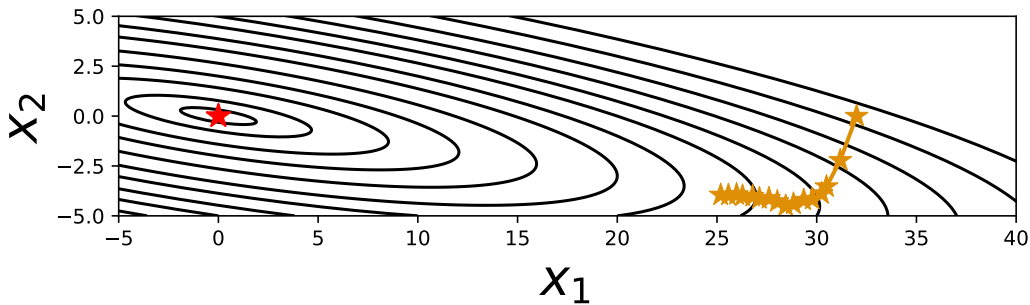
$$\mathbb{E}(f(w^{SGD}) - f(w^*)) = \Theta\left(\frac{p\Lambda^2 \log(T)\log(1/\delta)}{\mu n^2 \epsilon^2}\right)$$

$\Rightarrow$ and the result is *tight* (under these assumptions)

*R. Bassily et al. "Private ERM: Efficient Algorithms and Tight Error Bounds". 2014.

# The Problem of DP-SGD

It fails on imbalanced problems...



We need to refine measure of regularity of $f$:

* smoothness:

$$\|\nabla f(w+t) - \nabla f(w)\| \leq M\|t\|$$

* Lipschitzness:

$$\|\nabla f(w)\| \leq \Lambda$$

We need to refine measure of regularity of $f$:

* coordinate-wise smoothness:

$$|\nabla_j f(w + te_j) - \nabla_j f(w)| \leq M_j |t|$$

* coordinate-wise Lipschitzness:

$$|\nabla_j f(w)| \leq L_j$$

We need to refine measure of regularity of $f$:

* coordinate-wise smoothness:

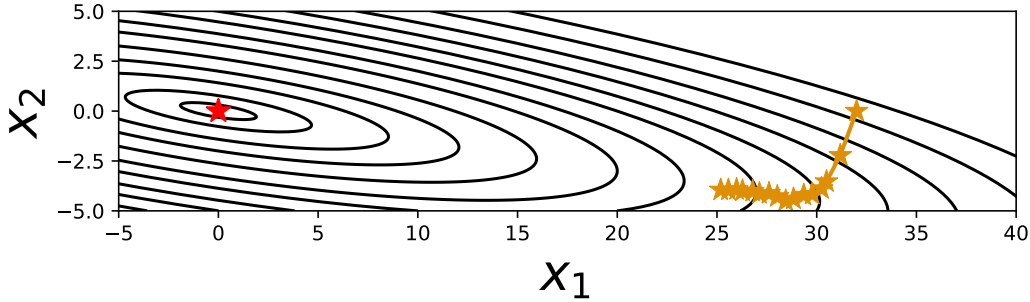
$$|\nabla_j f(w + te_j) - \nabla_j f(w)| \leq M_j |t|$$

* coordinate-wise Lipschitzness:

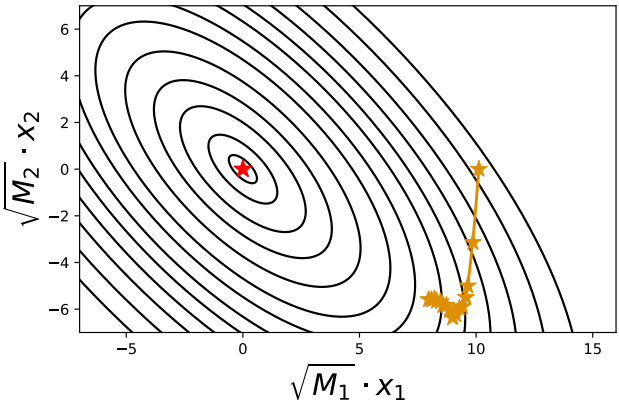
$$|\nabla_j f(w)| \leq L_j$$

Important: $M_j \leq M$, and $L_j \leq \Lambda$

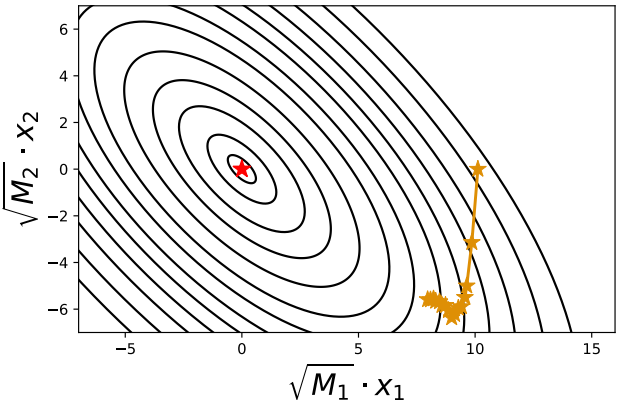
We can now use a more appropriate measure of our space!



We can now use a more appropriate measure of our space!



We can now use a more appropriate measure of our space!




Scaled norm: $\|w\|_{M,q} = \left( \sum_{j=1}^{p} M_j^{\frac{q}{2}} |w_j|^q \right)^{\frac{1}{q}}$ for $q \in \{1, 2\}$

And measure strong convexity appropriately:

* $\mu_2$-strong-convexity w.r.t $\|\cdot\|$

  $\ell(w;d) \geq \ell(w';d) + \langle \nabla \ell(w';d), w - w' \rangle + \frac{\mu_2}{2}\|w - w'\|^2$

* $\mu_{M,q}$-strong-convexity w.r.t $\|\cdot\|_{M,q}$

  $\ell(w;d) \geq \ell(w';d) + \langle \nabla \ell(w';d), w - w' \rangle + \frac{\boldsymbol{\mu_{M,q}}}{2}\|w - w'\|^2_{\boldsymbol{M,q}}$

It holds that $\boldsymbol{\mu_{M,2} \geq \mu_2}$

Using the scaled norm: $\|w\|_{M,q} = \Big(\sum_{j=1}^{p} M_j^{\frac{q}{2}} |w_j^q|\Big)^{\frac{1}{q}}$ for $q \in \{1,2\}$

# Differentially Private Coordinate Descent*

For $t = 0$ to $T - 1$:

* Choose a *coordinate* $j \in [p]$
* Draw noise $\eta_j^t \sim \mathcal{N}\left(0; \sigma_j^2\right)$
* Update $w_j^{t+1} = w_j^t - \gamma_j(\nabla_j f(w^t) + \eta_j^t)$, typically $\gamma_j \propto \frac{1}{M_j}$

Return $w^{CD} = \frac{1}{T}\sum_{t=1}^{T} w^t$

*P. Mangold et al. "Differentially Private Coordinate Descent for Composite ERM". 2022.

# Differentially Private Coordinate Descent*

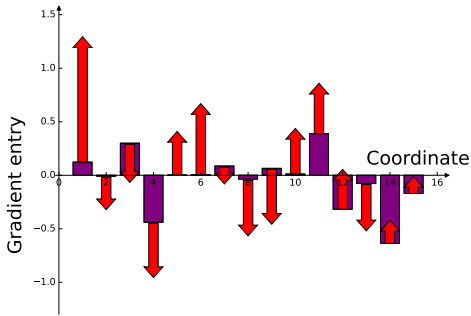
For $t = 0$ to $T - 1$:

- Choose a *coordinate* $j \in [p]$
- Draw noise $\eta_j^t \sim \mathcal{N}\left(0; O\left(\frac{L_j T}{n^2 \epsilon^2}\right)\right)$
- Update $w_j^{t+1} = w_j^t - \gamma_j(\nabla_j f(w^t) + \eta_j^t)$, typically $\gamma_j \propto \frac{1}{M_j}$

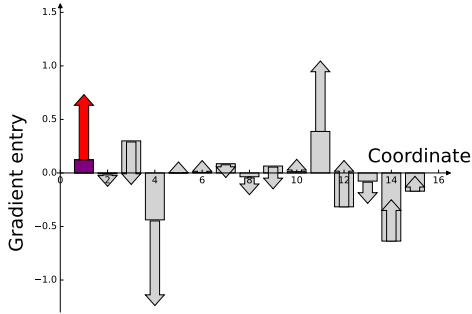
Return $w^{CD} = \frac{1}{T}\sum_{t=1}^{T} w^t$

*P. Mangold et al. "Differentially Private Coordinate Descent for Composite ERM". 2022.

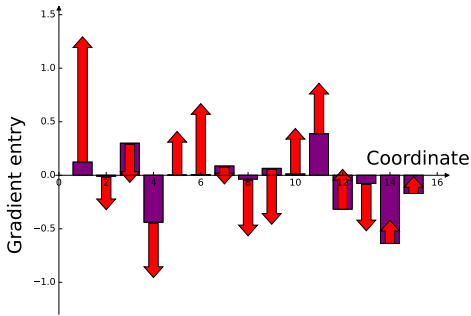
DP-SGD noise:



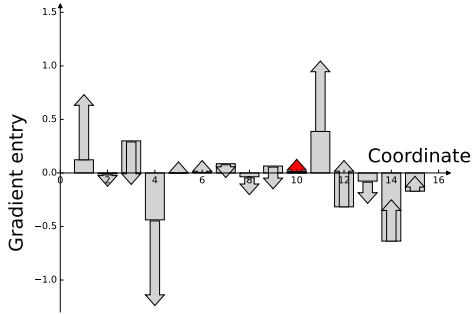

DP-CD noise:



DP-SGD noise:




DP-CD noise:



# Utility of DP-CD

For $\mu_{M,2}$-strongly-convex functions

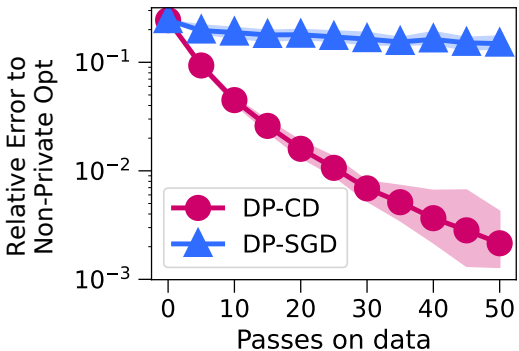
$$\mathbb{E}(f(w^{CD}) - f(w^*)) \leq O\left(\frac{p\log(1/\delta)}{\mu_{M,2} n^2 \epsilon^2} \|L\|^2_{M^{-1}}\right)$$

Recall that for DP-SGD:

$$\mathbb{E}(f(w^{SGD}) - f(w^*)) \leq O\left(\frac{p\log(1/\delta)}{\mu_2 n^2 \epsilon^2} \Lambda^2\right)$$

# Numerical Illustration

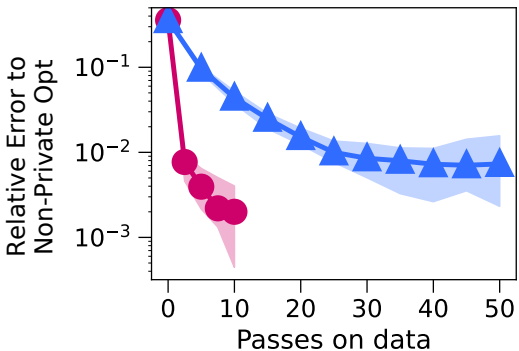
DP-CD uses more appropriate step sizes




* Regularized logistic regression
* Raw (imbalanced) data
* $n = 45,312$ records
* $p = 8$ features
* $\epsilon = 1$, $\delta = 1/n^2$

# Numerical Illustration

DP-CD does not require amplification by sampling




- Regularized logistic regression
- Standardized data
- $n = 45,312$ records
- $p = 8$ features
- $\epsilon = 1$, $\delta = 1/n^2$

# Practical Considerations

* Clipping:

$$\text{clip}(\nabla_j f(w), C_j) = \text{sign}(\nabla_j f(w)) \min(|\nabla_j f(w)|, C_j)$$

$\rightarrow$ guarantees that $\Delta(\nabla_j f) \leq 2C_j$

$\rightarrow$ scaling $C_j = \sqrt{\frac{M_j}{\sum_j M_j}} C$ works well

* Estimation of constants: $M_j$'s contain sensitive information...

# Choice of updated coordinate?

In DP-CD, we chose updated coordinate as:

* Choose a *coordinate* $j \in [p]$

To propose something different, we need another DP ingredient: **report noisy max**

$$j = \arg\max_{j' \in [p]} |\nabla_{j'} f(w^t) + \zeta_{j'}| \ , \ \text{with } \zeta_j^t \sim \text{Lap}\left(0; O\left(\frac{L_j T}{n^2 \epsilon^2}\right)\right)$$

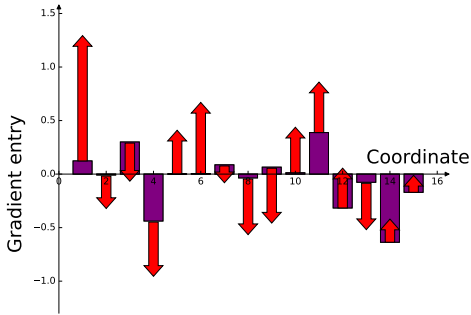
# Differentially Private **Greedy** CD

For $t = 0$ to $T - 1$:

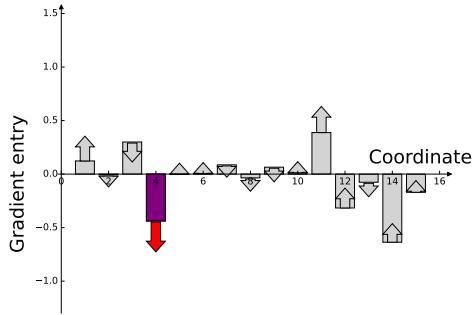
* Draw noise $\eta_j^t, \zeta_j^t \sim \text{Lap}\left(0; O\left(\frac{L_j T}{n^2 \epsilon^2}\right)\right)$
* Choose $j = \underset{j' \in [p]}{\arg\max} |\nabla_{j'} f(w^t) + \zeta_{j'}|$
* Update $w^{t+1} = w^t - \gamma_j (\nabla_j f(w^t) + \eta_j^t)$

Return $w^{GCD} = w^T$

DP-SGD noise:




DP-GCD noise:



# Utility of DP-GCD

When $f$ is $\mu_{M,1}$-strongly convex w.r.t $\|\cdot\|_{M,1}$

$$\mathbb{E}(f(w^{SGD}) - f(w^*)) = O\left(\frac{L_{\max}^2 \log(1/p)\log(1/\delta)}{M_{\min}\mu_{M,1}^2 n^2\epsilon^2}\right)$$

Recall that for DP-SGD:

$$\mathbb{E}(f(w^{SGD}) - f(w^*)) \leq O\left(\frac{p\log(1/\delta)}{\mu_2 n^2\epsilon^2}\Lambda^2\right)$$

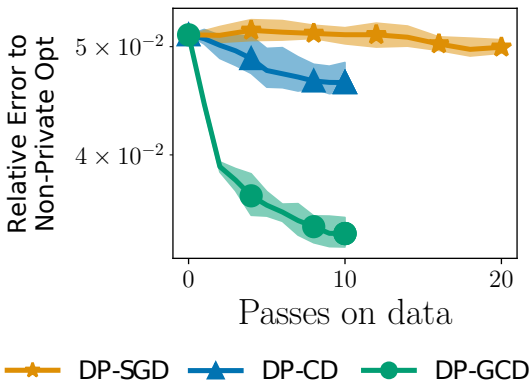
# Utility of DP-GCD
## Problems with sparse solution

When $f$ is $\mu_{M,2}$-strongly-convex w.r.t $\|\cdot\|_M$, and solution is (close to) $\tau$-sparse

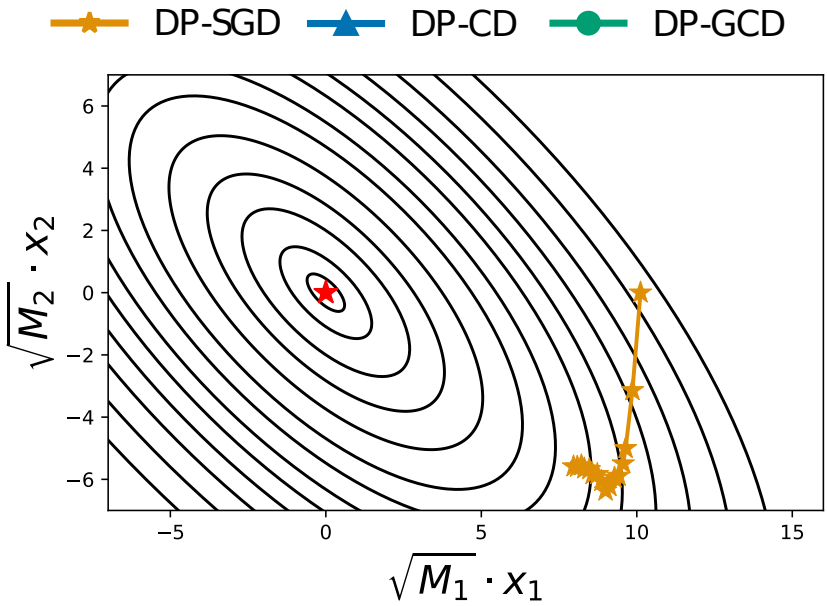
$$\mathbb{E}(f(w^{SGD}) - f(w^*)) = O\left(\frac{L_{\max}^2 \tau^2 \log(1/p)\log(1/\delta)}{M_{\min}\mu_{M,2}^2 n^2 \epsilon^2}\right)$$

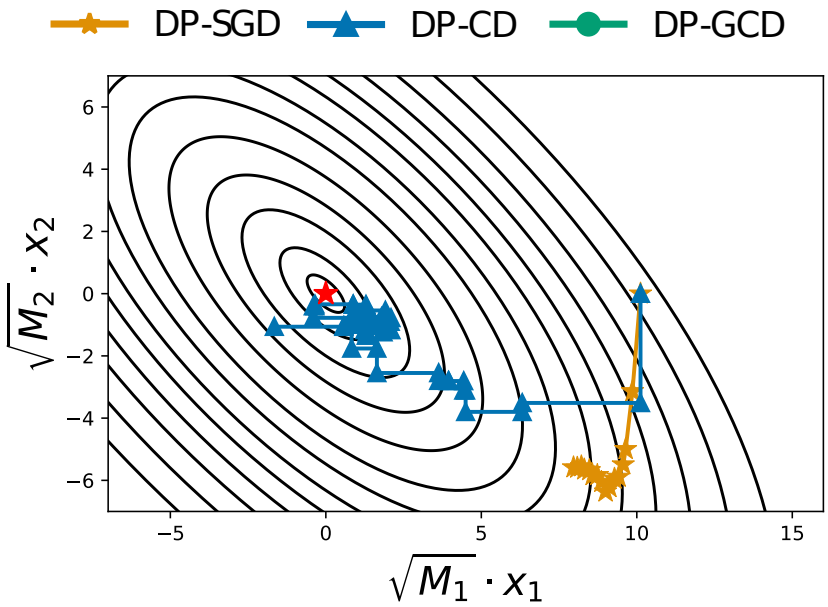
# Numerical Illustration

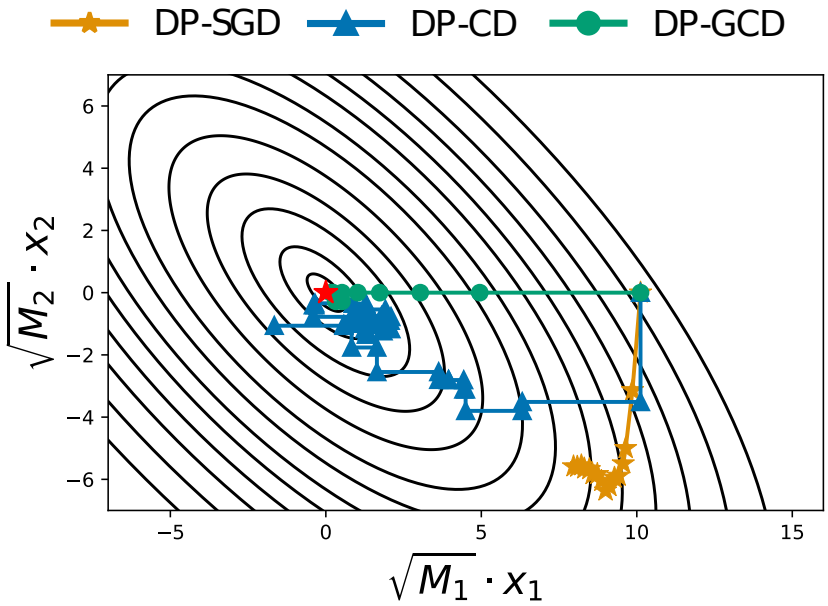
## DP-GCD can focus on relevant coordinates




* Regularized logistic regression
* Standardized data
* $n = 2,600$ records
* $p = 501$ features
* $\epsilon = 1$, $\delta = 1/n^2$


DP-SGD
DP-CD
DP-GCD
6
4
2
0
−2
−4
−6
−5
0
5
10
15
$\sqrt{M_2} \cdot x_2$
$\sqrt{M_1} \cdot x_1$


DP-SGD
DP-CD
DP-GCD
6
4
2
0
−2
−4
−6
−5
0
5
10
15
$\sqrt{M_2} \cdot x_2$
$\sqrt{M_1} \cdot x_1$


DP-SGD
DP-CD
DP-GCD
6
4
2
0
−2
−4
−6
−5
0
5
10
15
$\sqrt{M_2} \cdot x_2$
$\sqrt{M_1} \cdot x_1$

# Conclusion

Private coordinate descent methods can exploit:

- imbalance in parameter scales and variations
- imbalance/sparsity of the solution
- adapt to underlying structure

# Conclusion

Private coordinate descent methods can exploit:

* imbalance in parameter scales and variations
* imbalance/sparsity of the solution
* adapt to underlying structure

Open questions: adaptive step sizes and clipping, better sampling of coordinates, analyze proximal greedy CD...

# Thank you!

See the papers:

- P. Mangold, A. Bellet, J. Salmon, and M. Tommasi. "Differentially Private Coordinate Descent for Composite ERM". 2022 (ICML)

- P. Mangold, A. Bellet, J. Salmon, and M. Tommasi. "High-Dimensional Private ERM by Greedy Coordinate Descent". 2023 (AISTATS)