Coordinate Descent for Private Composite Empirical Risk Minimization

Paul Mangold

(Aurélien Bellet, Marc Tommasi, Joseph Salmon)

Magnet Seminar

February 17th, 2022

Supervised learning:

$$D = \{d_1, \dots, d_n\} \subseteq \mathcal{X} \times \mathcal{Y}.$$

Learn *good parameters* $w \in \mathbb{R}^p$ for

$$h_w : \mathcal{X} \mapsto \mathcal{Y}.$$

Empirical Risk Minimization

$$\arg\min_{w \in \mathbb{R}^p} F(w) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(w; d_i)}_{f(w)} .$$

Assumptions:

- $\ell(\cdot; d)$ convex, component-Lipschitz $\forall d \in \mathcal{X} \times \mathcal{Y}$.
- $\ell(\cdot; d_i)$ component-smooth $\forall d_i \in D$.

$\ell(\cdot; d)$ is component-Lipschitz for $d \in \mathcal{X} \times \mathcal{Y}$:

$$|\ell(w + te_j; d) - \ell(w; d)| \leq L_j \, |t| \,.$$

$\ell(\cdot; d)$ is component-Lipschitz for $d \in \mathcal{X} \times \mathcal{Y}$:

$$|\ell(w + te_j; d) - \ell(w; d)| \leq L_j \, |t| \, .$$

$\Rightarrow$ for $w \in \mathbb{R}^p$, $|\nabla_j \ell(w; d)| \leq L_j$.

$\ell(\cdot; d)$ is component-Lipschitz for $d \in \mathcal{X} \times \mathcal{Y}$:

$$|\ell(w + te_j; d) - \ell(w; d)| \leq L_j \, |t| \, .$$

$\Rightarrow$ for $w \in \mathbb{R}^p$, $|\nabla_j \ell(w; d)| \leq L_j$.

for $w, w' \in \mathbb{R}^p$ and $d, d' \in \mathcal{X} \times \mathcal{Y}$,

$|\nabla_j \ell(w; d) - \nabla_j \ell(w'; d')| \leq 2L_j$.

$\ell(\cdot; d)$ is component-smooth for $d \in D$:

$$\left|\nabla_j \ell(w + te_j; d) - \nabla_j \ell(w; d)\right| \leq M_j \left|t\right|.$$

$\ell(\cdot; d)$ is component-smooth for $d \in D$:

$$\left|\nabla_j \ell(w + te_j; d) - \nabla_j \ell(w; d)\right| \leq M_j \left|t\right|.$$

$\Rightarrow$ for $w, w' \in \mathbb{R}^p$,

$$f(w') \leq f(w) + \langle \nabla f(w), w' - w \rangle + \frac{1}{2} \left\| w' - w \right\|_M^2.$$

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

Composite ERM

$$\underset{w \in \mathbb{R}^p}{\arg\min}\, F(w) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(w; d_i)}_{f(w)} + \psi(w).$$

Assumptions:

- $\ell(\cdot; d)$ convex, component-Lipschitz $\forall d \in \mathcal{X} \times \mathcal{Y}$.
- $\ell(\cdot; d_i)$ component-smooth $\forall d_i \in D$.
- $\psi(w) = \sum_{j=1}^{p} \psi_j(w_j)$ convex and separable.

Example: LASSO

$$\underset{w \in \mathbb{R}^p}{\arg\min} \frac{1}{2n} \left\| Xw - y \right\|_2^2 + \lambda \left\| w \right\|_1 .$$

Example: LASSO

$$\underset{w \in \mathbb{R}^p}{\arg\min} \frac{1}{2n} \|Xw - y\|_2^2 + \lambda \|w\|_1 .$$

$$M_j = \frac{1}{n} \sum_{i=1}^{n} |X_{i,j}|^2 .$$

$\mathcal{A} : D \mapsto w$ is (ϵ, δ)-*Differentially Private*

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

$(D$ and D' differ on one element.)

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

Differentially Private Composite ERM:

$$\arg\min_{w \in \mathbb{R}^p} F(w) = f(w) + \psi(w)$$

$$\text{such that } w \text{ is } (\epsilon, \delta)\text{-DP.}$$

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

Solving DP-ERM?

$$\arg\min_{w \in \mathbb{R}^p} F(w) = f(w) + \psi(w)$$
$$\text{such that } w \text{ is } (\epsilon, \delta)\text{-DP.}$$

Solving DP-ERM?

$$\arg\min_{w \in \mathbb{R}^p} F(w) = f(w) + \psi(w)$$
$$\text{such that } w \text{ is } (\epsilon, \delta)\text{-DP.}$$

The classical: DP-SGD.

Solving DP-ERM?

$$\arg\min_{w \in \mathbb{R}^p} F(w) = f(w) + \psi(w)$$
$$\text{such that } w \text{ is } (\epsilon, \delta)\text{-DP}.$$

The classical: DP-SGD.
The challenger: DP-CD.

Stochastic Gradient Descent

$$w^{t+1} = \mathrm{prox}_{\eta\psi}\left(w^t - \eta \boldsymbol{\xi}\right),$$

where $\mathbb{E}[\boldsymbol{\xi}] = \nabla f(w^t)$.

– A. Beck and M. Teboulle, *"A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems"*, 2009.

– K. Mishchenko, A. Khaled, and P. Richtárik, *"Proximal and Federated Random Reshuffling"*, 2021.

Private Stochastic Gradient Descent

$$w^{t+1} = \operatorname{prox}_{\eta\psi}\left(w^t - \eta\left(\boldsymbol{\xi} + \boldsymbol{\mathcal{N}}(\boldsymbol{\sigma^2}\mathbb{1})\right)\right),$$

where $\mathbb{E}[\boldsymbol{\xi}] = \nabla f(w^t)$.

– R. Bassily, A. Smith, and A. Thakurta, *"Differentially 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"*, 2017.

Coordinate Descent

$$w_j^{t+1} = \mathrm{prox}_{\eta_j \psi_j}\left(w_j^t - \eta_j \boldsymbol{\nabla}_{\boldsymbol{j}} \boldsymbol{f}(\boldsymbol{w}^{\boldsymbol{t}})\right),$$

where $j \in \{1, \ldots, p\}$ is chosen randomly.

– P. Richtárik and M. Takáč, *"Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function"*, 2014.

Private Coordinate Descent

$$w_j^{t+1} = \mathrm{prox}_{\eta_j \psi_j}\left(w_j^t - \eta_j \left(\boldsymbol{\nabla_j f(w^t)} + \boldsymbol{\mathcal{N}(\sigma_j^2)} \right) \right),$$

where $j \in \{1, \ldots, p\}$ is chosen randomly.

DP-SGD:

$$w^{t+1} = \operatorname{prox}_{\boldsymbol{\eta}\psi}\left(w^t - \boldsymbol{\eta}\left(\boldsymbol{\xi} + \boldsymbol{\mathcal{N}}(\boldsymbol{\sigma}^2\mathbb{1})\right)\right),$$

$$\text{with } \mathbb{E}[\xi] = \nabla f(w^t).$$

DP-CD:

$$w_j^{t+1} = \operatorname{prox}_{\boldsymbol{\eta_j}\psi_j}\left(w_j^t - \boldsymbol{\eta_j}\left(\boldsymbol{\nabla_j f(w^t)} + \boldsymbol{\mathcal{N}}(\boldsymbol{\sigma_j^2})\right)\right).$$

	DP-SGD	DP-CD
Assumptions on f	Λ-Lipschitz β-smooth	L-comp.-Lipschitz M-comp.-smooth
Step sizes		

	DP-SGD	DP-CD
Assumptions on f	Λ-Lipschitz β-smooth	L-comp.-Lipschitz M-comp.-smooth
Step sizes	$\eta \propto 1/\beta$	

	DP-SGD	DP-CD
Assumptions on f	Λ-Lipschitz β-smooth	L-comp.-Lipschitz M-comp.-smooth
Step sizes	$\eta \propto 1/\beta$	$\eta_j \propto 1/M_j$

	DP-SGD	DP-CD
Assumptions on f	Λ-Lipschitz β-smooth	L-comp.-Lipschitz M-comp.-smooth
Step sizes	$\eta \propto 1/\beta$	$\eta_j \propto 1/M_j$
Noise scale		

	DP-SGD	DP-CD
Assumptions on f	Λ-Lipschitz β-smooth	L-comp.-Lipschitz M-comp.-smooth
Step sizes	$\eta \propto 1/\beta$	$\eta_j \propto 1/M_j$
Noise scale	$\sigma^2 = O\left(\frac{\boldsymbol{\Lambda^2} T}{n^2\epsilon^2}\right)$	

Since $\sup_{d,d'} \|\nabla \ell(\cdot; d) - \nabla \ell(\cdot; d')\|_2 \leq 2\Lambda$.

	DP-SGD	DP-CD
Assumptions on f	Λ-Lipschitz β-smooth	L-comp.-Lipschitz M-comp.-smooth
Step sizes	$\eta \propto 1/\beta$	$\eta_j \propto 1/M_j$
Noise scale	$\sigma^2 = O\left(\frac{\boldsymbol{\Lambda^2} T}{n^2\epsilon^2}\right)$	$\sigma_j^2 = O\left(\frac{\boldsymbol{L_j^2} T}{n^2\epsilon^2}\right)$

Since $\sup_{d,d'} |\nabla_j \ell(\cdot; d) - \nabla_j \ell(\cdot; d')| \le 2L_j$.

What about this $O\left(\,\cdot\,\right)$?

DP-SGD: sampling rate $q \leq 1$,

$$\epsilon \leq \frac{1}{\alpha - 1} \log \left(\frac{1}{2} \sum_{k=0}^{\infty} \binom{\alpha}{k} (1-q)^{\alpha - k} q^k \exp\left(\frac{k^2 - k}{2\sigma^2} \right) \left(\operatorname{erfc}\left(\frac{z_1 - k}{\sqrt{2}\sigma} \right) + \operatorname{erfc}\left(\frac{k - z_1}{\sqrt{2}\sigma} \right) \right) \right) + \frac{\log(1/\delta)}{\alpha - 1}.$$

What about this $O\Big(\cdot\Big)$?

DP-SGD: sampling rate $q \leq 1$,

$$\epsilon \leq \frac{1}{\alpha-1}\log\left(\frac{1}{2}\sum_{k=0}^{\infty}\binom{\alpha}{k}(1-q)^{\alpha-k}q^{k}\exp\left(\frac{k^2-k}{2\sigma^2}\right)\left(\operatorname{erfc}\left(\frac{z_1-k}{\sqrt{2}\sigma}\right)+\operatorname{erfc}\left(\frac{k-z_1}{\sqrt{2}\sigma}\right)\right)\right)+\frac{\log(1/\delta)}{\alpha-1}.$$

DP-CD: *not needed!*

– I. Mironov, K. Talwar, and L. Zhang, *"Rényi Differential Privacy of the Sampled Gaussian Mechanism"*, 2019.

Utility?

$$\mathbb{E}\big[F(w^T) - F^*\big] \leq \ ?$$

Ad hoc dual norms.

$$\|w\|_M = \sqrt{\sum_{j=1}^{p} M_j w_j^2} \qquad \|w\|_{M^{-1}} = \sqrt{\sum_{j=1}^{p} \frac{1}{M_j} w_j^2}.$$

	DP-SGD	DP-CD
Assumptions on f	convex Λ-Lipschitz β-smooth	convex L-comp.-Lipschitz M-comp.-smooth
Utility ($\mathbb{E}[F(w) - F^*] \leq ...$)		

	DP-SGD	DP-CD
Assumptions on f	convex Λ-Lipschitz β-smooth	convex L-comp.-Lipschitz M-comp.-smooth
Utility ($\mathbb{E}[F(w) - F^*] \leq ...$)	$O\left(\frac{\boldsymbol{\Lambda R_I}\sqrt{p}}{n\epsilon}\right)$	

Where $R_I^2 = \max(F(w^0) - F(w^*), \|w^0 - w^*\|_2^2)$.

	DP-SGD	DP-CD
Assumptions on f	convex Λ-Lipschitz β-smooth	convex L-comp.-Lipschitz M-comp.-smooth
Utility ($\mathbb{E}[F(w) - F^*] \leq ...$)	$O\left(\frac{\boldsymbol{\Lambda R_I}\sqrt{p}}{n\epsilon}\right)$	$O\left(\frac{\boldsymbol{\|L\|_{M^{-1}} R_M}\sqrt{p}}{n\epsilon}\right)$

Where
$$\|L\|_{M^{-1}} = \left(\textstyle\sum_{j=1}^{p} L_j^2/M_j\right)^{1/2},$$
$$R_I^2 = \max(F(w^0) - F(w^*), \|w^0 - w^*\|_2^2),$$
$$R_M^2 = \max(F(w^0) - F(w^*), \|w^0 - w^*\|_M^2).$$

	DP-SGD	DP-CD
Assumptions on f	μ_I-strongly-convex $w.r.t.$ $\|\cdot\|_2$	μ_M-strongly-convex $w.r.t.$ $\|\cdot\|_M$
	Λ-Lipschitz	L-comp.-Lipschitz
	β-smooth	M-comp.-smooth
Utility ($\mathbb{E}[F(w) - F^*] \leq ...$)		

	DP-SGD	DP-CD
Assumptions on f	μ_I-strongly-convex *w.r.t.* $\lVert\cdot\rVert_2$	μ_M-strongly-convex *w.r.t.* $\lVert\cdot\rVert_M$
	Λ-Lipschitz	L-comp.-Lipschitz
	β-smooth	M-comp.-smooth
Utility ($\mathbb{E}[F(w) - F^*] \leq ...$)	$O\left(\frac{\boldsymbol{\Lambda^2} p}{\boldsymbol{\mu_I} n^2 \epsilon^2}\right)$	

	DP-SGD	DP-CD
Assumptions on f	μ_I-strongly-convex *w.r.t.* $\lVert\cdot\rVert_2$	μ_M-strongly-convex *w.r.t.* $\lVert\cdot\rVert_M$
	Λ-Lipschitz	L-comp.-Lipschitz
	β-smooth	M-comp.-smooth
Utility ($\mathbb{E}[F(w) - F^*] \leq \dots$)	$O\left(\frac{\boldsymbol{\Lambda^2} p}{\boldsymbol{\mu_I} n^2 \epsilon^2}\right)$	$O\left(\frac{\boldsymbol{\lVert L\rVert^2_{M^{-1}}} p}{\boldsymbol{\mu_M} n^2 \epsilon^2}\right)$

Where $\lVert L\rVert_{M^{-1}} = (\sum_{j=1}^p L_j^2/M_j)^{1/2}$.

CD vs. SGD: Who wins?

Balanced M_j's:

$\rightarrow$ DP-SGD up to $\boldsymbol{p}$ times better.

CD vs. SGD: Who wins?

Imbalanced M_j's:

$\rightarrow$ DP-CD up to $\frac{\max_j M_j}{\min_j M_j}$ times better.

Practical Comments.

- Gradient Clipping.
- Hyperparameters.
- Private Smoothness Constants?

$$|\nabla_j \ell(w; d)| \leq L_j \quad \text{for } d \in \mathcal{X} \times \mathcal{Y}.$$

$$|\nabla_j \ell(w; d)| \le L_j \quad \text{for } d \in \mathcal{X} \times \mathcal{Y}.$$

$$\Rightarrow \sigma_j^2 = O\left(\frac{\boldsymbol{L_j}^2 T}{n^2 \epsilon^2}\right).$$

Clip!

$$\mathrm{clip}(\nabla_j \ell(w;d), C_{\boldsymbol{j}}) = \begin{cases} \pm C_{\boldsymbol{j}}, & \text{if } |\nabla_j \ell(w;d)| > C_{\boldsymbol{j}}, \\ \nabla_j \ell(w;d), & \text{otherwise.} \end{cases}$$

Clip!

$$\mathrm{clip}(\nabla_j \ell(w;d), C_{\boldsymbol{j}}) = \begin{cases} \pm C_{\boldsymbol{j}}, & \text{if } |\nabla_j \ell(w;d)| > C_{\boldsymbol{j}}, \\ \nabla_j \ell(w;d), & \text{otherwise.} \end{cases}$$

$$|\mathrm{clip}(\nabla_j \ell, C_{\boldsymbol{j}})| \leq C_{\boldsymbol{j}} \Rightarrow \sigma_j = O\left(\frac{\boldsymbol{C_j}^2 T}{n^2 \epsilon^2}\right).$$

	DP-SGD	DP-CD
Clipping	C	
Step sizes	$\eta = \gamma/\beta$	

	DP-SGD	DP-CD
Clipping	C	$C_j = \sqrt{\frac{M_j}{\mathrm{tr}(M)}} C$
Step sizes	$\eta = \gamma / \beta$	$\eta_j = \gamma / M_j$

$$\underset{w \in \mathbb{R}^p}{\arg\min} \overbrace{\frac{1}{n} \sum_{i=1}^{n} \underbrace{\ell(w, d_i)}_{M^{(i)}\text{-comp-smooth}}}^{f(w)} + \psi(w)$$

$$\operatorname*{arg\,min}_{w \in \mathbb{R}^p} \overbrace{\frac{1}{n} \sum_{i=1}^{n} \underbrace{\ell(w, d_i)}_{M^{(i)}\text{-comp-smooth}}}^{f(w)} + \psi(w)$$

$$M_j = \frac{1}{n} \sum_{i=1}^{n} M_j^{(i)}.$$

$$\operatorname*{arg\,min}_{w\in\mathbb{R}^p} \overbrace{\frac{1}{n}\sum_{i=1}^{n} \underbrace{(X_{i,:}^T w - y_i)^2}_{M^{(i)}\text{-comp-smooth}}}^{f(w)} + \lambda \|w\|_1$$

$$M_j = \frac{1}{n}\sum_{i=1}^{n} |X_{i,j}|^2 \, .$$

$$\underset{w \in \mathbb{R}^p}{\arg\min} \overbrace{\frac{1}{n} \sum_{i=1}^{n} \underbrace{(X_{i,:}^T w - y_i)^2}_{M^{(i)}\text{-comp-smooth}}}^{f(w)} + \lambda \|w\|_1$$

$$M_j = \frac{1}{n} \sum_{i=1}^{n} |X_{i,j}|^2 .$$

Remark: standardized data $\rightarrow M_j = 1$.

$$\operatorname*{arg\,min}_{w\in\mathbb{R}^p} \overbrace{\frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell(w,d_i)}_{M^{(i)}\text{-comp-smooth}}}^{f(w)} + \psi(w)$$

Let $\epsilon' \leq \epsilon$ $(e.g.,\ \epsilon' = 0.1\epsilon)$.

$$M_j^{priv} = \frac{1}{n}\sum_{i=1}^{n} M_j^{(i)} + \mathrm{Lap}\left(\frac{p \cdot \boldsymbol{\max_i M_j^{(i)}}}{n\epsilon'}\right).$$

$$\arg\min_{w\in\mathbb{R}^p} \overbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(w)}^{f(w)} + \psi(w)$$

Let $\epsilon' \leq \epsilon$ $(e.g.,\ \epsilon' = 0.1\epsilon)$.

$$M_j^{priv} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{clip}(M_j^{(i)}, \boldsymbol{b_j}) + \mathrm{Lap}\left(\frac{p \cdot \boldsymbol{b_j}}{n\epsilon'}\right).$$

Experiments.

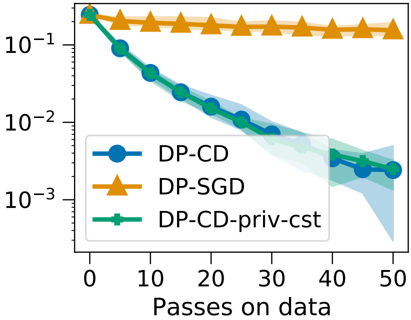
- On logistic regression and LASSO.

- Tune $\left\{ \begin{array}{l} \text{step size,} \\ \text{clipping threshold,} \\ \text{number of iterations.} \end{array} \right.$

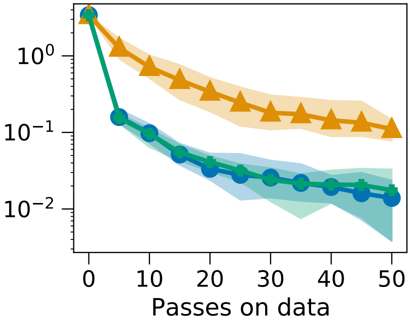
- Average over 10 runs.

Imbalanced Dataset

Imbalanced Dataset



ELECTRICITY (RAW)
$n = 45,312 \quad p = 8$
Logistic Regression

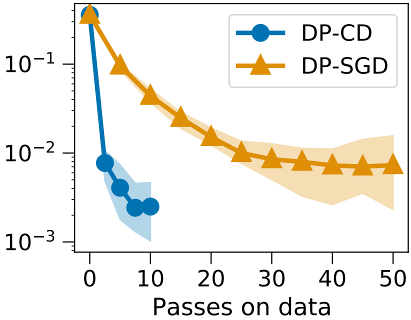


CALIFORNIA (RAW)
$n = 20,640 \quad p = 8$
LASSO

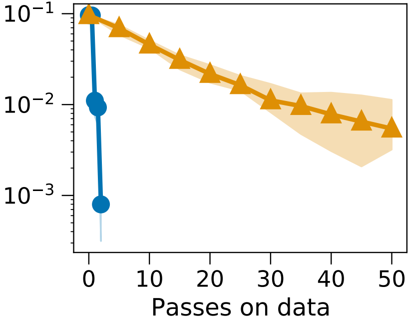
$$\epsilon = 1,\ \delta = 1/n^2.$$

Balanced Dataset

Balanced Dataset



ELECTRICITY (SCALED)
$n = 45,312 \quad p = 8$
Logistic Regression

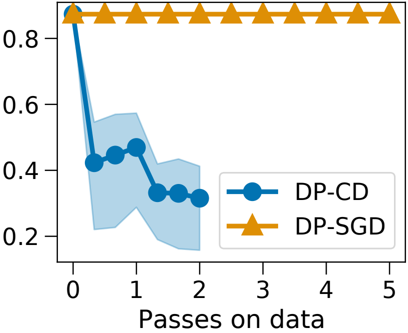


CALIFORNIA (SCALED)
$n = 20,640 \quad p = 8$
LASSO

$$\epsilon = 1,\ \delta = 1/n^2.$$

Higher dimension

Higher dimension



SPARSE SOLUTION

$n = 1,000 \quad p = 1,000$

LASSO

$\epsilon = 10$, $\delta = 1/n^2$ and $\|w^*\|_0 = 10$.

Partial Conclusion.

- Partial gradients without amplification.
- Large learning rates.
- Good practical performance.

Is DP-CD optimal?

Is DP-CD optimal? Kind of.

Loss	Convex L-comp.-Lipschitz	Strongly-Convex L-comp.-Lipschitz
Utility ($\mathbb{E}[F(w) - F^*] \geq ...$)	$\Omega\left(\frac{L_{\min}}{L_{\max}} \frac{\lVert L \rVert_2 \lVert w^* \rVert_2 \sqrt{p}}{n\epsilon}\right)$	$\Omega\left(\frac{L_{\min}^2}{L_{\max}^2} \frac{\lVert L \rVert_2^2 p}{\mu_I n^2 \epsilon^2}\right)$

Is DP-CD optimal? Yes, if...

Loss	Convex L-comp.-Lipschitz	Strongly-Convex L-comp.-Lipschitz
Utility $(\mathbb{E}[F(w) - F^*] \geq ...)$	$\Omega\left(\frac{\|L\|_2 \|w^*\|_2 \sqrt{p}}{n\epsilon}\right)$	$\Omega\left(\frac{\|L\|_2^2 p}{\mu_I n^2 \epsilon^2}\right)$

If $\sum_{j \in \mathcal{S}} L_j^2 = \Omega(\|L\|_2^2)$ for $\mathcal{S}$ with $\operatorname{card} \mathcal{S} \geq \frac{p}{75}$.

Perspectives!

Choose j wisely. (Greedily?)

- Match lower bounds?
- Reduce dependence on p.

- J. Nutini et al., *"Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection"*, 2015.
- S. P. Karimireddy et al., *"Efficient Greedy Coordinate Descent for Composite Problems"*, 2019.
- H. Fang et al., *"Greed Meets Sparsity: Understanding and Improving Greedy Coordinate Descent for Sparse Optimization"*, 2020.

Sparse Approximation?

- After T iterations: $\left\|w^T\right\|_0 \leq T$.
- Active sets?

– M. Massias, A. Gramfort, and J. Salmon, *"Celer: A Fast Solver for the Lasso with Dual Extrapolation"*, 2018.

– K. L. Clarkson, *"Coresets, Sparse Greedy Approximation, and the Frank-Wolfe Algorithm"*, 2010.

DP-CD as a subroutine.

e.g., in Iteratively Reweighted Least Squares

$$w^{t+1} = \arg\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \boldsymbol{\alpha}_{\boldsymbol{i}}^{\boldsymbol{t}} \left| x_i \beta - y_i \right|^2 .$$

– P. W. Holland and R. E. Welsch, *"Robust Regression Using Iteratively Reweighted Least-Squares"*, 1977.

– E. J. Candès, M. B. Wakin, and S. P. Boyd, *"Enhancing Sparsity by Reweighted L1 Minimization"*, 2008.

Coordinate-Wise Clipping.

- Even in DP-SGD!
- Could improve Fairness?

– V. Pichapati et al., *"AdaCliP: Adaptive Clipping for Private SGD"*, 2019.

– D. Xu, W. Du, and X. Wu, *"Removing Disparate Impact of Differentially Private Stochastic Gradient Descent on Model Accuracy"*, 2020.

Thank you! Questions? :)

See our paper:

– P. Mangold et al., *"Differentially Private Coordinate Descent for Composite Empirical Risk Minimization"*, 2021.