On the Benefits of Accelerated Optimization in Robust and Private Estimation

Laurentiu Marchis

University of Cambridge, Department of Pure Mathematics and Mathematical Statistics

StatMathAppli 2025

September 4, 2025

Joint work with Po-Ling Loh

Outline

- Part 1: Frank-Wolfe and differential privacy
- Part 2: Nesterov's momentum and heavy-tailed robustness

Part 1: Frank-Wolfe and differential privacy

Basics of (ϵ, δ)-DP

- We work with datasets $X \in \mathcal{E}^n$ and mechanisms taking values in $\mathbb{R}^p$.

Basics of (ϵ, δ)-DP

- We work with datasets $X \in \mathcal{E}^n$ and mechanisms taking values in $\mathbb{R}^p$.
- Gaussian noise addition: For a function $\mathcal{G}$ with ℓ_2-sensitivity

$$\Delta_{\mathcal{G}} := \sup_{X \sim X'} \left\| \mathcal{G}(X) - \mathcal{G}(X') \right\|_2 ,$$

the mechanism

$$\mathcal{G}(X) + N\left(0, \frac{2\Delta_{\mathcal{G}}^2 \log(2/\delta)}{\epsilon^2} I_p\right)$$

is (ϵ, δ)-DP

Basics of (ϵ, δ)-DP

- We work with datasets $X \in \mathcal{E}^n$ and mechanisms taking values in $\mathbb{R}^p$.
- Gaussian noise addition: For a function $\mathcal{G}$ with ℓ_2-sensitivity

$$\Delta_{\mathcal{G}} := \sup_{X \sim X'} \left\| \mathcal{G}(X) - \mathcal{G}(X') \right\|_2 ,$$

  the mechanism

$$\mathcal{G}(X) + N\left(0, \frac{2\Delta_{\mathcal{G}}^2 \log(2/\delta)}{\epsilon^2} I_p\right)$$

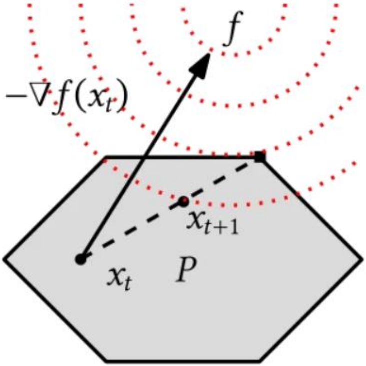
  is (ϵ, δ)-DP
- Recall the advanced composition (Dwork et al. (2010)): For $\epsilon \in (0, 0.9]$, and δ, $T > 0$, the class of

$$\left(\frac{\epsilon}{2\sqrt{2T\log(2/\delta)}}, \frac{\delta}{2T}\right) - DP$$

  mechanisms is (ϵ, δ)-DP under T-fold adaptive composition

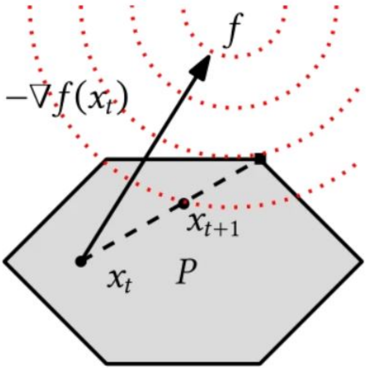
The Frank–Wolfe method: A projection-free approach

- Projected gradient descent requires projections onto the constraint set $\mathcal{C}$



The Frank–Wolfe method: A projection-free approach

- Projected gradient descent requires projections onto the constraint set $\mathcal{C}$
- Problem: computationally expensive for complicated sets $\mathcal{C}$

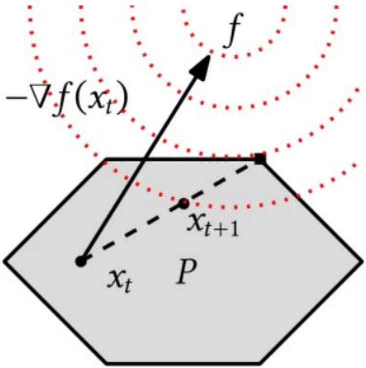


The Frank–Wolfe method: A projection-free approach

- Projected gradient descent requires projections onto the constraint set $\mathcal{C}$
- Problem: computationally expensive for complicated sets $\mathcal{C}$
- Solution: linearize the convex objective f and move towards a minimizer $v_t \in \mathcal{C}$ (e.g. Frank & Wolfe 1956):

$$v_t = \arg\min_{v \in \mathcal{C}} \nabla f(x_t)^T v,$$

$$x_{t+1} = (1 - \eta_t)\, x_t + \eta_t\, v_t$$



Frank–Wolfe (continued)

- Frank–Wolfe has seen renewed interest in ML (Jaggi 2013, Lacoste-Julien & Jaggi 2015, Asi et al. 2021, Raff et al. 2023); the LASSO (constrained form), for instance:

$$\min_{\|\theta\|_1 \le D} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2$$

Frank–Wolfe (continued)

- Frank–Wolfe has seen renewed interest in ML (Jaggi 2013, Lacoste-Julien & Jaggi 2015, Asi et al. 2021, Raff et al. 2023); the LASSO (constrained form), for instance:

$$\min_{\|\theta\|_1 \le D} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2$$

- Recent variants target strongly convex objectives (Li et al. 2020) and alternative geometries, improving convergence

Frank–Wolfe (continued)

- Frank–Wolfe has seen renewed interest in ML (Jaggi 2013, Lacoste-Julien & Jaggi 2015, Asi et al. 2021, Raff et al. 2023); the LASSO (constrained form), for instance:

$$\min_{\|\theta\|_1 \le D} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2$$

- Recent variants target strongly convex objectives (Li et al. 2020) and alternative geometries, improving convergence
- An accelerated Frank–Wolfe for smooth convex f over ℓ_2-balls (more generally, strongly convex sets; Garber & Hazan 2015) yields

$$f(x_t) - f(x_*) \lesssim e^{-\Theta(t)}, \qquad x_* = \arg\min_{x \in \mathcal{C}} f(x)$$

vs. $O(1/t)$ in the classical case

Motivation: (ϵ, δ)-DP ERM via convex optimization

- Problem: privately minimize the ER: $\mathcal{L}(\theta, \mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta, z_i)$

Motivation: (ϵ, δ)-DP ERM via convex optimization

- Problem: privately minimize the ER: $\mathcal{L}(\theta, \mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta, z_i)$
- Privacy-by-noise (Bassily et al. 2014, Jain & Thakurta 2014, Wang et al. 2017, Smith et al. 2017, Cai et al. 2021):

Motivation: (ϵ, δ)-DP ERM via convex optimization

- Problem: privately minimize the ER: $\mathcal{L}(\theta, \mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta, z_i)$
- Privacy-by-noise (Bassily et al. 2014, Jain & Thakurta 2014, Wang et al. 2017, Smith et al. 2017, Cai et al. 2021):

  Gradient perturbation add noise to gradients at each step (most popular)

Motivation: (ϵ, δ)-DP ERM via convex optimization

- Problem: privately minimize the ER: $\mathcal{L}(\theta, \mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta, z_i)$
- Privacy-by-noise (Bassily et al. 2014, Jain & Thakurta 2014, Wang et al. 2017, Smith et al. 2017, Cai et al. 2021):

  Gradient perturbation add noise to gradients at each step (most popular)
- Talwar et al. (2015): private Frank–Wolfe using noisy gradients for minimizing an L_2-Lipschitz, $\beta_{\mathcal{L}}$-smooth loss over a convex set $\mathcal{C}$ (diameter $\|\mathcal{C}\|_2 < \infty$)

Motivation: private Frank-Wolfe

- $v_t = \arg\min_{v \in \mathcal{C}} (\nabla \mathcal{L}(\theta_t, \mathcal{D}_n) + \xi_t)^T v$, and $\xi_t \sim N\left(0, \frac{32 L_2^2 T \log^2(T/\delta)}{n^2 \epsilon^2} I_p\right)$

Motivation: private Frank-Wolfe

- $v_t = \arg\min_{v \in \mathcal{C}} (\nabla \mathcal{L}(\theta_t, \mathcal{D}_n) + \xi_t)^T v$, and $\xi_t \sim N\left(0, \frac{32 L_2^2 T \log^2(T/\delta)}{n^2 \epsilon^2} I_p\right)$
- $\theta_{t+1} = (1 - \eta_t)\theta_t + \eta_t v_t$, and $\eta_t = \frac{2}{2+t}$

Motivation: private Frank-Wolfe

- $v_t = \arg\min_{v \in \mathcal{C}} (\nabla \mathcal{L}(\theta_t, \mathcal{D}_n) + \xi_t)^T v$, and $\xi_t \sim N\left(0, \frac{32 L_2^2 T \log^2(T/\delta)}{n^2 \epsilon^2} I_p\right)$
- $\theta_{t+1} = (1 - \eta_t)\theta_t + \eta_t v_t$, and $\eta_t = \frac{2}{2+t}$
- Running for $T \asymp \left(\frac{\beta_{\mathcal{L}} ||\mathcal{C}||_2^2 n\epsilon}{L_2 G_{\mathcal{C}}}\right)^{2/3}$ gives

$$\mathbb{E}\left[\mathcal{L}(\theta_T, \mathcal{D}_n) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta, \mathcal{D}_n)\right] = \tilde{O}\left(\frac{\beta_{\mathcal{L}}^{1/3} (||\mathcal{C}||_2 L_2 G_{\mathcal{C}})^{2/3}}{(n\epsilon)^{2/3}}\right),$$

where $G_{\mathcal{C}} = \mathbb{E}_{b \sim N(0, I_p)}\left[\sup_{\theta \in \mathcal{C}} \theta^T b\right]$

Motivation: continued

- Can we get better rates and iteration counts T? If we can, then how?

Motivation: continued

- Can we get better rates and iteration counts T? If we can, then how?
- Our method: Run the private Frank-Wolfe method with an adequate choice of η (independent of t and see later) and optimize over a strongly convex set:

Motivation: continued

- Can we get better rates and iteration counts T? If we can, then how?
- Our method: Run the private Frank-Wolfe method with an adequate choice of η (independent of t and see later) and optimize over a strongly convex set:
  - $v_t = \arg\min_{v \in \mathcal{C}} (\nabla \mathcal{L}(\theta_t, \mathcal{D}_n) + \xi_t)^T v$, and $\xi_t \sim N\left(0, \frac{32 L_2^2 T \log^2(T/\delta)}{n^2 \epsilon^2} I_p\right)$

Motivation: continued

- Can we get better rates and iteration counts T? If we can, then how?
- Our method: Run the private Frank-Wolfe method with an adequate choice of η (independent of t and see later) and optimize over a strongly convex set:
  - $v_t = \arg\min_{v \in \mathcal{C}} (\nabla \mathcal{L}(\theta_t, \mathcal{D}_n) + \xi_t)^T v$, and $\xi_t \sim N\left(0, \frac{32 L_2^2 T \log^2(T/\delta)}{n^2 \epsilon^2} I_p\right)$
  - $\theta_{t+1} = (1-\eta)\theta_t + \eta v_t$

Analysis of accelerated Frank-Wolfe

- Our approach: Develop a general optimization framework to allow noise, tailor the learning rate η to accelerate, and take more advantage of the geometry of $\mathcal{C}$

Analysis of accelerated Frank-Wolfe

- Our approach: Develop a general optimization framework to allow noise, tailor the learning rate η to accelerate, and take more advantage of the geometry of $\mathcal{C}$
- Relaxed and accelerated Frank-Wolfe:

$$x_{t+1} = (1-\eta)x_t + \eta v_t,$$

where $v_t \in \mathcal{C}$ s.t. $v_t^T \nabla f(x_t) \leq \min_{v \in \mathcal{C}} v^T \nabla f(x_t) + \Delta$

Relaxed and accelerated Frank-Wolfe

Theorem 1 (Marchis and Loh 2025)

Let $\mathcal{C} = \mathbb{B}_2(D)$ and f a convex, β_f-smooth function such that $0 < r \leq ||\nabla f(x)||_2$ for all $x \in \mathcal{C}$. Then, for $t \geq 1$,

$$f(x_t) - f(x_*) \leq c^t \left(f(x_0) - f(x_*)\right) + \frac{3\Delta\eta}{2(1-c)},$$

where $x_* \in \arg\min_{x \in \mathcal{C}} f(x)$, $c = \max\left\{\frac{1}{2}, 1 - \frac{r}{8D\beta_f}\right\}$, and $\eta = \min\left\{1, \frac{r}{4D\beta_f}\right\}$.

- Modification of arguments from Garber and Hazan 2015

Relaxed and accelerated Frank-Wolfe

Theorem 1 (Marchis and Loh 2025)

Let $\mathcal{C} = \mathbb{B}_2(D)$ and f a convex, β_f-smooth function such that $0 < r \leq ||\nabla f(x)||_2$ for all $x \in \mathcal{C}$. Then, for $t \geq 1$,

$$f(x_t) - f(x_*) \leq c^t \left(f(x_0) - f(x_*)\right) + \frac{3\Delta\eta}{2(1-c)},$$

where $x_* \in \arg\min_{x \in \mathcal{C}} f(x)$, $c = \max\left\{\frac{1}{2}, 1 - \frac{r}{8D\beta_f}\right\}$, and $\eta = \min\left\{1, \frac{r}{4D\beta_f}\right\}$.

- Modification of arguments from Garber and Hazan 2015
- More generally, one can take $\mathcal{C}$ to be strongly convex

Relaxed and accelerated Frank-Wolfe

Theorem 1 (Marchis and Loh 2025)

Let $\mathcal{C} = \mathbb{B}_2(D)$ and f a convex, β_f-smooth function such that $0 < r \leq ||\nabla f(x)||_2$ for all $x \in \mathcal{C}$. Then, for $t \geq 1$,

$$f(x_t) - f(x_*) \leq c^t \left(f(x_0) - f(x_*)\right) + \frac{3\Delta\eta}{2(1-c)},$$

where $x_* \in \arg\min_{x \in \mathcal{C}} f(x)$, $c = \max\left\{\frac{1}{2}, 1 - \frac{r}{8D\beta_f}\right\}$, and $\eta = \min\left\{1, \frac{r}{4D\beta_f}\right\}$.

- Modification of arguments from Garber and Hazan 2015
- More generally, one can take $\mathcal{C}$ to be strongly convex
- Next, we use this in the context of privacy

Private and accelerated Frank-Wolfe for distribution-free data

- Squared error loss $\mathcal{L}(\theta, (x_i, y_i)) = \frac{1}{2}(y_i - x_i^T\theta)^2$, $|y_i|, ||x_i||_\infty \leq 1$

Private and accelerated Frank-Wolfe for distribution-free data

- Squared error loss $\mathcal{L}(\theta, (x_i, y_i)) = \frac{1}{2}(y_i - x_i^T\theta)^2$, $|y_i|, ||x_i||_\infty \leq 1$
- $\mathcal{C} = \mathbb{B}_2(D)$, i.e. ridge regression

Private and accelerated Frank-Wolfe for distribution-free data

- Squared error loss $\mathcal{L}(\theta, (x_i, y_i)) = \frac{1}{2}(y_i - x_i^T\theta)^2$, $|y_i|, ||x_i||_\infty \leq 1$
- $\mathcal{C} = \mathbb{B}_2(D)$, i.e. ridge regression
- Note $\beta_{\mathcal{L}} = \frac{1}{n}\left\|\sum_{i=1}^n x_i x_i^T\right\|_2 \leq p$, $L_2 \asymp \sqrt{p} + pD$

Private and accelerated Frank-Wolfe for distribution-free data

- Squared error loss $\mathcal{L}(\theta, (x_i, y_i)) = \frac{1}{2}(y_i - x_i^T\theta)^2$, $|y_i|, ||x_i||_\infty \leq 1$
- $\mathcal{C} = \mathbb{B}_2(D)$, i.e. ridge regression
- Note $\beta_{\mathcal{L}} = \frac{1}{n}\left\|\sum_{i=1}^n x_i x_i^T\right\|_2 \leq p$, $L_2 \asymp \sqrt{p} + pD$
- Assume $\inf\limits_{\theta \in \mathcal{C}} \frac{||\nabla\mathcal{L}(\theta, \mathcal{D}_n)||_2}{D\beta_{\mathcal{L}}} \gtrsim 1$ (can be satisfied with an appropriate linear model), we get for $T \asymp \log(n)$ that

$$\mathbb{E}\left[\mathcal{L}(\theta_T, \mathcal{D}_n) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta, \mathcal{D}_n)\right] = \widetilde{O}\left(\frac{(\sqrt{p} + pD)D\sqrt{p}}{n\epsilon}\right)$$

Comparisons

- For squared error loss, $|y_i|, ||x_i||_\infty \leq 1$, and $\mathcal{C} = \mathbb{B}_2(D)$, Talwar et al. 2015 gives a rate of $\widetilde{O}\left(\left(\frac{(\sqrt{p}+pD)D^2 p}{n\epsilon}\right)^{2/3}\right)$, for $T \asymp \left(\frac{n\epsilon D}{1+2\sqrt{p}D}\right)^{2/3}$

Comparisons

- For squared error loss, $|y_i|, ||x_i||_\infty \leq 1$, and $\mathcal{C} = \mathbb{B}_2(D)$, Talwar et al. 2015 gives a rate of $\widetilde{O}\left(\left(\frac{(\sqrt{p}+pD)D^2 p}{n\epsilon}\right)^{2/3}\right)$, for $T \asymp \left(\frac{n\epsilon D}{1+2\sqrt{p}D}\right)^{2/3}$
- Minimax optimality:

Comparisons

- For squared error loss, $|y_i|, ||x_i||_\infty \leq 1$, and $\mathcal{C} = \mathbb{B}_2(D)$, Talwar et al. 2015 gives a rate of $\widetilde{O}\left(\left(\frac{(\sqrt{p}+pD)D^2p}{n\epsilon}\right)^{2/3}\right)$, for $T \asymp \left(\frac{n\epsilon D}{1+2\sqrt{p}D}\right)^{2/3}$
- Minimax optimality:
  - For $D \asymp \frac{1}{\sqrt{p}}$, our acceleration improves the rate of $\left(\frac{\sqrt{p}}{n\epsilon}\right)^{2/3}$ to $\frac{\sqrt{p}}{n\epsilon}$ and the iteration count T from $\left(\frac{n\epsilon}{\sqrt{p}}\right)^{2/3}$ to $\log(n)$

Comparisons

- For squared error loss, $|y_i|, ||x_i||_\infty \leq 1$, and $\mathcal{C} = \mathbb{B}_2(D)$, Talwar et al. 2015 gives a rate of $\widetilde{O}\left(\left(\frac{(\sqrt{p}+pD)D^2 p}{n\epsilon}\right)^{2/3}\right)$, for $T \asymp \left(\frac{n\epsilon D}{1+2\sqrt{p}D}\right)^{2/3}$
- Minimax optimality:
  - For $D \asymp \frac{1}{\sqrt{p}}$, our acceleration improves the rate of $\left(\frac{\sqrt{p}}{n\epsilon}\right)^{2/3}$ to $\frac{\sqrt{p}}{n\epsilon}$ and the iteration count T from $\left(\frac{n\epsilon}{\sqrt{p}}\right)^{2/3}$ to $\log(n)$
  - For $n \asymp \frac{p^{3/2}}{\log(p)}$ we can show the optimality of our accelerated method

GLMs: setup

- Data: $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ are i.i.d. from P_{θ^*}:

$$P_{\theta^*}(y|x) \propto \exp\left(\frac{yx^T\theta^* - \Phi(x^T\theta^*)}{c(\sigma)}\right), \quad ||x||_2, |y| \lesssim 1, \quad \mathbb{E}[xx^T] \succ 0$$

GLMs: setup

- Data: $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ are i.i.d. from P_{θ^*}:

$$P_{\theta^*}(y|x) \propto \exp\left(\frac{yx^T\theta^* - \Phi(x^T\theta^*)}{c(\sigma)}\right), \quad ||x||_2, |y| \lesssim 1, \quad \mathbb{E}[xx^T] \succ 0$$

- Mild assumptions on Φ: $\log(1 + e^z)$ satisfies these (logistic regression)

GLMs: setup

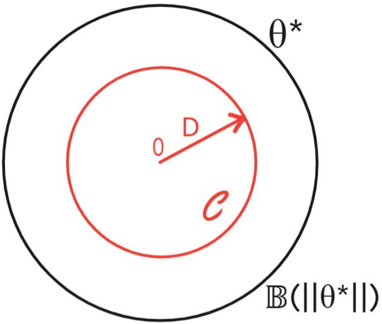
- Data: $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ are i.i.d. from P_{θ^*}:

$$P_{\theta^*}(y|x) \propto \exp\left(\frac{yx^T\theta^* - \Phi(x^T\theta^*)}{c(\sigma)}\right), \quad ||x||_2, |y| \lesssim 1, \quad \mathbb{E}[xx^T] \succ 0$$

- Mild assumptions on Φ: $\log(1+e^z)$ satisfies these (logistic regression)
- Privately minimize the negative log-likelihood loss

$$\mathcal{L}(\theta, \mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i^T\theta) - y_i x_i^T\theta \quad \text{over} \quad \mathcal{C} = \mathbb{B}_2(D)$$

GLMs: main results



- Vanishing bias: For $\|\theta^*\|_2 - D \asymp \frac{1}{n^{2/5}}$ and $T = \widetilde{\Theta}\left(n^{2/5}\right)$,

$$\cdot\, \mathcal{L}(\theta_T, \mathcal{D}_n) - \min_{\theta \in \mathbb{B}_2(||\theta^*||_2)} \mathcal{L}(\theta, \mathcal{D}_n) = \widetilde{O}\left(\frac{1}{n^{4/5}\epsilon}\right)$$

$$\cdot\, ||\theta_T - \theta^*||_2 = \widetilde{O}\left(\frac{1}{n^{1/2}} + \frac{1}{n^{2/5}\epsilon^{1/2}}\right), w.h.p.$$

GLMs: comparisons to Talwar et al. (2015)

- For $\epsilon \gtrsim n^{-2/5}$,

GLMs: comparisons to Talwar et al. (2015)

- For $\epsilon \gtrsim n^{-2/5}$,
  - acceleration $\frac{1}{n^{4/5}\epsilon}$ outperforms the $\frac{1}{(n\epsilon)^{2/3}}$ rate

GLMs: comparisons to Talwar et al. (2015)

- For $\epsilon \gtrsim n^{-2/5}$,
  - acceleration $\frac{1}{n^{4/5}\epsilon}$ outperforms the $\frac{1}{(n\epsilon)^{2/3}}$ rate
  - the iteration count T improves: $n^{2/5}$ vs. $(n\epsilon)^{2/3}$

Part 2: Nesterov's momentum and heavy-tailed robustness

Heavy-tailed robustness: setup & motivation

- Data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ i.i.d. from

$$y = x^T\theta^* + w, \quad w \perp\!\!\!\perp x, \quad \mathbb{E}[w] = 0, \quad \mathbb{E}[w^2] \asymp 1,$$

with $\mathbb{E}[x] = 0$, $\mathbb{E}[xx^T] \succ 0$, and mild moment assumptions on x

Heavy-tailed robustness: setup & motivation

- Data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ i.i.d. from

$$y = x^T \theta^* + w, \quad w \perp\!\!\!\perp x, \quad \mathbb{E}[w] = 0, \quad \mathbb{E}[w^2] \asymp 1,$$

  with $\mathbb{E}[x] = 0$, $\mathbb{E}[xx^T] \succ 0$, and mild moment assumptions on x
- Goal: estimate θ^* robustly under heavy tails

Heavy-tailed robustness: setup & motivation

- Data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ i.i.d. from

$$y = x^T \theta^* + w, \quad w \perp\!\!\!\perp x, \quad \mathbb{E}[w] = 0, \quad \mathbb{E}[w^2] \asymp 1,$$

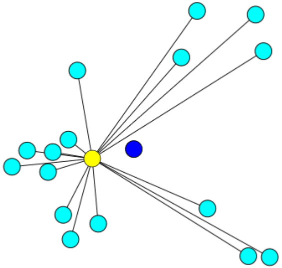
  with $\mathbb{E}[x] = 0$, $\mathbb{E}[xx^T] \succ 0$, and mild moment assumptions on x
- Goal: estimate θ^* robustly under heavy tails
- Naive GD:

$$\theta_{t+1} = \theta_t - \frac{\eta}{n} \sum_{i=1}^{n} (x_i^T \theta_t - y_i) x_i,$$

  but sample-mean gradients can be suboptimal without sub-Gaussian tails

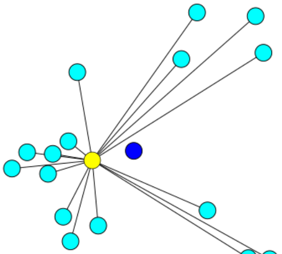
Geometric median of means (G_{MOM})

- Use a "high-dimensional median":



Geometric median of means (G_{MOM})

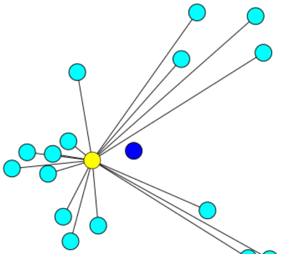
- Use a "high-dimensional median":
- Partition gradients $\{(x_i^T\theta - y_i)x_i\}_{i=1}^n$ into b buckets, compute bucket means $\{\widehat{\mu}_j(\theta)\}_{j=1}^b$



Geometric median of means (G_{MOM})

- Use a "high-dimensional median":
- Partition gradients $\{(x_i^T\theta - y_i)x_i\}_{i=1}^n$ into b buckets, compute bucket means $\{\widehat{\mu}_j(\theta)\}_{j=1}^b$
- Define the geometric median of means:

$$g(\theta) = \arg\min_{\mu} \sum_{j=1}^{b} \|\mu - \widehat{\mu}_j(\theta)\|_2$$

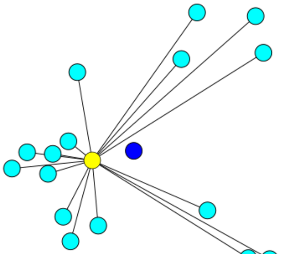


Geometric median of means (G_{MOM})

- Use a "high-dimensional median":
- Partition gradients $\{(x_i^T\theta - y_i)x_i\}_{i=1}^n$ into b buckets, compute bucket means $\{\widehat{\mu}_j(\theta)\}_{j=1}^b$
- Define the geometric median of means:

$$g(\theta) = \arg\min_{\mu} \sum_{j=1}^{b} \|\mu - \widehat{\mu}_j(\theta)\|_2$$

- Gives sub-Gaussian-type concentration under few moments (Minsker 2015)



Robust gradient descent (GD)

- Replace sample mean by $g(\theta)$:

$$\theta_{t+1} = \theta_t - \eta\, g(\theta_t)$$

Robust gradient descent (GD)

- Replace sample mean by $g(\theta)$:

$$\theta_{t+1} = \theta_t - \eta\, g(\theta_t)$$

- Let $\tau_\ell = \lambda_{\min}(\mathbb{E}[xx^T])$, $\tau_u = \lambda_{\max}(\mathbb{E}[xx^T])$ be constants

Robust gradient descent (GD)

- Replace sample mean by $g(\theta)$:

$$\theta_{t+1} = \theta_t - \eta\, g(\theta_t)$$

- Let $\tau_\ell = \lambda_{\min}(\mathbb{E}[xx^T])$, $\tau_u = \lambda_{\max}(\mathbb{E}[xx^T])$ be constants
- Prasad et al. (2020): for $\eta = \frac{2}{\tau_u + \tau_\ell}$ and $n \gtrsim Tp \log(T/\zeta)$,

$$\|\theta_t - \theta^*\|_2 \lesssim k^t + \sqrt{\frac{pT \log(T/\zeta)}{n}}, \qquad \forall t \in [T],$$

w.p. $\geq 1 - \zeta$, for some $k < \frac{\tau_u}{\tau_u + \tau_\ell}$

Robust gradient descent (GD)

- Replace sample mean by $g(\theta)$:

$$\theta_{t+1} = \theta_t - \eta\, g(\theta_t)$$

- Let $\tau_\ell = \lambda_{\min}(\mathbb{E}[xx^T])$, $\tau_u = \lambda_{\max}(\mathbb{E}[xx^T])$ be constants
- Prasad et al. (2020): for $\eta = \frac{2}{\tau_u + \tau_\ell}$ and $n \gtrsim Tp \log(T/\zeta)$,

$$\|\theta_t - \theta^*\|_2 \lesssim k^t + \sqrt{\frac{pT \log(T/\zeta)}{n}}, \qquad \forall t \in [T],$$

w.p. $\geq 1 - \zeta$, for some $k < \frac{\tau_u}{\tau_u + \tau_\ell}$

- Nesterov's momentum:

$$\theta_{t+1} = \theta_t + \lambda(\theta_t - \theta_{t-1}) - \eta\, g\big(\theta_t + \lambda(\theta_t - \theta_{t-1})\big)$$

Nesterov's acceleration (AGD) & comparisons

Theorem (Marchis & Loh, 2025)

If $1 < \frac{\tau_u}{\tau_\ell} < 1.76$ and $n \gtrsim Tp\log(T/\zeta)$, with $\eta = \frac{2}{\tau_u}$ and $\lambda = \frac{\sqrt{\tau_u} - \sqrt{\tau_\ell}}{\sqrt{\tau_u} + \sqrt{\tau_\ell}}$, then w.p.$\geq 1 - \zeta$,

$$\|\theta_t - \theta^*\|_2 \lesssim \left(1 - \sqrt{\frac{\tau_\ell}{\tau_u}}\right)^{t/2} + \sqrt{\frac{pT\log(T/\zeta)}{n}} \quad \forall t \in [T].$$

- Can show $k > (1 - \sqrt{\tau_\ell/\tau_u})^{1/2}$

Nesterov's acceleration (AGD) & comparisons

Theorem (Marchis & Loh, 2025)

If $1 < \frac{\tau_u}{\tau_\ell} < 1.76$ and $n \gtrsim Tp \log(T/\zeta)$, with $\eta = \frac{2}{\tau_u}$ and $\lambda = \frac{\sqrt{\tau_u} - \sqrt{\tau_\ell}}{\sqrt{\tau_u} + \sqrt{\tau_\ell}}$, then w.p.$\geq 1 - \zeta$,

$$\|\theta_t - \theta^*\|_2 \lesssim \left(1 - \sqrt{\frac{\tau_\ell}{\tau_u}}\right)^{t/2} + \sqrt{\frac{pT \log(T/\zeta)}{n}} \quad \forall t \in [T].$$

- Can show $k > (1 - \sqrt{\tau_\ell/\tau_u})^{1/2}$
- Hence AGD reduces the contraction parameter while keeping the statistical error the same

Contributions

Private estimation

- Acceleration results in smaller iteration counts, giving faster rates and requiring less noise for privacy

Heavy-tailed robustness

Contributions

Private estimation

- Acceleration results in smaller iteration counts, giving faster rates and requiring less noise for privacy
- For a lower bound on the gradient and an ℓ_2-ball, Frank-Wolfe can achieve $T \asymp \log(n)$

Heavy-tailed robustness

Contributions

Private estimation

- Acceleration results in smaller iteration counts, giving faster rates and requiring less noise for privacy
- For a lower bound on the gradient and an ℓ_2-ball, Frank-Wolfe can achieve $T \asymp \log(n)$
- The growing constraint set for GLMs gives the benefits of acceleration, while removing the bias as $n \to \infty$.

Heavy-tailed robustness

Contributions

Private estimation

- Acceleration results in smaller iteration counts, giving faster rates and requiring less noise for privacy
- For a lower bound on the gradient and an ℓ_2-ball, Frank-Wolfe can achieve $T \asymp \log(n)$
- The growing constraint set for GLMs gives the benefits of acceleration, while removing the bias as $n \to \infty$.

Heavy-tailed robustness

- Using G_{MOM} results in sub-Gaussian concentration of gradients

Contributions

Private estimation

- Acceleration results in smaller iteration counts, giving faster rates and requiring less noise for privacy
- For a lower bound on the gradient and an ℓ_2-ball, Frank-Wolfe can achieve $T \asymp \log(n)$
- The growing constraint set for GLMs gives the benefits of acceleration, while removing the bias as $n \to \infty$.

Heavy-tailed robustness

- Using G_{MOM} results in sub-Gaussian concentration of gradients
- Acceleration via Nesterov improves the exponential decay with t

Extras & reference

- We also study Frank-Wolfe + heavy-tailed robustenss and Nesterov + privacy
- Laurentiu Marchis & Po-Ling Loh (2025). On the benefits of accelerated optimization in robust and private estimation. arXiv preprint arXiv:2506.03044.

Thank you!