UNIVERSITY OF TWENTE.

Non-Euclidean Differentially Private Stochastic Convex Optimization

Cristóbal Guzmán
Statistics Group, DAMUT, University of Twente
c.guzman@utwente.nl

École Polytechnique, Paris, France
April 21, 2022

Joint work with

Raef Bassily
Ohio State U

Michael Menart
Ohio State U

Anupama Nandi
Ohio State U

- *Non-Euclidean Differentially Private Stochastic Convex Optimization.* **COLT 2021**
- *Differentially Private Stochastic Optimization: New Results in Convex and Non-Convex Settings.* **NeurIPS 2021**

Privacy in Data Analysis

Massive datasets are a key element of current technological revolution

Massive datasets are a key element of current technological revolution

Datasets often contain **sensitive user data**

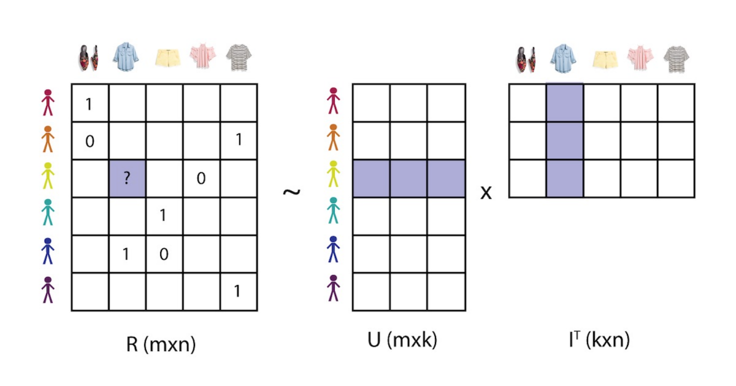
Massive datasets are a key element of current technological revolution

Datasets often contain **sensitive user data**

Q: How to **learn** from data without infringing users' **privacy**?

Privacy Attacks: The Netflix Case

- *Netflix Prize* competition, US$1,000,000 (2006-09)
- **Goal:** based on historical user scores, provide movie recommendations for users
- **Data:** $100,480,507$ ratings by $\sim 500,000$ users on $\sim 18,000$ movies

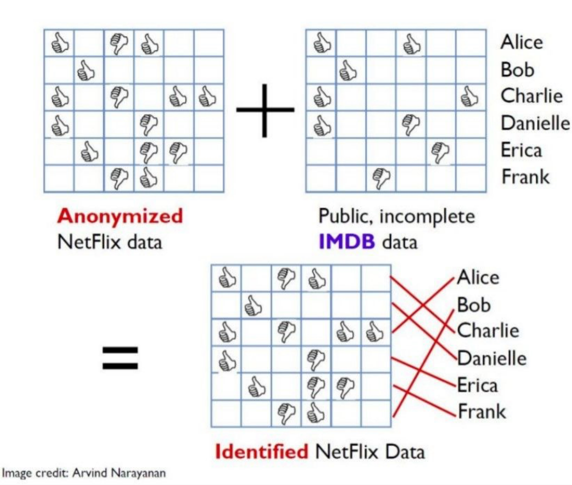

1
0
1
?
0
1
1
0
1
~
x
R (mxn)
U (mxk)
I^T (kxn)

Privacy Attacks: The Netflix Case (cont'd)

- Anonymized data, *in full accordance with the law*
- Narayanan and Shmatikov, 2008 showed how cross references with (public) IMDB exposed the identity of Netflix users

Privacy Attacks: The Netflix Case (cont'd)

- Anonymized data, *in full accordance with the law*
- Narayanan and Shmatikov, 2008 showed how cross references with (public) IMDB exposed the identity of Netflix users

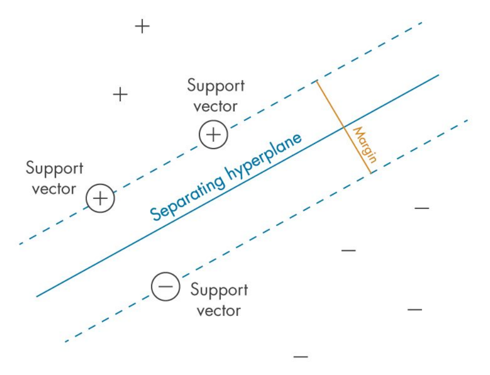


Privacy in ML Models

- Perhaps releasing a private dataset is difficult
- But what about *models*?

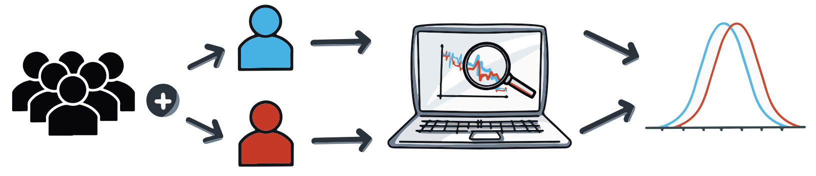
Privacy in ML Models

- Perhaps releasing a private dataset is difficult
- But what about *models*?
- Even more modest ML models (SVM, linear regression, etc.) can suffer from privacy risks



Differential Privacy

Differential Privacy (DP)

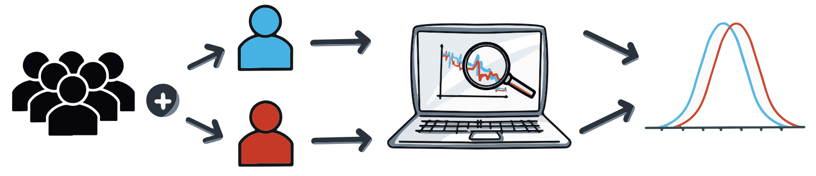


Definition [Differential Privacy (DP)]

Two datasets $S = (z_1, \ldots, z_n)$ and $S' = (z'_1, \ldots, z'_n)$ in $\mathcal{Z}^n$ are *neighbors* (denoted $S \simeq S'$) iff

There exists at most one $i \in [n]$ s.t. $z_i \neq z'_i$

Differential Privacy (DP)



Definition [Differential Privacy (DP)]

Two datasets $S = (z_1, \ldots, z_n)$ and $S' = (z'_1, \ldots, z'_n)$ in $\mathcal{Z}^n$ are *neighbors* (denoted $S \simeq S'$) iff

There exists at most one $i \in [n]$ s.t. $z_i \neq z'_i$

Randomized algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \mathcal{X}$ is (ε, δ)-differentially private if

$$\mathbb{P}(\mathcal{A}(S) \in E) \leq e^{\varepsilon} \cdot \mathbb{P}(\mathcal{A}(S') \in E) + \delta \qquad (\forall S \simeq S')(\forall E \subseteq \mathcal{X})$$

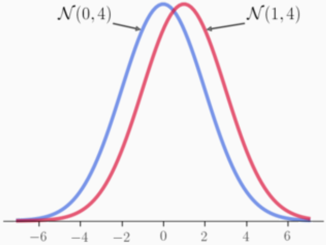
The Gaussian Mechanism

Start with a deterministic algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbb{R}^d$

e.g., empirical mean $\mathcal{A}(S) = \frac{1}{n}\sum_{i=1}^n z_i$

Gaussian Mechanism

- Hypothesis: ℓ_2-sensitivity $\|\mathcal{A}(S) - \mathcal{A}(S')\|_2 \leq \Delta_2$
- Mechanism: $\mathcal{A}_{\text{Gauss}}(S) \sim \mathcal{N}(\mathcal{A}(S), \sigma^2)$
- Guarantee: (ε, δ)-DP (for $\sigma^2 = O(\Delta_2^2 \ln(1/\delta)/\varepsilon^2)$)



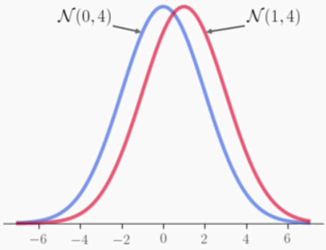
The Gaussian Mechanism

Start with a deterministic algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbb{R}^d$

e.g., empirical mean $\mathcal{A}(S) = \frac{1}{n}\sum_{i=1}^{n} z_i$

Gaussian Mechanism

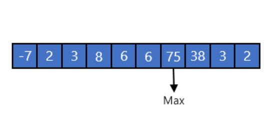
- Hypothesis: ℓ_2-sensitivity
  $\|\mathcal{A}(S) - \mathcal{A}(S')\|_2 \leq \Delta_2$
- Mechanism:
  $\mathcal{A}_{\text{Gauss}}(S) \sim \mathcal{N}(\mathcal{A}(S), \sigma^2)$
- Guarantee: (ε, δ)-DP
  (for $\sigma^2 = O(\Delta_2^2 \ln(1/\delta)/\varepsilon^2)$)



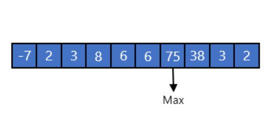
Note: Error of GM, $\mathbb{E}\|\mathcal{A}(S) - \mathcal{A}_{\text{Gauss}}(S)\|_2 = \Theta(\sqrt{d}\sigma)$

Differentially Private Selection

Goal: Select the largest element from an array

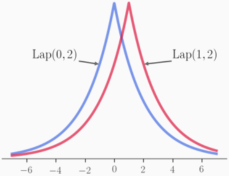


Goal: Select the largest element from an array



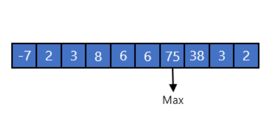
Report Noisy Max Mechanism

- Hypothesis: ℓ_∞-sensitivity $\|\mathcal{A}(S) - \mathcal{A}(S')\|_\infty \leq \Delta_\infty$
- Guarantee: $(\varepsilon, 0)$-DP



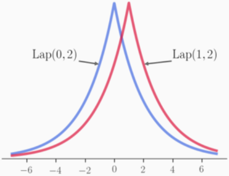
Differentially Private Selection

Goal: Select the largest element from an array



Report Noisy Max Mechanism

- Hypothesis: ℓ_∞-sensitivity $\|\mathcal{A}(S) - \mathcal{A}(S')\|_\infty \leq \Delta_\infty$
- Guarantee: $(\varepsilon, 0)$-DP

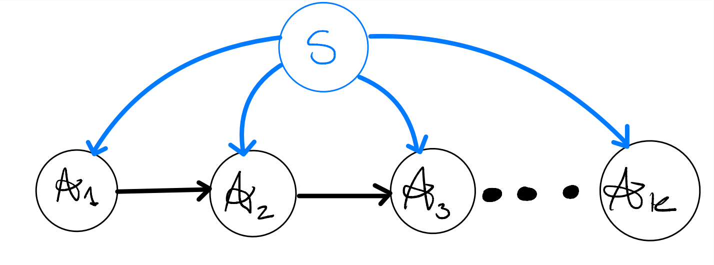


- Mechanism: $\mathcal{A}_{\text{RNM}}(S) = \arg\max_{j \in [d]} \left\{ \mathcal{A}_j(S) + \text{Lap}(0, \Delta_\infty/\varepsilon) \right\}$
- Accuracy: w.h.p. $|\mathcal{A}_{\text{RNM}}(S) - \max_{j \in [d]} \mathcal{A}(S)| = O\left(\frac{\Delta_\infty \ln d}{\varepsilon}\right)$

Composition in Differential Privacy

- Let $\mathcal{A}_1(S), \mathcal{A}_2(S, a_1), \ldots, \mathcal{A}_k(S, a_{k-1})$ mechanisms that are (ε, δ)-DP w.r.t. their first input
- Define inductively, $\mathcal{B}_1 = \mathcal{A}_1$, and

$$\mathcal{B}_j(S) = \mathcal{A}_j(S, \mathcal{B}_{j-1}(S)) \quad (\forall j = 2, \ldots, k)$$



Composition in Differential Privacy

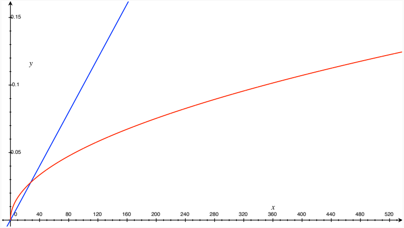
- Let $\mathcal{A}_1(S), \mathcal{A}_2(S, a_1), \ldots, \mathcal{A}_k(S, a_{k-1})$ mechanisms that are (ε, δ)-DP w.r.t. their first input
- Define inductively, $\mathcal{B}_1 = \mathcal{A}_1$, and

$$\mathcal{B}_j(S) = \mathcal{A}_j(S, \mathcal{B}_{j-1}(S)) \quad (\forall j = 2, \ldots, k)$$

Theorem (Basic Composition)

$(\mathcal{B}_1, \ldots, \mathcal{B}_k)$ is $(k\varepsilon, k\delta)$-DP

Composition in Differential Privacy



source: J. Ullman lecture notes

Theorem (Basic Composition)

$(\mathcal{B}_1, \ldots, \mathcal{B}_k)$ is $(k\varepsilon, k\delta)$-DP

Theorem (Advanced Composition) [Dwork, Rothblum & Vadhan:'10]

If $k < 1/\varepsilon^2$. Then for any $0 < \delta' \leq 1$, $(\mathcal{B}_1, \ldots, \mathcal{B}_k)$ is

$$\left(O(\varepsilon\sqrt{k \ln(1/\delta')}), k\delta + \delta'\right)\text{-DP}$$

Stochastic Convex Optimization

Stochastic Convex Optimization (SCO)

$$\textbf{(SCO)} \quad \min_{x \in \mathcal{X}} \{F_{\mathcal{D}}(x) := \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[f(x, \mathbf{z})]\} = F^*_{\mathcal{D}}$$

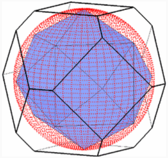
- $(\mathbb{R}^d, \|\cdot\|)$: d-dimensional normed space
- $\mathcal{X} \subseteq \mathcal{B}_{\|\cdot\|}(0, D)$, compact and convex
- $\mathcal{Z}$ any set
- $\mathcal{D}$ probability distribution supported on $\mathcal{Z}$

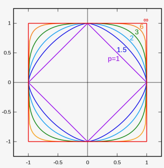
Stochastic Convex Optimization (SCO)

$$\textbf{(SCO)} \quad \min_{x \in \mathcal{X}} \{ F_{\mathcal{D}}(x) := \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[f(x, \mathbf{z})] \} = F_{\mathcal{D}}^*$$

- $(\mathbb{R}^d, \|\cdot\|)$: d-dimensional normed space
- $\mathcal{X} \subseteq \mathcal{B}_{\|\cdot\|}(0, D)$, compact and convex
- $\mathcal{Z}$ any set
- $\mathcal{D}$ probability distribution supported on $\mathcal{Z}$
- Convex loss $f(\cdot, z)$
  - L_0-Lipschitz: $|f(x,z) - f(y,z)| \le L_0 \|x - y\|$
  - L_1-Lipschitz gradient: $\|\nabla f(x,z) - \nabla f(y,z)\|_* \le L_1 \|x - y\|$

Recall dual norm: $\|w\|_* = \sup_{\|x\| \le 1} \langle w, x \rangle$, *and dual norm of* $\|\cdot\|_p$ *is* $\|\cdot\|_q$ ($1/p + 1/q = 1$)





Stochastic Convex Optimization (SCO)

(SCO) $\min_{x \in \mathcal{X}} \{F_{\mathcal{D}}(x) := \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[f(x, \mathbf{z})]\} = F_{\mathcal{D}}^*$

Excess Risk: Given data $\mathrm{S} = (\mathbf{z}_1, \ldots, \mathbf{z}_n) \overset{i.i.d.}{\sim} \mathcal{D}^n$
Does there exist an algorithm $\mathcal{A} : \bigcup_n \mathcal{Z}^n \mapsto \mathcal{X}$ s.t.

$$\underbrace{\mathbb{E}_{\mathcal{A}} \mathbb{E}_{\mathbf{S} \sim \mathcal{D}^n} \left[F_{\mathcal{D}}(\mathcal{A}(\mathbf{S})) - F_{\mathcal{D}}^* \right]}_{\textit{excess (population) risk}} \overset{n \to \infty}{\longrightarrow} 0$$

Stochastic Convex Optimization (SCO): Excess Risk Rates

- ℓ_p-setup: $\|\cdot\| = \|\cdot\|_p$, $1 \le p \le \infty$

$p = 1$	$p \in (1, 2]$	$p \in (2, \infty)$	$p = \infty$
$\Theta\left(\sqrt{\frac{\ln d}{n}}\right)$	$\Theta\left(\frac{1}{\sqrt{n}}\right)$	$\Theta\left(\min\left\{\frac{1}{n^{1/p}}, \frac{d^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{n}}\right\}\right)$	$\Theta\left(\sqrt{\frac{d}{n}}\right)$

[Nemirovsky & Yudin:1983]

Stochastic Convex Optimization (SCO): Excess Risk Rates

- ℓ_p-setup: $\|\cdot\| = \|\cdot\|_p$, $1 \le p \le \infty$

$p = 1$	$p \in (1, 2]$	$p \in (2, \infty)$	$p = \infty$
$\Theta\left(\sqrt{\frac{\ln d}{n}}\right)$	$\Theta\left(\frac{1}{\sqrt{n}}\right)$	$\Theta\left(\min\left\{\frac{1}{n^{1/p}}, \frac{d^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{n}}\right\}\right)$	$\Theta\left(\sqrt{\frac{d}{n}}\right)$

[Nemirovsky & Yudin:1983]

- Upper bounds are achieved by Stochastic Mirror Descent (SMD)
- Algorithms run with a single pass over the data: $O(n)$ time
- Not only in expectation, but w/high probability (*regular norms*)

Differentially Private Stochastic Convex Optimization (DP-SCO)

Differentially-Private Stochastic Convex Optimization (DP-SCO)

DP-SCO: Given data $\mathrm{S} = (\mathbf{z}_1, \ldots, \mathbf{z}_n) \overset{i.i.d.}{\sim} \mathcal{D}^n$

Does there exist an (ε, δ)-DP algorithm $\mathcal{A} : \bigcup_n \mathcal{Z}^n \mapsto \mathcal{X}$ s.t.

$$\underbrace{\mathbb{E}_{\mathcal{A}} \mathbb{E}_{\mathbf{S} \sim \mathcal{D}^n} \Big[F_{\mathcal{D}}(\mathcal{A}(\mathbf{S})) - F^*_{\mathcal{D}} \Big]}_{\textit{excess (population) risk}} \overset{n \to \infty}{\longrightarrow} 0$$

Differentially-Private Stochastic Convex Optimization (DP-SCO)

p	Upper Bound	Lower bound
1	$\tilde{O}\left(\sqrt{\frac{\log d}{n}} + \left(\frac{\log d}{\varepsilon n}\right)^{2/3}\right)$	$\Omega\left(\sqrt{\frac{\log d}{n}} + \left(\frac{1}{\varepsilon n}\right)^{2/3}\right)$
$(1,2]$	$\tilde{O}\left(\sqrt{\frac{\kappa}{n}} + \frac{\kappa\sqrt{d}}{\varepsilon n}\right)$	$\Omega\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\kappa\varepsilon n}\right)$
$(2,\infty)$	$\tilde{O}\left(\frac{d^{1/2-1/p}}{\sqrt{n}} + \frac{d^{1-1/p}}{\varepsilon n}\right)$	$\Omega\left(\min\left\{\frac{d^{1/2-1/p}}{\sqrt{n}}, \frac{1}{(\varepsilon n)^{1/p}}, \frac{d^{1-1/p}}{n\varepsilon}\right\}\right)$
∞	$\tilde{O}\left(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}\right)$	$\Omega\left(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}\right)$

Notes: *[BFTT:'19, AFKT:'21, BGN:'21, ABGMU:'22]*

- ℓ_1-setup also requires smoothness
- $\kappa = 1/(p-1)$: strong convexity of ℓ_p, $1 < p \leq 2$
- Upper bounds also hold w/high probability
- For smooth case, algorithms are single pass and projection free

DP-SCO: ℓ_1-setup

Avoiding the Curse of Dimensionality in DP-SCO

Is the polynomial dimension-dependence in DP-SCO risk avoidable?

Optimal excess risk ℓ_2-setup *[Bassily, Feldman, Talwar & Thakurta:'19]*

$$\Theta\Big(L_0 D\Big(\underbrace{\frac{1}{\sqrt{n}}}_{\text{SCO}} + \underbrace{\frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}}_{\text{DP-ERM [BST:'14]}}\Big)\Big)$$

- Need not to release high-dimensional vectors *[Steinke & Ullman:'15]*

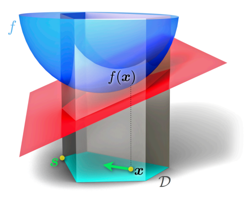
Avoiding the Curse of Dimensionality in DP-SCO

Is the polynomial dimension-dependence in DP-SCO risk avoidable?

- Need not to release high-dimensional vectors *[Steinke & Ullman:'15]*
- There exists one optimization algorithm with implicit updates:

$$v^{t+1} = \arg\min\{\langle \nabla f(x^t), v\rangle : v \in \mathsf{ext}(\mathcal{X})\}$$

Conditional gradient (a.k.a. Frank-Wolfe) algorithm



Avoiding the Curse of Dimensionality in DP-SCO

Is the polynomial dimension-dependence in DP-SCO risk avoidable?

- There exists one optimization algorithm with implicit updates:

$$v^{t+1} = \arg\min\{\langle \nabla f(x^t), v\rangle : v \in \mathsf{ext}(\mathcal{X})\}$$

Conditional gradient (a.k.a. Frank-Wolfe) algorithm

- Can be made private by adding Laplace noise on each $\langle \nabla f(x^t), v\rangle$, and minimizing the noisy evaluations

Avoiding the Curse of Dimensionality in DP-SCO

Is the polynomial dimension-dependence in DP-SCO risk avoidable?

- There exists one optimization algorithm with implicit updates:

$$v^{t+1} = \arg\min\{\langle \nabla f(x^t), v\rangle : v \in \mathsf{ext}(\mathcal{X})\}$$

Conditional gradient (a.k.a. Frank-Wolfe) algorithm

- (Full-batch) Private FW on ERM achieves nearly-optimal error

[Talwar, Thakurta & Zhang:'14-'15]

- Conversion to SCO excess risk guarantees always suboptimal
- Stochastic FW has suboptimal rates, even nonprivately!

[Hazan & Luo:'16]

Polyhedral Stochastic Frank-Wolfe w/Variance Reduction

Non-privately proposed in [Hassani, Karbasi, Mokhtari & Shen:'19; Zhang, Shen, Mokhtari, Hassani & Karbasi:'20]

$g \leftarrow$ minibatch gradient $\approx \nabla F_{\mathcal{D}}(x^t)$

$\Delta \leftarrow$ grad variation $\approx \nabla F_{\mathcal{D}}(x^t) - \nabla F_{\mathcal{D}}(x^{t-1})$

VarianceReducedGradient(Δ, g, ∇^{t-1}**)**

$$\nabla^t = (1-\eta)\left(\nabla^{t-1} + \Delta\right) + \eta g$$

FrankWolfeUpdate(∇, x^t**)**

$$v = \arg\min_{v \in \mathsf{ext}(\mathcal{X})} \left[\langle \nabla, v \rangle\right]$$

$$x^{t+1} \longleftarrow (1-\eta)x^t + \eta v$$

Polyhedral Stochastic Frank-Wolfe w/Variance Reduction

Poly-SFW *[Bassily, G. & Nandi:'21]*

Starting batch size: $n/2$ ($\to$ sensitivity control)

Batch size 1 for updates ($\to n/2$ iterations)

$g \leftarrow$ minibatch gradient $\approx \nabla F_{\mathcal{D}}(x^t)$

$\Delta \leftarrow$ grad variation $\approx \nabla F_{\mathcal{D}}(x^t) - \nabla F_{\mathcal{D}}(x^{t-1})$

VarianceReducedGradient(Δ, g, ∇^{t-1})

$$\nabla^t = (1-\eta)\left(\nabla^{t-1} + \Delta\right) + \eta g$$

PrivateFrankWolfeUpdate(∇, x^t)

$$v = \arg\min_{v \in \mathsf{ext}(\mathcal{X})} \left[\langle \nabla, v\rangle + \mathsf{Lap}\left(\frac{2s_t}{\varepsilon}\sqrt{n\log(1/\delta)}\right)\right]$$

$$x^{t+1} \longleftarrow (1-\eta)x^t + \eta v$$

Note: s_t is the sensitivity of $\langle \nabla^t, v\rangle$ w.r.t. $\mathbf{S}$

Poly-SFW: Privacy Analysis

Lemma: Sensitivity Bound

For the Poly-SFW algorithm, let (*global sensitivity*)

$$s_t := \max_{v \in \mathsf{ext}(\mathcal{X})} \max_{S \simeq S'} |\langle v, \nabla_t(S) - \nabla_t(S') \rangle|$$

Then

$$s_t \leq \max \left\{ \tfrac{2L_0 D}{n}(1-\eta)^t, 2\eta(L_1 D^2 + L_0 D) \right\}$$

Poly-SFW: Privacy Analysis

Lemma: Sensitivity Bound

For the Poly-SFW algorithm, let (*global sensitivity*)

$$s_t := \max_{v \in \mathsf{ext}(\mathcal{X})} \max_{S \simeq S'} |\langle v, \nabla_t(S) - \nabla_t(S') \rangle|$$

Then

$$s_t \leq \max \left\{ \tfrac{2L_0 D}{n}(1-\eta)^t, 2\eta(L_1 D^2 + L_0 D) \right\}$$

Corollary

The Poly-SFW algorithm is (ε, δ)-DP

Poly-SFW: Privacy Analysis

Lemma: Sensitivity Bound

For the Poly-SFW algorithm, let (*global sensitivity*)

$$s_t := \max_{v \in \text{ext}(\mathcal{X})} \max_{S \simeq S'} |\langle v, \nabla_t(S) - \nabla_t(S') \rangle|$$

Then

$$s_t \leq \max\left\{\tfrac{2L_0 D}{n}(1-\eta)^t, 2\eta(L_1 D^2 + L_0 D)\right\}$$

Corollary

The Poly-SFW algorithm is (ε, δ)-DP

Proof idea.

- By Lemma, any given step t is $(\varepsilon/\sqrt{n \ln(1/\delta)}, 0)$-DP

  (Report Noisy Max)
- Advanced composition of DP gives (ε, δ)-DP $\square$

Poly-SFW: Convergence Analysis

Lemma: Variance-Reduced Gradient Estimate

For Poly-SFW, the recursive gradient estimator ∇^t satisfies

$$\mathbb{E}_{\mathcal{A},\mathbf{S}\sim\mathcal{D}^n}\|\nabla^t-\nabla F_{\mathcal{D}}(x^t)\|_\infty \leq 4L_0\sqrt{\frac{\ln d}{n}}(1-\eta)^t+4\eta\sqrt{2t\ln(d)}(L_1D+L_0)$$

Proof idea. Use that ℓ_∞ is $(2\ln d)$-*regular* and solve recursive estimator (2nd moment) bounds *[Juditsky & Nemirovski:2009]*

Poly-SFW: Convergence Analysis

Lemma: Variance-Reduced Gradient Estimate

For Poly-SFW, the recursive gradient estimator ∇^t satisfies

$$\mathbb{E}_{\mathcal{A},\mathbf{S}\sim\mathcal{D}^n}\|\nabla^t - \nabla F_{\mathcal{D}}(x^t)\|_\infty \leq 4L_0\sqrt{\frac{\ln d}{n}}(1-\eta)^t + 4\eta\sqrt{2t\ln(d)}(L_1 D + L_0)$$

Proof idea. Use that ℓ_∞ is $(2\ln d)$-*regular* and solve recursive estimator (2nd moment) bounds *[Juditsky & Nemirovski:2009]*

$(\mathbb{R}^d, \|\cdot\|_\infty) \longrightarrow (\mathbb{R}^d, \|\cdot\|_q)$, with $q = \ln d$. These norms are equivalent, and $\|\cdot\|_q$ is $(\ln d)$-smooth

$$\|x+y\|_q^2 \leq \|x\|_q^2 + \langle \nabla(\|\cdot\|_q^2)(x), y\rangle + (\ln d)\|y\|_q^2$$

Poly-SFW: Convergence Analysis

Lemma: Variance-Reduced Gradient Estimate

For Poly-SFW, the recursive gradient estimator ∇^t satisfies

$$\mathbb{E}_{\mathcal{A},\mathbf{S}\sim\mathcal{D}^n}\|\nabla^t - \nabla F_{\mathcal{D}}(x^t)\|_\infty \leq 4L_0\sqrt{\frac{\ln d}{n}}(1-\eta)^t + 4\eta\sqrt{2t\ln(d)}(L_1 D + L_0)$$

Theorem [Bassily, G. & Nandi:'21]

Poly-SFW algorithm attains excess risk

$$\mathbb{E}_{\mathcal{A},\mathbf{S}\sim\mathcal{D}^n}[F_{\mathcal{D}}(\mathcal{A}(\mathbf{S})) - F^*_{\mathcal{D}}] = O\left((L_1 D^2 + L_0 D)\frac{\ln(d)\ln\left(\frac{n}{\ln d}\right)\sqrt{\ln(1/\delta)}}{\varepsilon\sqrt{n}}\right)$$

Note: Both gradient estimator error and accuracy can be bounded with high probability, by leveraging regularity

ℓ_1-Setup: Further Improvements

- Our bound is nearly-optimal, as long as $\varepsilon = \Theta(1)$
- What about $\varepsilon = o(1)$?

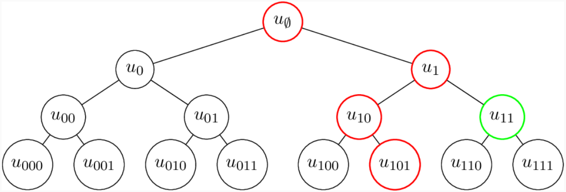
ℓ_1-Setup: Further Improvements

[Asi, Feldman, Koren & Talwar:'21]

- Provide a (ε, δ)-DP algorithm with improved excess risk

$$O\Bigg(\underbrace{(L_0 D + L_1 D^2)\sqrt{\frac{\log d}{n}}}_{\text{SCO}} \log n + \underbrace{L_1 D^2 \left[\frac{\log(d) \log^2(n) \log(1/\delta)}{\varepsilon n}\right]^{2/3}}_{\text{DP-ERM [TTZ:'15]}} \Bigg)$$

- Similar to Poly-SFW, combined with *tree-aggregation for prefix sums* + priv. amplification by shuffling



ℓ_p **Setup:** $1 < p < 2$

Lower Bounds for ℓ_p Setup: $1 < p < 2$

Theorem [Bassily, G. & Nandi:'21]

Consider ℓ_p-setup, $1 < p \leq 2$. If $\mathcal{A} : \mathcal{Z}^n \mapsto \mathcal{X}$ is (ε, δ)-DP, then

- DP-SCO excess risk $\tilde{\Omega}\Big(\frac{1}{\sqrt{n}} + (p-1)\frac{\sqrt{d}}{\varepsilon n}\Big)$
- ERM error is $\Omega\Big((p-1)\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\Big)$

Lower Bounds for ℓ_p Setup: $1 < p < 2$

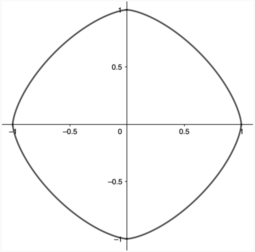
Theorem [Bassily, G. & Nandi:'21]

Consider ℓ_p-setup, $1 < p \leq 2$. If $\mathcal{A} : \mathcal{Z}^n \mapsto \mathcal{X}$ is (ε, δ)-DP, then

- DP-SCO excess risk $\tilde{\Omega}\Big(\frac{1}{\sqrt{n}} + (p-1)\frac{\sqrt{d}}{\varepsilon n}\Big)$
- ERM error is $\Omega\Big((p-1)\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\Big)$

Remarks:

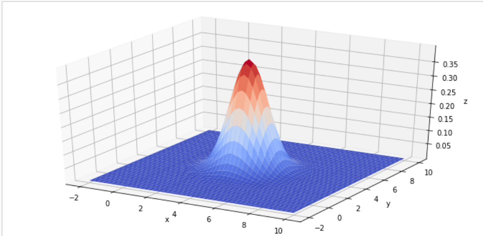
- Sudden transition in the excess risk when $p = 1 + \Omega(1)$
- LB is tight, up to $(p-1)$ factor *[AFKT:'21; BGN:'21]*
- Proof uses strong convexity of ℓ_p *[Ball, Carlen & Lieb:'94]*



Upper Bounds for ℓ_p-setups: Generalized Gaussian Mechanism

- Recall the (isotropic) Gaussian density

$$g(z) = C\exp\{-\|z-\mu\|_2^2/[2\sigma^2]\}$$



Upper Bounds for ℓ_p-setups: Generalized Gaussian Mechanism

- Recall the (isotropic) Gaussian density

$$g(z) = C\exp\{-\|z-\mu\|_2^2/[2\sigma^2]\}$$

- Let $(\mathbf{E}, \|\cdot\|_*)$ be κ-regular w/smooth norm $\|\cdot\|_+$, and an algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbf{E}$ with $\|\cdot\|_*$-sensitivity

$$\Delta = \sup_{S \simeq S'} \|\mathcal{A}(S) - \mathcal{A}(S')\|_*$$

- **Generalized Gaussian (GG) Mechanism:**

$\mathcal{A}_{\mathcal{GG}}(S)$ w/density $g(z) = C\exp\{-\|z-\mathcal{A}(S)\|_+^2/[2\sigma^2]\}$

Upper Bounds for ℓ_p-setups: Generalized Gaussian Mechanism

- **Generalized Gaussian (GG) Mechanism:**

$$\mathcal{A}_{\mathcal{GG}}(S) \text{ w/density } g(z) = C \exp\{-\|z - \mathcal{A}(S)\|_+^2/[2\sigma^2]\}$$

Proposition [Bassily, G. & Nandi:'21]

- If $\sigma^2 = 2\kappa \log(1/\delta)\Delta^2/\varepsilon^2$, the GG mechanism is (ε, δ)-DP
- $\mathbb{E}[\|\mathcal{A}(S) - \mathcal{A}_{\mathcal{GG}}(S)\|_*^2] \le d\sigma^2$

Consequence: GG mechanism allows use of Noisy stochastic first-order algorithms for spaces whose dual is κ-regular

GG Mechanism: Analysis

- **Rényi DP:** Let $\mathbb{P} = \mathcal{A}(S)$ and $\mathbb{Q} = \mathcal{A}(S')$

$$\begin{aligned}
& \exp\{(\alpha - 1)D_\alpha(\mathbb{P}||\mathbb{Q})\} \\
= \quad & C \int_{\mathbb{R}^d} \left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)^\alpha d\mathbb{Q} \\
= \quad & C \int_{\mathbb{R}^d} \exp\left\{ - \frac{\alpha}{2\sigma^2} \|z - \mu_1\|_+^2 + \frac{\alpha - 1}{2\sigma^2} \|z - \mu_2\|_+^2 \right\} dz \\
= \quad & C \int_{\mathbb{R}^d} \exp\left\{ - \frac{\alpha}{2\sigma^2} \|z - \mu_1 + \mu_2\|_+^2 + \frac{\alpha - 1}{2\sigma^2} \|z\|_+^2 \right\} dz.
\end{aligned}$$

GG Mechanism: Analysis

- **Rényi DP:** Let $\mathbb{P} = \mathcal{A}(S)$ and $\mathbb{Q} = \mathcal{A}(S')$

$$
\begin{aligned}
& \exp\{(\alpha-1)D_\alpha(\mathbb{P}||\mathbb{Q})\} \\
= \;& C\int_{\mathbb{R}^d} \left(\tfrac{d\mathbb{P}}{d\mathbb{Q}}\right)^\alpha d\mathbb{Q} \\
= \;& C\int_{\mathbb{R}^d} \exp\Big\{-\tfrac{\alpha}{2\sigma^2}\|z-\mu_1\|_+^2 + \tfrac{\alpha-1}{2\sigma^2}\|z-\mu_2\|_+^2\Big\}dz \\
= \;& C\int_{\mathbb{R}^d} \exp\Big\{-\tfrac{\alpha}{2\sigma^2}\|z-\mu_1+\mu_2\|_+^2 + \tfrac{\alpha-1}{2\sigma^2}\|z\|_+^2\Big\}dz.
\end{aligned}
$$

Let $\mu = \mu_1 - \mu_2$ and $p(\cdot) = \|\cdot\|_+^2$. Use convexity and smoothness of $\|\cdot\|_+^2$

$$
\begin{aligned}
-\alpha\|z-\mu\|_+^2 &\leq -\alpha\|z\|_+^2 + \langle \nabla p(z), \alpha\mu\rangle \\
&\leq -\alpha\|z\|_+^2 + [\|z\|_+^2 - \|z-\alpha\mu\|_+^2 + \kappa_+\|\alpha\mu\|_+^2]
\end{aligned}
$$

GG Mechanism: Analysis

- **Rényi DP:** Let $\mathbb{P} = \mathcal{A}(S)$ and $\mathbb{Q} = \mathcal{A}(S')$

$$\begin{aligned} & \exp\{(\alpha-1)D_\alpha(\mathbb{P}\|\mathbb{Q})\} \\ = \quad & C\int_{\mathbb{R}^d}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)^\alpha d\mathbb{Q} \\ = \quad & C\int_{\mathbb{R}^d}\exp\Big\{-\tfrac{\alpha}{2\sigma^2}\|z-\mu_1\|_+^2+\tfrac{\alpha-1}{2\sigma^2}\|z-\mu_2\|_+^2\Big\}dz \\ = \quad & C\int_{\mathbb{R}^d}\exp\Big\{-\tfrac{\alpha}{2\sigma^2}\|z-\mu_1+\mu_2\|_+^2+\tfrac{\alpha-1}{2\sigma^2}\|z\|_+^2\Big\}dz. \end{aligned}$$

Let $\mu = \mu_1 - \mu_2$ and $p(\cdot) = \|\cdot\|_+^2$. Use convexity and smoothness of $\|\cdot\|_+^2$

$$\begin{aligned} -\alpha\|z-\mu\|_+^2 \quad & \leq \quad -\alpha\|z\|_+^2 + \langle\nabla p(z), \alpha\mu\rangle \\ & \leq \quad -\alpha\|z\|_+^2 + [\|z\|_+^2 - \|z-\alpha\mu\|_+^2 + \kappa_+\|\alpha\mu\|_+^2] \end{aligned}$$

- Plugging the bound,

$\exp\{(\alpha-1)D_\alpha(\mathbb{P}\|\mathbb{Q})\} \leq \frac{\kappa_+\alpha^2}{2\sigma^2(\alpha-1)}\|\mu_1-\mu_2\|_+^2 \leq \frac{\kappa\alpha^2}{2\sigma^2(\alpha-1)}\|\mu_1-\mu_2\|^2\square$

Upper Bounds: Noisy Variance-Reduced Stochastic Frank-Wolfe

- Follows a similar strategy to the polyhedral case, but:
  - Add GG noise to the gradient estimator
  - Solve the linear optimization subroutine exactly

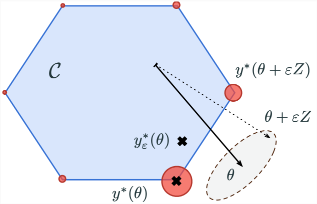


figure from [Berthet et al.'20]

Upper Bounds: Noisy Variance-Reduced Stochastic Frank-Wolfe

- Follows a similar strategy to the polyhedral case, but:
  - Add GG noise to the gradient estimator
  - Solve the linear optimization subroutine exactly

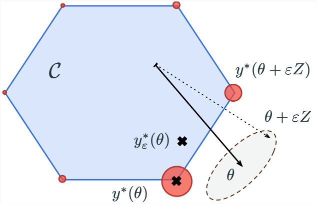


figure from [Berthet et al.:'20]

- Naive version of this algorithm is suboptimal $O\left(\frac{\kappa}{\sqrt{n}} + \frac{\kappa\sqrt{d}}{\varepsilon n^{3/4}}\right)$
- In combination with tree aggregation we get **optimal rates in a single pass,** $O\left(\sqrt{\frac{\kappa}{n}} + \frac{\kappa\sqrt{d}}{\varepsilon n}\right)$
- No privacy amplification needed
- Algorithm is general: works for any space whose dual is κ-regular

Upper Bounds: Noisy Variance-Reduced Stochastic Frank-Wolfe

- Follows a similar strategy to the polyhedral case, but:
  - Add GG noise to the gradient estimator
  - Solve the linear optimization subroutine exactly
- Naive version of this algorithm is suboptimal $O\big(\frac{\kappa}{\sqrt{n}} + \frac{\kappa\sqrt{d}}{\varepsilon n^{3/4}}\big)$
- In combination with tree aggregation we get **optimal rates in a single pass**, $O\big(\sqrt{\frac{\kappa}{n}} + \frac{\kappa\sqrt{d}}{\varepsilon n}\big)$
- No privacy amplification needed
- Algorithm is general: works for any space whose dual is κ-regular

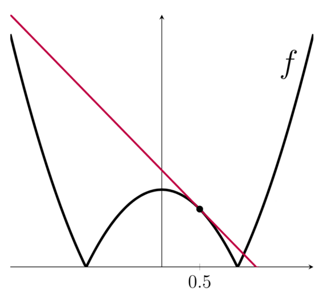
Note: AFKT:'21 obtained same rates for $1 < p \leq 2$ nonsmooth case, but their oracle complexity is superlinear in n

Nonconvex Losses

DP Stochastic Nonconvex Optimization

- Vanishing excess risk is provably hard
- Stationarity measures:
  - **Unconstrained smooth:** $\mathbb{E}_{\mathcal{A},\mathbf{S}} \|\nabla F_{\mathcal{D}}(\mathcal{A}(\mathbf{S}))\|_*$

    [Wang, Chen & Xu:'19; Wang, Xu:'19; Song, Steinke, Thakkar & Thakurta:'21; Zhou, Chen, Hong & Wu:'20]
  - **Constrained smooth:** $\mathbb{E}_{\mathcal{A},\mathbf{S}} \sup_{x\in\mathcal{X}} \langle \nabla F_{\mathcal{D}}(\mathcal{A}(\mathbf{S})), \mathcal{A}(\mathbf{S}) - x \rangle$
  - **Weakly convex (nonsmooth):** close to a near-stationary point



DP Stochastic Nonconvex Optimization

- Vanishing excess risk is provably hard
- Stationarity measures:
  - **Unconstrained smooth:** $\mathbb{E}_{\mathcal{A},\mathbf{S}}\|\nabla F_{\mathcal{D}}(\mathcal{A}(\mathbf{S}))\|_*$

    [Wang, Chen & Xu:'19; Wang, Xu:'19; Song, Steinke, Thakkar & Thakurta:'21; Zhou, Chen, Hong & Wu:'20]
  - **Constrained smooth:** $\mathbb{E}_{\mathcal{A},\mathbf{S}} \sup_{x\in\mathcal{X}}\langle\nabla F_{\mathcal{D}}(\mathcal{A}(\mathbf{S})), \mathcal{A}(\mathbf{S})-x\rangle$
  - **Weakly convex (nonsmooth):** close to a near-stationary point

Setting	ℓ_p Setup	Rate	Linear Time?
Smooth Constrained	$p=1$	$\frac{\log^{2/3} d}{(n\varepsilon)^{1/3}}$	✓
	$1<p\leq 2$	$\frac{1}{n^{1/3}}+\left(\frac{\sqrt{d}}{n\varepsilon}\right)^{2/5}$	✓
Weakly Convex (Nonsmooth)	$1\leq p\leq 2$	$\frac{1}{n^{1/4}}+\left(\frac{\sqrt{d}}{n\varepsilon}\right)^{1/3}$	No

[Bassily, G. & Menart:'21]

DP Stochastic Nonconvex Optimization

Setting	ℓ_p Setup	Rate	Linear Time?
Smooth Constrained	$p = 1$	$\frac{\log^{2/3} d}{(n\varepsilon)^{1/3}}$	✓
	$1 < p \leq 2$	$\frac{1}{n^{1/3}} + \left(\frac{\sqrt{d}}{n\varepsilon}\right)^{2/5}$	✓
Weakly Convex (Nonsmooth)	$1 \leq p \leq 2$	$\frac{1}{n^{1/4}} + \left(\frac{\sqrt{d}}{n\varepsilon}\right)^{1/3}$	No

[Bassily, G. & Menart:'21]

Open Problem: Lower bounds for stationarity?

DP Stochastic Weakly Convex Optimization

- **Key Observation:** Weakly convex functions can be convexified by strongly convex regularization

DP Stochastic Weakly Convex Optimization

- **Key Observation:** Weakly convex functions can be convexified by strongly convex regularization
- Algorithm based on a sequence of *stochastic proximal steps*

$$\mathsf{prox}_{1/\beta}(x^t) = \arg\min_y \left\{ F_{\mathcal{D}}(x) + \frac{\beta}{2}\|x - x^t\|^2 \right\}$$

- Each subproblem solved with optimal risk by *phased noisy stochastic mirror-descent*, with disjoint data batches *[AFKT:'21]*

DP Stochastic Weakly Convex Optimization

- **Key Observation:** Weakly convex functions can be convexified by strongly convex regularization
- Algorithm based on a sequence of *stochastic proximal steps*

$$\mathsf{prox}_{1/\beta}(x^t) = \arg\min_y \left\{ F_{\mathcal{D}}(x) + \frac{\beta}{2}\|x - x^t\|^2 \right\}$$

- Each subproblem solved with optimal risk by *phased noisy stochastic mirror-descent*, with disjoint data batches *[AFKT:'21]*
- **Guarantee:** Randomly chosen iterate $\hat{x}$ satisfies *close to near stationarity* in expectation, for $\vartheta = \tilde{O}\Big(\frac{1}{n^{1/4}} + \big(\frac{\sqrt{d}}{n\varepsilon}\big)^{1/3}\Big)$,

$$\exists x \in \mathcal{X} : \quad \|\hat{x} - x\| \le \vartheta, \qquad \inf_{g \in \partial F_{\mathcal{D}}(x)} \sup_{y \in \mathcal{X}} \langle g, x - y \rangle \le \vartheta$$

- $O(n^{-1/4})$ is best rate known nonprivately
  [Davis, Grimmer:'19; Davis, Drusvyatskiy:'19]
- Oracle complexity is $\tilde{O}(\min\{n^{3/2}, n^2\varepsilon/\sqrt{d}\})$

Summary

Summary

Conclusions

- Provide new algorithms for DP-SCO in ℓ_p-setups
- Novel and sharp lower bounds for DP-SCO in ℓ_p-setups
- Introduce new DP mechanism for regular normed spaces
- Extensions to stationary points for nonconvex settings

Summary

Conclusions

- Provide new algorithms for DP-SCO in ℓ_p-setups
- Novel and sharp lower bounds for DP-SCO in ℓ_p-setups
- Introduce new DP mechanism for regular normed spaces
- Extensions to stationary points for nonconvex settings

Future Directions

- What other sets, aside from polytopes, can avoid the $\tilde{\Omega}(\sqrt{d}/[\varepsilon n])$ lower bound for DP-SCO?
- Universally optimal algorithm for DP-SCO for general norms?
- Oracle complexity for nonsmooth DP-SCO
- Lower bounds for nonconvex DP-SO

Thank you!