M-estimation, noisy optimization and user-level local privacy

Marco Avella Medina

Meeting in Mathematical Statistics, CIRM

December 20, 2023

Columbia University
in the City of New York

Motivation

- In the recent years certain versions of differential privacy are being deployed by Microsoft, Apple, Mozilla, Google and the US Census Bureau

Motivation

- In the recent years certain versions of differential privacy are being deployed by Microsoft, Apple, Mozilla, Google and the US Census Bureau
- Lack of general differentially private tools for parametric inference
- Establish connections between privacy-preserving data analysis and robust statistics
- Study private counterparts of most commonly implemented algorithms for M-estimators in statistical software.

Diffferentially private inference via noisy optimization

Based on joint work with Casey Bradshaw and Po-Ling Loh

Our contribution

- Global finite-sample convergence analysis of private gradient descent and Newton method.
- The theory relies on local strong convexity and self-concordance.
- Identify loss functions that avoid bounded data, bounded parameter space and truncation arguments.
- Propose differentially private asymptotic confidence regions.

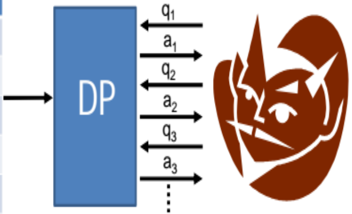
Related work

- DP and noisy optimization : Song et al. (2013), Bassily et al. (2014), Duchi et al. (2018), Feldman et al. (2020), Cai et al. (2021) among many many others...
- Private confidence intervals : Wang, Kifer and Lee (2019) proposes a similar technique. Other work includes Sheffet (2017), Karwa and Vadhan (2017), Barrientos et al. (2019), Canonne et al. (2019), Avella-Medina (2021)...

Differential privacy framework

- *Setting* : a trusted curator holds a sensitive database constituted by n individual rows.
- *Goal* : protect every individual row while allowing statistical analysis of the database as a whole

Sex	Blood	...	HIV
F	B	...	Y
M	A	...	N
M	O	...	N
M	O	...	Y
F	A	...	N
M	B	...	Y



Gaussian differential privacy

Dong, Roth and Su (2022, JRSS B)

Interpretation : telling whether someone is in the dataset is harder than telling apart $N(0, 1)$ and $N(\mu, 1)$

Gaussian differential privacy

Dong, Roth and Su (2022, JRSS B)

Interpretation : telling whether someone is in the dataset is harder than telling apart $N(0,1)$ and $N(\mu,1)$

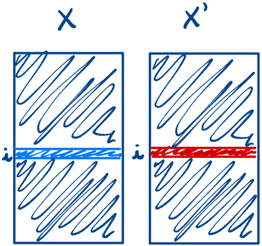
- New intuitive definition of differential privacy via hypothesis testing
  - Gaussian mechanism : $\tilde{m}(x_1,\ldots,x_n) = m(x_1,\ldots,x_n) + \frac{1}{\mu}\mathrm{GS}(m)N(0,1)$
  - Gaussian differential privacy : $H_0 : P = N(0,1)$ V. $H_1 : P = N(\mu,1)$

Gaussian differential privacy

Dong, Roth and Su (2022, JRSS B)

Interpretation : telling whether someone is in the dataset is harder than telling apart $N(0,1)$ and $N(\mu,1)$

- New intuitive definition of differential privacy via hypothesis testing
  - Gaussian mechanism : $\tilde{m}(x_1,\ldots,x_n) = m(x_1,\ldots,x_n) + \frac{1}{\mu}\mathrm{GS}(m)N(0,1)$
  - Gaussian differential privacy : $H_0 : P = N(0,1)$ V. $H_1 : P = N(\mu,1)$



$$\mathrm{GS}(m) = \sup_{x,x',\, d_H(x,x')=1} \| m(x) - m(x') \|_2$$

Ex: median

Gaussian differential privacy

Dong, Roth and Su (2022, JRSS B)

Interpretation : telling whether someone is in the dataset is harder than telling apart $N(0,1)$ and $N(\mu,1)$

- New intuitive definition of differential privacy via hypothesis testing
  - Gaussian mechanism : $\tilde{m}(x_1,\ldots,x_n) = m(x_1,\ldots,x_n) + \frac{1}{\mu}\mathrm{GS}(m)N(0,1)$
  - Gaussian differential privacy : $H_0 : P = N(0,1)$ V. $H_1 : P = N(\mu,1)$
- Nice characterization of composition
  - Product : $G_{\mu_1} \otimes G_{\mu_2} \cdots \otimes G_{\mu_K} = G_{\sqrt{\sum_{k=1}^{K} \mu_k^2}}$
  - CLT : $f_1 \otimes \cdots \otimes f_K \approx G_\mu$

M-estimators

An M-estimator $\hat{\theta} = T(F_n)$ of $\theta_0 \in \mathbb{R}^p$ (Huber, 1964) is defined as

$$\hat{\theta} = \text{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(z_i, \theta) = \text{argmin}_{\theta \in \mathbb{R}^p} E_{F_n}[\rho(Z, \theta)],$$

or by an implicit equation as

$$\frac{1}{n} \sum_{i=1}^{n} \Psi(z_i, \hat{\theta}) = E_{F_n}[\Psi(Z, \hat{\theta})] = 0.$$

M-estimators : properties

- For M-estimators the IF is proportional to Ψ :

$$IF(z; F, T) = M(\Psi, F)^{-1}\Psi(z; F, T)$$

i.e. bounded if $\Psi(z; F, T)$ is bounded.

- M-estimators are asymptotically normal :

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V(\Psi, F)),$$

where

$$\begin{aligned}
V(\Psi, F) &= M(\Psi, F)^{-1}Q(\Psi, F)M(\Psi, F)^{-1} \\
M(\Psi, F) &= -\frac{\partial}{\partial\theta}E_F[\Psi(Z, \theta)]\Big|_{\theta=T(F)} \\
Q(\Psi, F) &= E_F[\Psi(Z, T(F)) \cdot \Psi(Z, T(F))^\top].
\end{aligned}$$

Noisy Gradient Descent

- Noisy gradient descent :

$$\theta^{(k+1)} = \theta^{(k)} - \eta\left(\frac{1}{n}\sum_{i=1}^{n}\Psi(x_i, \theta^{(k)}) + \frac{2\sup\|\Psi\|_2 \cdot \sqrt{K}}{n\mu} Z_k\right)$$

$$\{Z_k\} \overset{iid}{\sim} N(0, I_p)$$

Noisy Gradient Descent

- Noisy gradient descent :

$$\theta^{(k+1)} = \theta^{(k)} - \eta\left(\frac{1}{n}\sum_{i=1}^{n}\Psi(x_i, \theta^{(k)}) + \frac{2\sup\|\Psi\|_2 \cdot \sqrt{K}}{n\mu} Z_k\right)$$

$\{Z_k\} \overset{iid}{\sim} N(0, I_p)$

GS(gradient)

Noisy Gradient Descent

- Noisy gradient descent :

$$\theta^{(k+1)} = \theta^{(k)} - \eta\left(\frac{1}{n}\sum_{i=1}^{n}\Psi(x_i, \theta^{(k)}) + \frac{2\sup\|\Psi\|_2 \cdot \sqrt{K}}{n\mu}Z_k\right)$$

$$\{Z_k\} \overset{iid}{\sim} N(0, I_p)$$

Theorem. Assuming local strong convexity, after $K \geq C \log n$ iterations of NGD we have that

1. $\theta^{(K)}$ is μ-GDP
2. $\theta^{(K)} - \theta_0 = \hat{\theta} - \theta_0 + O_p\left(\frac{\sqrt{K}p}{\mu n}\right)$
3. $\sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F))$

Noisy Gradient Descent

- Noisy gradient descent :

$$\theta^{(k+1)} = \theta^{(k)} - \eta\left(\frac{1}{n}\sum_{i=1}^{n}\Psi(x_i, \theta^{(k)}) + \frac{2\sup\|\Psi\|_2 \cdot \sqrt{K}}{n\mu} Z_k\right)$$

$$\{Z_k\} \overset{iid}{\sim} N(0, I_p)$$

Theorem. Assuming local strong convexity, after $K \geq C\log n$ iterations of NGD we have that

$O_p\left(\sqrt{\frac{p}{n}}\right)$ statistical error

1. $\theta^{(K)}$ is μ-GDP
2. $\theta^{(K)} - \theta_0 = \hat{\theta} - \theta_0 + O_p\left(\frac{\sqrt{K}p}{\mu n}\right)$ privacy error
3. $\sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F))$

Noisy Gradient Descent

- Noisy gradient descent :

$$\theta^{(k+1)} = \theta^{(k)} - \eta\left(\frac{1}{n}\sum_{i=1}^{n}\Psi(x_i, \theta^{(k)}) + \frac{2\sup\|\Psi\|_2 \cdot \sqrt{K}}{n\mu} Z_k\right)$$

$$\{Z_k\} \overset{iid}{\sim} N(0, I_p)$$

Theorem. Assuming local strong convexity, after $K \geq C \log n$ iterations of NGD we have that

1. $\theta^{(K)}$ is μ-GDP
2. $\theta^{(K)} - \theta_0 = \hat{\theta} - \theta_0 + O_p\left(\frac{\sqrt{K}p}{\mu n}\right)$
3. $\sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F))$

Remark

Optimal rates of convergence : our estimators attain near minimax rates of covergence under (ε, δ)-DP according to Cai, Wang and Zhang (2021, AoS)

$$\inf_{A \in \mathcal{A}_{\varepsilon,\delta}} \sup_{P \in \mathcal{P}(\sigma,p)} \mathbb{E}\|A(F_n) - \theta_0\| \gtrsim \sigma\left(\sqrt{\frac{p}{n}} + \frac{p\sqrt{\log(1/\delta)}}{n\varepsilon}\right)$$

Remark

Optimal rates of convergence : our estimators attain near minimax rates of covergence under (ε, δ)-DP according to Cai, Wang and Zhang (2021, AoS)

$$\inf_{A \in \mathcal{A}_{\varepsilon,\delta}} \sup_{P \in \mathcal{P}(\sigma,p)} \mathbb{E}\|A(F_n) - \theta_0\| \gtrsim \sigma \left(\sqrt{\frac{p}{n}} + \frac{p\sqrt{\log(1/\delta)}}{n\varepsilon} \right)$$

$= \frac{1}{\mu}$

Example : linear regression

- Consider a linear regression model

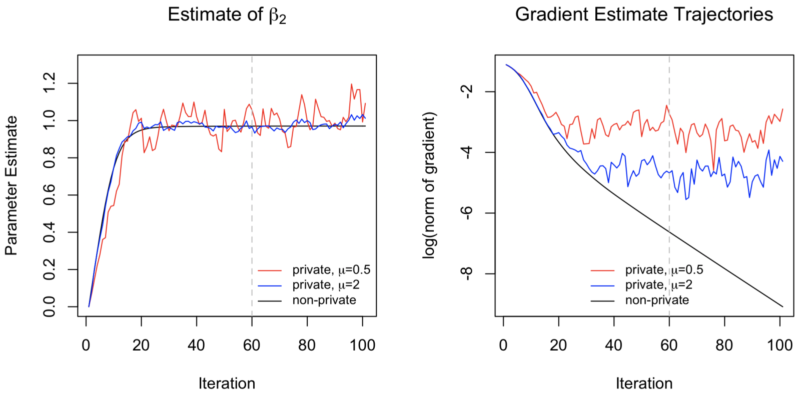
$$
\begin{aligned}
& y_i = x_i^T \beta + u_i \text{ for } i = 1, \ldots, n \\
& x_i \in \mathbb{R}^p \\
& u_i \sim N(0, \sigma^2)
\end{aligned}
$$

- We want to solve

$$(\hat{\beta}, \hat{\sigma}) = \text{argmin}_{\beta,\sigma} \left[\frac{1}{n} \sum_{i=1}^{n} \sigma \rho_c \left(\frac{y_i - x_i^T \beta}{\sigma} \right) w(x_i) + \frac{1}{2} \kappa n \sigma \right]$$

where $w(x_i) = \min\left(1, \frac{1}{\|x_i\|_2}\right)$ and κ is a Fisher consistency constant.

Example : linear regression



Noisy Newton

- Noisy Newton :

$$\theta^{(k+1)} = \theta^{(k)} - \left(\frac{1}{n} \sum_{i=1}^{n} \dot{\Psi}(x_i, \theta^{(k)}) + \frac{2\bar{B}\sqrt{2K}}{\mu n} W_k \right)^{-1} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \Psi(x_i, \theta^{(k)}) + \frac{2B\sqrt{2K}}{\mu n} N_k \right)$$

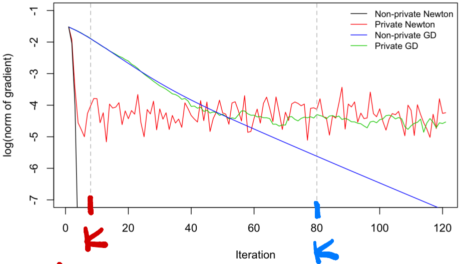
where $\{N_k\}$ and $\{W_k\}$ are i.i.d. sequences of vectors and symmetric matrices with i.i.d. standard normal components.

- Condition. Hessian of the form

$$\nabla^2 \mathcal{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} a(x_i, \theta) a(x_i, \theta)^\top,$$

where $\sup_{x,\theta} \|a(x, \theta)\|_2^2 \le \bar{B} < \infty$.

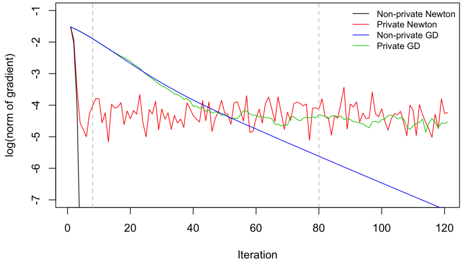
Noisy Newton theory



K for noisy Newton

K for noisy GD

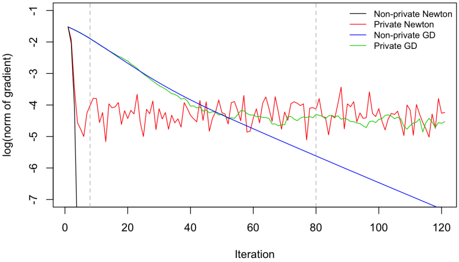
Noisy Newton theory



Theorem. Assuming local strong convexity, a Liptschitz continuous Hessian and $\|\nabla \mathcal{L}_n(\theta^{(0)})\| \leq \frac{\tau_1^2}{L}$, after $K \geq C \log\log n$ iterations of noisy Newton

1. $\theta^{(K)}$ is μ-GDP is differentially private
2. $\theta^{(K)} - \theta_0 = \hat{\theta} - \theta_0 + O_p\left(\frac{\sqrt{K}}{\mu}\frac{p}{n}\right)$
3. $\sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F))$

Noisy Newton theory



Theorem. Assuming local strong convexity, a Liptschitz continuous Hessian and $\|\nabla \mathcal{L}_n(\theta^{(0)})\| \leq \frac{\tau_1^2}{L}$, after $K \geq C \log\log n$ iterations of noisy Newton

1. $\theta^{(K)}$ is μ-GDP is differentially private
2. $\theta^{(K)} - \theta_0 = \hat{\theta} - \theta_0 + O_p\left(\frac{\sqrt{K}}{\mu}\frac{p}{n}\right)$
3. $\sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F))$

Discussion

Why is our approach interesting ?

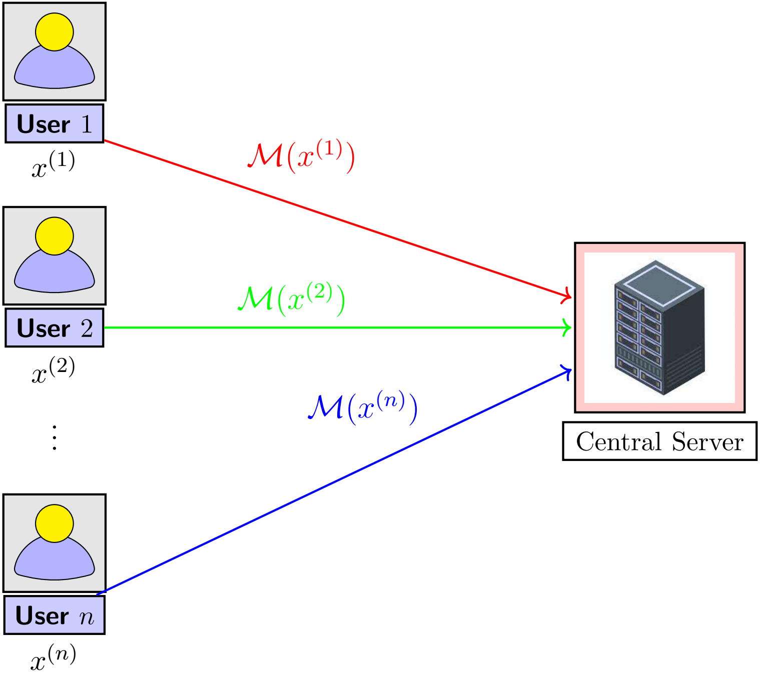
1. Algorithms are easy to implement and computationally efficient !
2. Importance of (local) strong convexity for optimal parametric rates of convergence
3. General framework for differentially private parametric inference
4. Connections between optimization, differential privacy and robust statistics.

M-estimators with user-level local differential privacy contraints

Based on joint work with Lekshmi Ramesh, Elise Han and Cindy Rush

Two variants of differential privacy

- Local Differential Privacy : Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith (STOC, 2008), Duchi, Jordan, Wainwright (JASA, 2018)
- User-level differential privacy : Liu, Suresh, Yu, Kumar, Riley (NeurIPS 2020), Levy, Sun, Amin, Kale, Kulesza, Mohri, Suresh. (NeurIPS 2021).



Local Differential Privacy

User-level privacy

- There are n users, and each user has m samples. We denote the samples of user i as

$$x^{(i)} = (x_1^{(i)}, \ldots, x_m^{(i)})$$

where $x_j^{(i)} \in \mathbb{R}^d$

User-level privacy

- There are n users, and each user has m samples. We denote the samples of user i as

$$x^{(i)} = (x_1^{(i)}, \dots, x_m^{(i)})$$

where $x_j^{(i)} \in \mathbb{R}^d$

- For a given $i \in [n]$, $x^{(i)} = (x_1^{(i)}, \dots, x_m^{(i)})$ and $x^{(i)\prime} = (x_1^{(i)\prime}, \dots, x_m^{(i)\prime})$ are user-level neighbors if there exists $\mathcal{S} \subseteq [m]$ such that

$$x_j^{(i)} \neq x_j^{(i)\prime}$$

for all $j \in \mathcal{S}$

User-level privacy

- A mechanism $\mathcal{M} : \mathbb{R}^{d \times m} \to \mathcal{Z}$ is said to be user-level (ε, δ)-LDP if, for every $x = (x_1, \ldots, x_m)$ and $x' = (x'_1, \ldots, x'_m)$ that are user-level neighbors and every $Z \subset \mathcal{Z}$, there exists $\varepsilon > 0$ and $\delta \in (0, 1)$ such that

$$\mathbb{P}(\mathcal{M}(x) \in Z) \leq e^{\varepsilon}\, \mathbb{P}(\mathcal{M}(x') \in Z) + \delta.$$

Empirical Risk Minimization

- Samples $\{x_j^{(i)}\}$ drawn i.i.d. from P_{θ_0} for $\theta_0 \in \Theta$
- Loss function $\ell : \mathbb{R}^d \times \Theta \to \mathbb{R}$
- Find a minimizer of the empirical risk

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(x_j^{(i)}, \theta) = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{L}_{n,m}(\theta).$$

Empirical Risk Minimization

- Samples $\{x_j^{(i)}\}$ drawn i.i.d. from P_{θ_0} for $\theta_0 \in \Theta$
- Loss function $\ell : \mathbb{R}^d \times \Theta \to \mathbb{R}$
- Find a minimizer of the empirical risk

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(x_j^{(i)}, \theta) = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{L}_{n,m}(\theta).$$

- We will assume ℓ is differentiable, smooth and locally strongly convex
- The per-sample gradients are bounded :

$$\|g_j^{(i)}(\theta)\|_2 = \|\nabla \ell(x_j^{(i)}, \theta)\|_2 \leq B$$

for all i, j, θ

User-level LDP ERM

- Users and the center communicate over multiple rounds to obtain $\hat{\theta}$

User-level LDP ERM

- Users and the center communicate over multiple rounds to obtain $\hat{\theta}$
- Round t involves the following steps :

User-level LDP ERM

- Users and the center communicate over multiple rounds to obtain $\hat{\theta}$
- Round t involves the following steps :
  - Users compute local gradients

$$g^{(i)}(\theta_t) = \frac{1}{m} \sum_{j=1}^{m} g_j^{(i)}(\theta_t)$$

User-level LDP ERM

- Users and the center communicate over multiple rounds to obtain $\hat{\theta}$
- Round t involves the following steps :
  - Users compute local gradients

$$g^{(i)}(\theta_t) = \frac{1}{m} \sum_{j=1}^{m} g_j^{(i)}(\theta_t)$$

  - Users and center run the user-level LDP mean estimation algorithm with $\{g^{(i)}(\theta_t)\}_{i \in [n]}$ as inputs to obtain $\hat{g}(\theta_t)$

User-level LDP ERM

- Users and the center communicate over multiple rounds to obtain $\hat{\theta}$
- Round t involves the following steps :
  - Users compute local gradients

$$g^{(i)}(\theta_t) = \frac{1}{m} \sum_{j=1}^{m} g_j^{(i)}(\theta_t)$$

  - Users and center run the user-level LDP mean estimation algorithm with $\{g^{(i)}(\theta_t)\}_{i \in [n]}$ as inputs to obtain $\hat{g}(\theta_t)$
  - Center updates parameter

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

  and sends it to all users

User-level LDP ERM

- The update rule can be rewritten as

$$\theta_{t+1} = \theta_t - \frac{\eta}{mn} \sum_{i,j} g_j^{(i)}(\theta_t) - \eta Z_{1,t} + \eta Z_{2,t}$$

where

$$\begin{aligned} Z_{1,t} &= \hat{g}(\theta_t) - \mathbb{E}[g_j^{(i)}(\theta_t)] \\ Z_{2,t} &= \frac{1}{mn} \sum_{i,j} g_j^{(i)}(\theta_t) - \mathbb{E}[g_j^{(i)}(\theta_t)] \end{aligned}$$

Mean estimation under user-level local privacy constraints

Problem setting

- Samples $\{x_j^{(i)}\}_{i\in[n],j\in[m]}$ drawn i.i.d. from a distribution with mean μ

Problem setting

- Samples $\{x_j^{(i)}\}_{i\in[n],j\in[m]}$ drawn i.i.d. from a distribution with mean μ
- User i communicates its sample through mechanism $\mathcal{M} : \mathbb{R}^{dm} \to \mathcal{Z}$ to a center

Problem setting

- Samples $\{x_j^{(i)}\}_{i\in[n],j\in[m]}$ drawn i.i.d. from a distribution with mean μ
- User i communicates its sample through mechanism $\mathcal{M} : \mathbb{R}^{dm} \to \mathcal{Z}$ to a center
- The center uses an estimator $f : \mathcal{Z}^n \to \mathbb{R}^d$ to output an estimate

$$\hat{\mu} = f(\mathcal{M}(x^{(1)}), \ldots, \mathcal{M}(x^{(n)}))$$

Problem setting

- Samples $\{x_j^{(i)}\}_{i\in[n],j\in[m]}$ drawn i.i.d. from a distribution with mean μ
- User i communicates its sample through mechanism $\mathcal{M} : \mathbb{R}^{dm} \to \mathcal{Z}$ to a center
- The center uses an estimator $f : \mathcal{Z}^n \to \mathbb{R}^d$ to output an estimate

$$\hat{\mu} = f(\mathcal{M}(x^{(1)}), \ldots, \mathcal{M}(x^{(n)}))$$

- Design mechanism $\mathcal{M}$ and an estimation procedure f such that

Problem setting

- Samples $\{x_j^{(i)}\}_{i\in[n],j\in[m]}$ drawn i.i.d. from a distribution with mean μ
- User i communicates its sample through mechanism $\mathcal{M}:\mathbb{R}^{dm}\to\mathcal{Z}$ to a center
- The center uses an estimator $f:\mathcal{Z}^n\to\mathbb{R}^d$ to output an estimate

$$\hat{\mu}=f(\mathcal{M}(x^{(1)}),\ldots,\mathcal{M}(x^{(n)}))$$

- Design mechanism $\mathcal{M}$ and an estimation procedure f such that
  1. The mechanism $\mathcal{M}$ is user-level (ε,δ)-LDP

Problem setting

- Samples $\{x_j^{(i)}\}_{i\in[n],j\in[m]}$ drawn i.i.d. from a distribution with mean μ
- User i communicates its sample through mechanism $\mathcal{M}:\mathbb{R}^{dm}\to\mathcal{Z}$ to a center
- The center uses an estimator $f:\mathcal{Z}^n\to\mathbb{R}^d$ to output an estimate

$$\hat{\mu}=f(\mathcal{M}(x^{(1)}),\ldots,\mathcal{M}(x^{(n)}))$$

- Design mechanism $\mathcal{M}$ and an estimation procedure f such that
  1. The mechanism $\mathcal{M}$ is user-level (ε,δ)-LDP
  2. The estimation error $\|\hat{\mu}-\mu\|_2$ is small with high probability

A naive estimator

- Each user sends a noisy version of its local mean estimate
- Assume $d = 1$ and $|x_j^{(i)}| \leq B$. Local mean
$$y_i = \frac{1}{m} \sum_{j=1}^{m} x_j^{(i)}$$
has sensitivity $2B$
- User i sends
$$\mathcal{M}(x^{(i)}) = y_i + w_i$$
where $w_i \overset{iid}{\sim} \mathcal{N}\left(\frac{2B^2}{\varepsilon^2} \ln \frac{2}{\delta} \right)$
- Center computes final estimate
$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{M}(x^{(i)})$$

A naive estimator

- $\mathcal{M}$ is user-level (ε, δ)-LDP
- The estimator has error

$$\mathbb{E}[\|\hat{\mu} - \mu\|_2] = \tilde{O}\left(\frac{B}{\sqrt{mn}} + \frac{B}{\sqrt{n\varepsilon}}\right)$$

- Error term due to privacy constraint does not improve with m

An improved estimator

- We will use the fact that the local averages y_i concentrate in an interval of size $O(B/\sqrt{m})$ around the mean with high probability
- Projecting y_i onto this interval reduces sensitivity (and therefore noise) by a factor of $1/\sqrt{m}$

An improved estimator

- We will use the fact that the local averages y_i concentrate in an interval of size $O(B/\sqrt{m})$ around the mean with high probability
- Projecting y_i onto this interval reduces sensitivity (and therefore noise) by a factor of $1/\sqrt{m}$
- Two-round estimator :
  - Round 1 : Center computes private estimate for an $O(B/\sqrt{m})$ sized interval containing the mean with high probability and sends it to users
  - Round 2 : Users send projected private local means to center which then computes the final average

Algorithm

- Round 1
  - At each user i : Divide the interval $[-B, B]$ into disjoint intervals of width $2B\sqrt{2\ln(2n/\xi)}/\sqrt{m}$. Find interval where y_i lies and send randomized bin index.
  - At center : find most popular interval $\tilde{I}$ and send to all users

Algorithm

- Round 1
  - At each user i : Divide the interval $[-B, B]$ into disjoint intervals of width $2B\sqrt{2\ln(2n/\xi)}/\sqrt{m}$. Find interval where y_i lies and send randomized bin index.
  - At center : find most popular interval $\tilde{I}$ and send to all users
- Round 2
  - At each user i : compute the noisy truncated mean

$$\tilde{\mu}_i = \mathtt{Proj}_{\tilde{I}}(y_i) + w_i,$$

    where $w_i \sim \mathcal{N}(0, 8\sigma^2 \ln(6/\delta)/\varepsilon'^2)$ where $\varepsilon' = \varepsilon/4\sqrt{\ln(3/\delta)}$.
  - At center : Aggregate local estimates :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\mu}_i$$

Theorem

The two-round mean estimation algorithm is user-level (ε, δ)-LDP. Moreover, the output $\hat{\mu}$ of the algorithm satisfies

$$\mathbb{P}\left(|\hat{\mu} - \mu| \geq C\left(\frac{B}{\sqrt{mn}}\sqrt{\ln\frac{1}{\xi}} + \frac{B}{\sqrt{mn}\varepsilon}\ln\frac{n}{\xi}\ln\frac{1}{\delta}\right)\right) \leq \xi,$$

provided $n = \tilde{\Omega}(1/\epsilon)$.

Results for the multivariate case

- Running the univariate algorithm coordinate-wise leads to an error of $\tilde{O}(d/\sqrt{mn}\varepsilon)$
- This can be improved to $\tilde{O}(\sqrt{d}/\sqrt{mn}\varepsilon)$ by using a preprocessing step

Results for the multivariate case

- Running the univariate algorithm coordinate-wise leads to an error of $\tilde{O}(d/\sqrt{mn}\varepsilon)$
- This can be improved to $\tilde{O}(\sqrt{d}/\sqrt{mn}\varepsilon)$ by using a preprocessing step
- Random rotation trick : Rotate local averages using matrix HD where H is a $d \times d$ Hadamard matrix and D is diagonal with i.i.d. Rademacher entries
- The rotation ensures that $\|HDy_i\|_\infty = \tilde{O}(B/\sqrt{d})$ for all $i \in [n]$ with high probability

Results for the multivariate case

Theorem

The algorithm described before is user-level (ε, δ)-LDP. Further, provided $n = \tilde{\Omega}(\sqrt{d}/\varepsilon)$, the output $\hat{\mu}$ of the algorithm satisfies

$$\|\hat{\mu} - \mu\|_2 = O\left(\frac{B}{\sqrt{mn}} \ln \frac{nd}{\xi} + \frac{B\sqrt{d}}{\sqrt{mn}\varepsilon}\left(\ln \frac{nd}{\xi} \ln \frac{d}{\delta}\right)^{1.5}\right),$$

with probability at least $1 - \xi$.

Back to ERM under user-level local privacy constraints

User-level LDP ERM

- The update rule can be rewritten as

$$\theta_{t+1} = \theta_t - \frac{\eta_t}{mn} \sum_{i,j} g_j^{(i)}(\theta_t) - \eta_t Z_{1,t} + \eta_t Z_{2,t}$$

where

$$Z_{1,t} = \hat{g}(\theta_t) - \mathbb{E}[g_j^{(i)}(\theta_t)]$$
$$Z_{2,t} = \frac{1}{mn} \sum_{i,j} g_j^{(i)}(\theta_t) - \mathbb{E}[g_j^{(i)}(\theta_t)]$$

Bounding the noise terms

- We want an upper bound on

$$\|Z_{1,t}\|_2 = \|\hat{g}(\theta_t) - \mathbb{E}[g_j^{(i)}(\theta_t)]\|_2$$

that holds for all $t \in [T]$

Bounding the noise terms

- We want an upper bound on

$$\|Z_{1,t}\|_2 = \|\hat{g}(\theta_t) - \mathbb{E}[g_j^{(i)}(\theta_t)]\|_2$$

that holds for all $t \in [T]$

- For a fixed $\theta \in \Theta$,

$$\|\hat{g}(\theta) - \mathbb{E}[g_j^{(i)}(\theta)]\|_2 = \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{mn\varepsilon}}\right)$$

with high probability (guarantee of the mean estimation algorithm)

Bounding the noise terms

- We want an upper bound on

$$\|Z_{1,t}\|_2 = \|\hat{g}(\theta_t) - \mathbb{E}[g_j^{(i)}(\theta_t)]\|_2$$

  that holds for all $t \in [T]$

- For a fixed $\theta \in \Theta$,

$$\|\hat{g}(\theta) - \mathbb{E}[g_j^{(i)}(\theta)]\|_2 = \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{mn}\varepsilon}\right)$$

  with high probability (guarantee of the mean estimation algorithm)

- But cannot use this guarantee for θ_t since the inputs $\{g^{(i)}(\theta_t)\}$ to the mean estimation algorithm are not independent anymore

Bounding the noise term : key steps

- Let Γ be a Δ-net for Θ. Using union bound

$$\mathbb{P}\left(\sup_{\theta}\|\hat{g}(\theta)-\mathbb{E}[g_j^{(i)}(\theta)]\|_2 \geq C\frac{B\sqrt{d}}{\sqrt{mn}\varepsilon}\left(\ln\frac{nd|\Gamma|}{\xi}\right)^{1.5}\ln\frac{d}{\delta}\right) \leq \xi$$

Bounding the noise term : key steps

- Let Γ be a Δ-net for Θ. Using union bound

$$\mathbb{P}\left(\sup_{\theta} \|\hat{g}(\theta) - \mathbb{E}[g_j^{(i)}(\theta)]\|_2 \geq C \frac{B\sqrt{d}}{\sqrt{mn}\varepsilon}\left(\ln \frac{nd|\Gamma|}{\xi}\right)^{1.5} \ln \frac{d}{\delta}\right) \leq \xi$$

- With probability at least $1 - \xi$,

$$\|Z_{1,t}\|_2 = O\left(\frac{B\sqrt{d}}{\sqrt{mn}\varepsilon}\left(\ln \frac{nd}{\xi} + d\ln\left(1 + \frac{\tau\sqrt{mn}\varepsilon}{d^2}\right)\right)^{1.5} \ln \frac{d}{\delta}\right) = r_{n,m}$$

provided $n = \tilde{\Omega}(\sqrt{d}/\varepsilon)$

Bounding the noise term : key steps

- Let Γ be a Δ-net for Θ. Using union bound

$$\mathbb{P}\left(\sup_{\theta} \|\hat{g}(\theta) - \mathbb{E}[g_j^{(i)}(\theta)]\|_2 \geq C\frac{B\sqrt{d}}{\sqrt{mn}\varepsilon}\left(\ln\frac{nd|\Gamma|}{\xi}\right)^{1.5}\ln\frac{d}{\delta}\right) \leq \xi$$

- With probability at least $1-\xi$,

$$\|Z_{1,t}\|_2 = O\left(\frac{B\sqrt{d}}{\sqrt{mn}\varepsilon}\left(\ln\frac{nd}{\xi} + d\ln\left(1+\frac{\tau\sqrt{mn}\varepsilon}{d^2}\right)\right)^{1.5}\ln\frac{d}{\delta}\right) = r_{n,m}$$

provided $n = \tilde{\Omega}(\sqrt{d}/\varepsilon)$

- Convergence of θ_t follows analysis of noisy gradient descent similar to the one seen in the central model.

Guarantees for user-level LDP ERM

- Noisy gradient descent :

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

Theorem

Suppose $\mathcal{L}_{n,m}$ is locally τ_1-strongly convex and τ_2-smooth. Further let $\eta \leq \frac{1}{2}\min\left\{\frac{1}{\tau_2}, 1\right\}$, $\sqrt{mn} = \tilde{\Omega}(Bd^2/\varepsilon)$ and $T = \Omega(\log n)$. Then, with probability at least $1-\xi$,

$$\|\theta_T - \hat{\theta}\|_2 \leq C\sqrt{T}r_{n,m},$$

where C is a constant depending on B, τ_1, τ_2, and η.

References

- M. Avella-Medina, C. Bradshaw & P.L. Loh (2023) "Differentially private inference via noisy optimization." *Annals of Statistics*
- L. Ramesh, E. Han, M. Avella-Medina, & C. Rush (2023) "M-estimators under user-level local differential privacy constraints." *ArXiv (soon !)*

Thank you !

Questions ? ? ?