Differentially private inference via noisy optimization

Marco Avella Medina
joint with Casey Bradshaw and Po-Ling Loh

Statistical Learning and Differential Privacy Workshop, Bath UK

September 12, 2022

Columbia University in the City of New York

Motivation

- Study private counterparts of most commonly implemented algorithms for M-estimators in statistical software.
- Lack of general differentially private tools for parametric inference.
- Establish connections between privacy-preserving data analysis and robust statistics

Our contribution

joint work with Casey Bradshaw and Po-Ling Loh

- 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 : recent work including Wang, Kifer and Lee (2019) proposes a similar technique. Other work Sheffet (2017), Karwa and Vadhan (2017), Barrientos et al. (2019), Canonne et al. (2019), Avella-Medina (2021)...

Anonymized data ?

- Latanya Sweeney showed that gender, date of birth, and ZIP code are sufficient to uniquely identify the vast majority of Americans. In 1997 she identified the Governor of Massachusetts in a public anonymous database and sent him his own personal health record to his office !
- Narayanan and Shmatikov (2008, SP) show that anonymization fails even when combined with sanitization. Successfully de-anonymized Netflix data and caused cancellation of second Netflix prize
  - Problem : auxiliary information and linkage attacks
  - We can't know what adversary knows or will know in the future.

Summary statistics can reveal individual information

- Homer et al. 2008 showed that commonly released minor allele frequencies (MAFs) i.e. sample means are not private.
- The plots below are taken from Zhang & Zhang (2020). They illustrate the problem with a heart disease data set consisting of 100 patients and 347,019 SNPs.

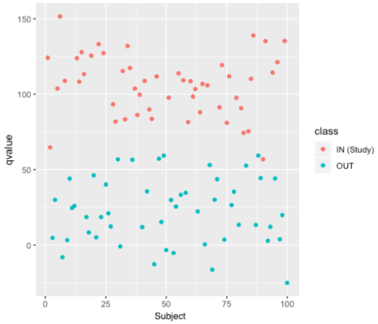


FIGURE – Standard q-score

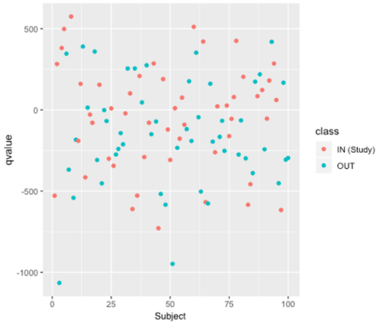


FIGURE – DP q-scores

Summary statistics can reveal individual information

- Homer et al. 2008 showed that commonly released minor allele frequencies (MAFs) i.e. sample means are not private.
- The plots below are taken from Zhang & Zhang (2020). They illustrate the problem with a heart disease data set consisting of 100 patients and 347,019 SNPs.

$\bar{X}^T y_i$ $\qquad$ $(\bar{X} + \text{noise})^T y_i$

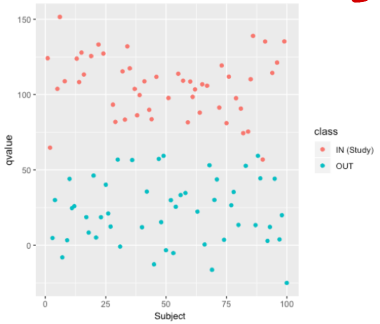


Figure – Standard q-score

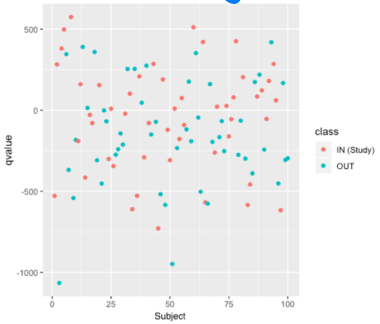
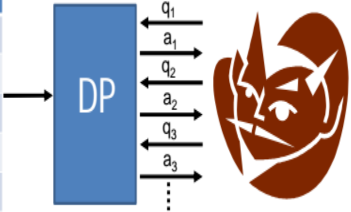


Figure – DP q-scores

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 $N(\mu,1)$

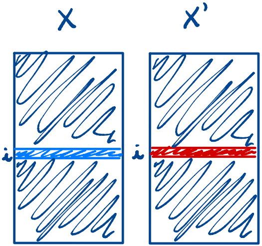
- New intuitive definition of differential privacy via hypothesis testing
  - Gaussian mechanism : $\tilde{m}(x_1,\ldots,x_n) = m(x_1,\ldots,x_n) + \frac{1}{\mu}\mathrm{GS}(m)N(0,1)$
  - Gaussian differential privacy : $H_0 : P = N(0,1)$ V. $H_1 : P = N(\mu,1)$

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 : $\tilde{m}(x_1,\dots,x_n) = m(x_1,\dots,x_n) + \frac{1}{\mu}\mathrm{GS}(m)N(0,1)$
  - Gaussian differential privacy : $H_0 : P = N(0,1)$ V. $H_1 : P = N(\mu,1)$



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

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 : $G_{\mu_1} \otimes G_{\mu_2} \cdots \otimes G_{\mu_K} = G_{\sqrt{\sum_{k=1}^K \mu_k^2}}$
  - Universality (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)

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

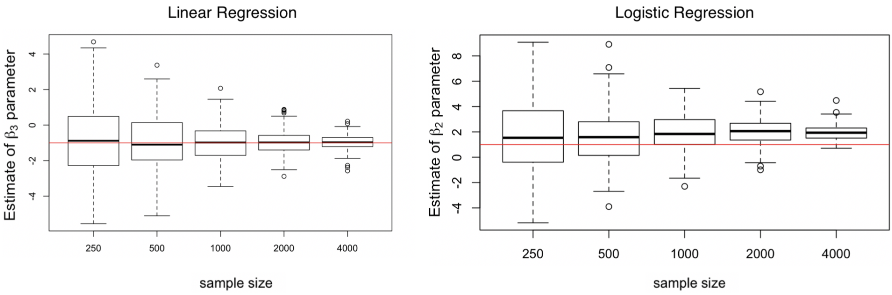
$= \frac{1}{\mu}$

A remark on clipping

- Clipped likelihood as M-estimator

$$\tilde{\theta} : \frac{1}{n}\sum_{i=1}^{n} h_c\left(\nabla \log f(x_i;\tilde{\theta})\right) = 0,$$

where $h_c(z) = z\min\{1, \frac{c}{\|z\|_2}\}$ is the multivariate Huber function.



Example : linear regression

- Consider a linear regression model

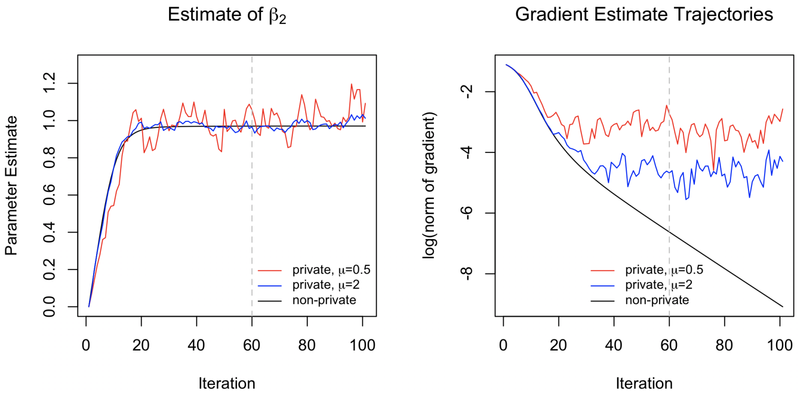
$$\begin{aligned} y_i &= x_i^T\beta + u_i \text{ for } i = 1, \ldots, n \\ x_i &\in \mathbb{R}^p \\ u_i &\sim N(0, \sigma^2) \end{aligned}$$

- We want to solve

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

where $w(x_i) = \min\left(1, \frac{1}{\|x_i\|_2^2}\right)$ and κ is a Fisher consistency constant.

Example : linear regression



Noisy Newton

- Noisy Newton :

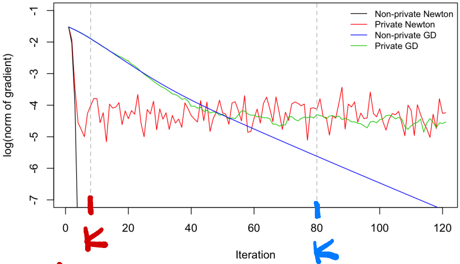
$$\theta^{(k+1)} = \theta^{(k)} - \left(\frac{1}{n}\sum_{i=1}^{n}\dot{\Psi}(x_i,\theta^{(k)}) + \frac{2\bar{B}\sqrt{2K}}{\mu n}W_k\right)^{-1} \cdot \left(\frac{1}{n}\sum_{i=1}^{n}\Psi(x_i,\theta^{(k)}) + \frac{2B\sqrt{2K}}{\mu n}N_k\right)$$

  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

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

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

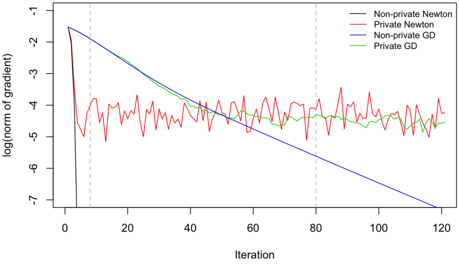
Noisy Newton theory



K for noisy Newton

K for noisy GD

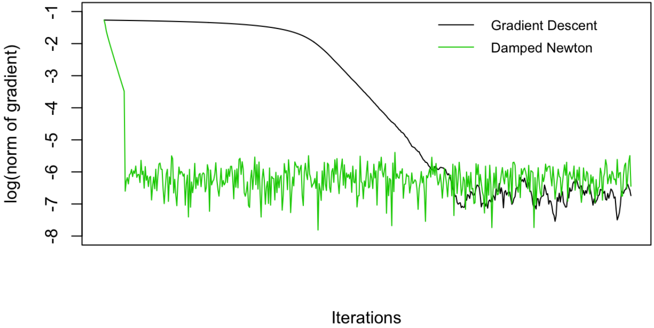
Noisy Newton theory



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. $\sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F))$

Damped Newton V. NGD



- Pure Newton threshold :
  - Local strong convexity : $\|\nabla \mathcal{L}_n(\theta^{(0)})\| \leq \frac{\tau_1^2}{L}$
  - Self-concordance : $\lambda_{\min}^{-1/2}(\nabla^2 \mathcal{L}_n(\theta^{(0)}))\lambda(\theta^{(0)}) \leq \frac{1}{16\gamma}$.

Self-concordance

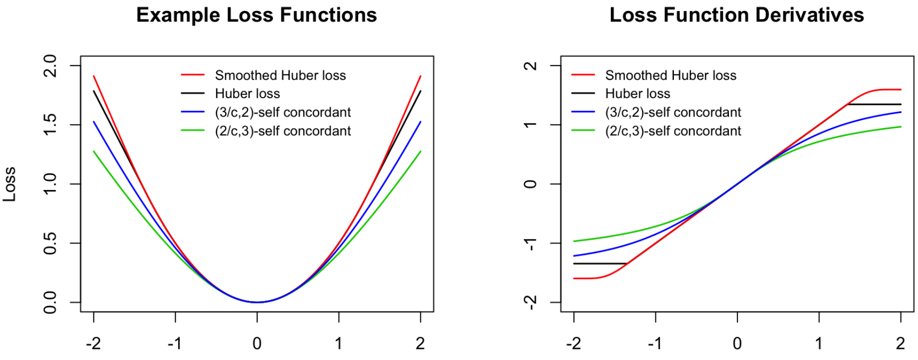
A univariate function $f : \mathbb{R} \to \mathbb{R}$ is (γ, ν)-*self-concordant* if

$$|f'''(x)| \leq \gamma \left(f''(x)\right)^{\nu/2},$$

for all x. A multivariate function $f : \mathbb{R}^p \to \mathbb{R}$ is (γ, ν)-self-concordant if

$$\left|\langle \nabla^3 f(x)[v]u, u\rangle\right| \leq \gamma \|u\|^2_{\nabla^2 f(x)} \|v\|^{\nu-2}_{\nabla^2 f(x)} \|v\|_2^{3-\nu},$$

for all $x, u, v \in \mathbb{R}^p$.



Asymptotic variance

What do the terms in the variance formula look like for our linear regression example ?

$$\begin{aligned} Q_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \psi_c^2\left(\frac{y_i - x_i^T\theta}{\sigma}\right) w(x_i)^2 x_i x_i^\top \\ &= \frac{1}{n}\sum_{i=1}^{n} z_i z_i^\top \end{aligned}$$

$$\begin{aligned} M_n(\theta) &= \frac{1}{n\sigma}\sum_{i=1}^{n} \dot{\psi}_c\left(\frac{y_i - x_i^\top\theta}{\sigma}\right) w(x_i) x_i x_i^\top \\ &= \frac{1}{n}\sum_{i=1}^{n} \tilde{z}_i \tilde{z}_i^\top \end{aligned}$$

where $\|z_i\| \leq B$ and $\|\tilde{z}_i\| \leq \bar{B}$.

Private sandwich formula

1. Plug private estimators $\theta^{(K)}$ and $\sigma^{(K)}$ in M_n and Q_n.
2. Matrix Gaussian mechanism : add symmetric matrix with i.i.d. Gaussians in upper triangular part of the matrix. (Dwork et al. 2014, STOC)

$$\tilde{M}_n(\theta^{(K)}) = M_n(\theta^{(K)}) + \frac{2\bar{B}}{\mu n} G_1 \quad \text{and} \quad \tilde{Q}_n(\theta^{(K)}) = Q_n(\theta^{(K)}) + \frac{2B^2}{\mu n} G_2$$

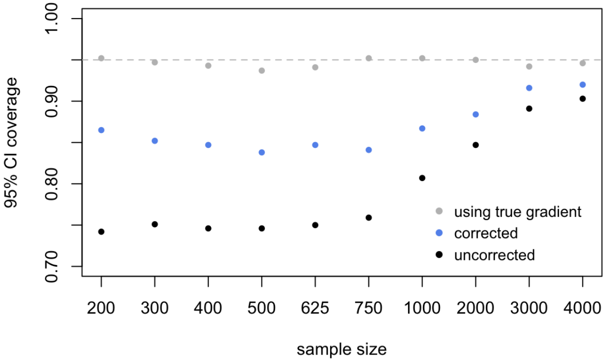
3. Compute $V_n(\theta^{(K)}) = \tilde{M}_n(\theta^{(K)})^{-1} \tilde{Q}_n(\theta^{(K)}) \tilde{M}_n(\theta^{(K)})^{-1}$

Proposition. $V_n(\theta^{(K)})$ is $\sqrt{3}\mu$-GDP and $\tilde{V}_n(\theta^{(K)}) \to_p V(\theta_0)$.

GDP Confidence Interval Coverage

Corrected variance formula :

$$\hat{V}_n(\theta^{(K)}) = \tilde{V}_n(\theta^{(K)}) + \frac{8\eta^2 B^2 K}{n\mu^2} I.$$



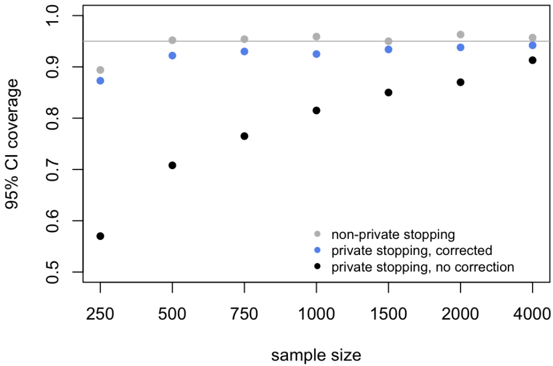
GDP Confidence Interval Coverage

Corrected variance formula for noisy Newton :

$$\hat{V}_n(\theta^{(K)}) = \tilde{V}_n(\theta^{(K)}) + nC_{Newton},$$

where

$$C_{Newton} := \eta^2 \left\{ \nabla^2 \mathcal{L}_n(\theta^{(k)}) + \tilde{W}_k \right\}^{-1} \left(\frac{2B\sqrt{2K}}{\mu n} \right)^2 \left\{ \nabla^2 \mathcal{L}_n(\theta^{(k)}) + \tilde{W}_k \right\}^{-1}.$$



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.

References

- M. Avella-Medina, C. Bradshaw & P.L. Loh (2021) "Differentially private inference via noisy optimization." (*arXiv*)

Thank you !

Questions ? ? ?