Differentially private statistical inference

Marco Avella Medina

CUSO Winter School, Les Diablerets

Feb 2-5, 2025

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Some references

- Cynthia Dwork and Aaron Roth (2014) “The algorithmic foundations of differential privacy”. *Foundations and Trends in Theoretical Computer Science*.
- Salil Vadhan (2017) “The complexity of differential privacy”. *Tutorials on the Foundations of Cryptography*.
- Adam Smith and Jonathan Ullman (2023) “Privacy in Statistics and Machine Learning” *Lectures notes and videos from course*.
- Gautam Kamath (2020, 2022) “Algorithms for Private Data Analysis” *Lectures notes and videos from course*.

Information control and privacy ?

- Conventional measures
  - Control access to information
  - Control for what the purpose the information is used
- Measures for data release ?
  - Anonymization
  - Sanitization : perturbation, suppression, generalization

Information control measures are not satisfactory...

Privacy attacks

Reidentification attacks

- In 1997 Latanya Sweeney showed that gender, date of birth, and ZIP code are sufficient to uniquely identify the vast majority of Americans (up to 87%). She could even identify the Governor of Massachusetts in a public anonymous hospital discharge records and sent him his own personal health record to his office !
- In 2013 in and her team also identified the names of more than 40% of a sample of anonymous participants in a high-profile DNA study, the Personal Genome Project.

Reconstruction attacks

Netflix Challenge Re-identification (Narayanan & Shmatikov 2008)

- Data : X = ratings on Netflix e.g. sparse, categorical and high-dimensional movie
- Adversary's input :
  - S=subset of individual observations, possibly slightly distorted
  - Y= public ratings on IMDB e.g auxiliary dataset Y containing information on certain individuals in dataset S.
- Adversary's goal : identify individuals in X by matching their records to those in Y.

Reconstruction attacks

Netflix Challenge Re-identification (Narayanan & Shmatikov 2008)

- Narayanan-Shmatikov Algorithm
  1. Calculate $score(Y, x_i)$ for each $x_i \in S$ as well as the standard deviation $\hat{\sigma}$ of the calculated scores.
  2. Let $r_1 = x_i$ and $r_2 = x_j$ be the records with the largest scores
  3. If $score(Y, r_1) - score(Y, r_2) > \phi\hat{\sigma}$, output r_1, else output "no match found".
- The authors recommend to use a score of the form

$$score(Y, x) = \min_i \sum_{j \in [n]} \frac{1}{\log |n_j|} d(y_{ij}, x_j),$$

where n is the number of individual records in Y, n_j is the number of users that rated movie j, and $d(y_{ij}, x_j)$ is a distance between y_j and x_j.

Reconstruction attacks

Netflix Challenge Re-identification (Narayanan & Shmatikov 2008)

Narayanan and Shmatikov (2008, SP) strikingly shows how anonymization fails even when combined with sanitization.

- Successfully de-anonymized Netflix data from individuals with public ratings on IMDB. Only approximate ratings and dates sufficed to identify individuals.
- They propose polynomial time algorithm that breaks privacy
- Problem : auxiliary information and linkage attacks
- This caused cancellation of second Netflix prize and resulted in a lawsuit.
- We can't know what adversary knows or will know in the future.

Membership attacks

- Genome wide association studies (Homer et al. 2008, PLoS genetics)
    - Release frequencies of SNP's (individual positions)
    - Determine whether individual "i" is in "case group" i.e. has a particular disease

Membership attacks

- Genome wide association studies (Homer et al. 2008, PLoS genetics)
  - Release frequencies of SNP's (individual positions)
  - Determine whether individual "i" is in "case group" i.e. has a particular disease
- Microtargeted ads (Korolova 2011, J. Privacy and Confidentiality)
  - Define a sufficiently narrow target profiles that allow to identify specific individuals.
  - Design k campaign adds on the likely unique subject that will see them and record impressions over a reasonable time period to infer a specific feature f_i mentioned in ith add. This strategy was succesfully deployed to figure out the age and sexual orientation of friends and friends' friends. Could be used to extract other features.

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.

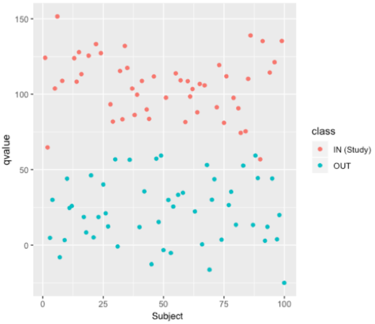


Figure – Standard q-score

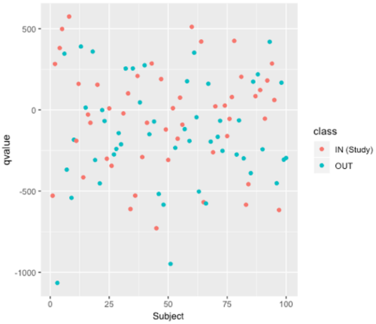


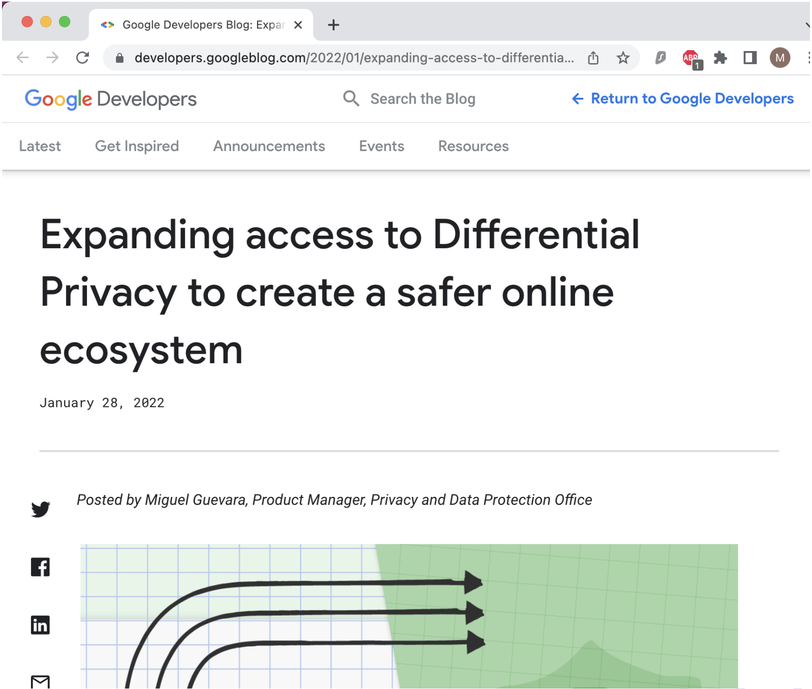
Figure – DP q-scores

Differential privacy : properties and basic algorithms

blogs.microsoft.com/on-the-issues/2020/12/10/differential-privacy-smartnoise-early-adopter-acceleration-progran
Microsoft
Microsoft On the Issues
Our Company
News and Stories
Topics
Cloud Principles
Pr

How differential privacy enhances Microsoft's privacy and security tools: SmartNoise Early Adopter Acceleration Program Launched

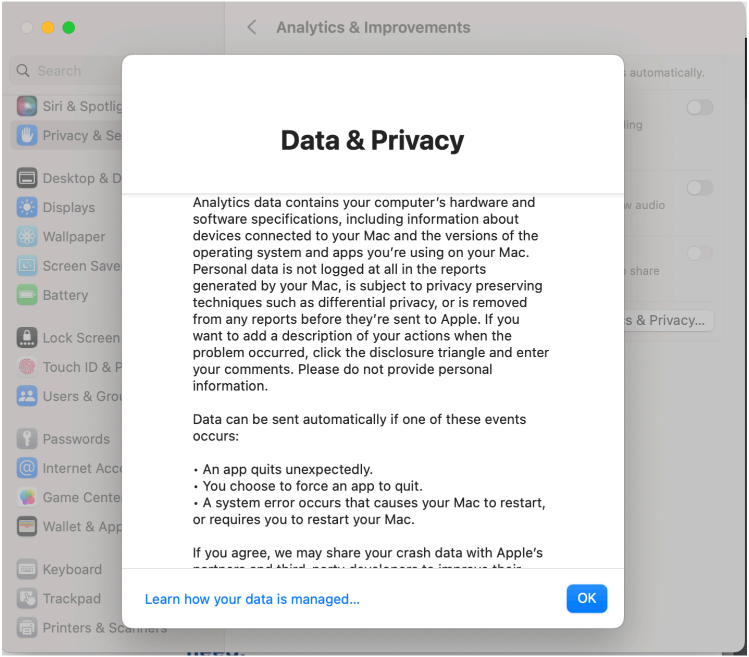
Dec 10, 2020 | John Kahan - VP, Chief Data Analytics Officer


Google Developers Blog: Expar
developers.googleblog.com/2022/01/expanding-access-to-differentia...
Google Developers
Search the Blog
Return to Google Developers
Latest
Get Inspired
Announcements
Events
Resources

Expanding access to Differential Privacy to create a safer online ecosystem

January 28, 2022

Posted by Miguel Guevara, Product Manager, Privacy and Data Protection Office



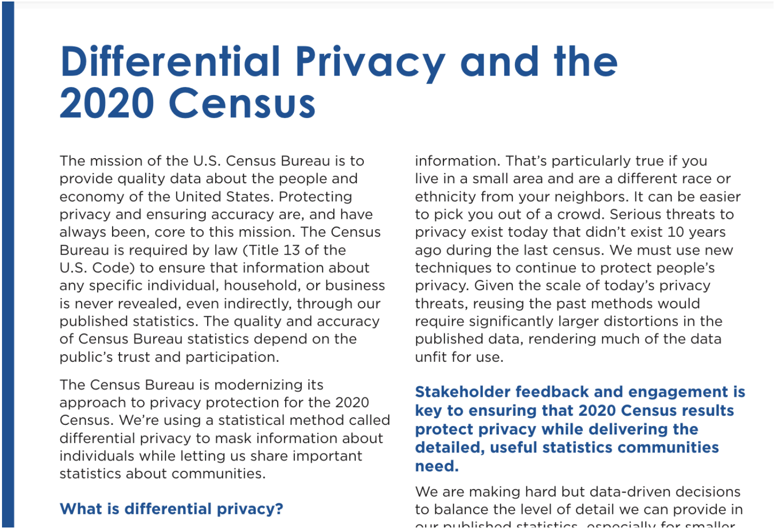
Data & Privacy

Analytics data contains your computer's hardware and software specifications, including information about devices connected to your Mac and the versions of the operating system and apps you're using on your Mac. Personal data is not logged at all in the reports generated by your Mac, is subject to privacy preserving techniques such as differential privacy, or is removed from any reports before they're sent to Apple. If you want to add a description of your actions when the problem occurred, click the disclosure triangle and enter your comments. Please do not provide personal information.

Data can be sent automatically if one of these events occurs:

- An app quits unexpectedly.
- You choose to force an app to quit.
- A system error occurs that causes your Mac to restart, or requires you to restart your Mac.

If you agree, we may share your crash data with Apple's partners and third-party developers to improve their



Differential Privacy and the 2020 Census

The mission of the U.S. Census Bureau is to provide quality data about the people and economy of the United States. Protecting privacy and ensuring accuracy are, and have always been, core to this mission. The Census Bureau is required by law (Title 13 of the U.S. Code) to ensure that information about any specific individual, household, or business is never revealed, even indirectly, through our published statistics. The quality and accuracy of Census Bureau statistics depend on the public's trust and participation.

The Census Bureau is modernizing its approach to privacy protection for the 2020 Census. We're using a statistical method called differential privacy to mask information about individuals while letting us share important statistics about communities.

What is differential privacy?

information. That's particularly true if you live in a small area and are a different race or ethnicity from your neighbors. It can be easier to pick you out of a crowd. Serious threats to privacy exist today that didn't exist 10 years ago during the last census. We must use new techniques to continue to protect people's privacy. Given the scale of today's privacy threats, reusing the past methods would require significantly larger distortions in the published data, rendering much of the data unfit for use.

Stakeholder feedback and engagement is key to ensuring that 2020 Census results protect privacy while delivering the detailed, useful statistics communities need.

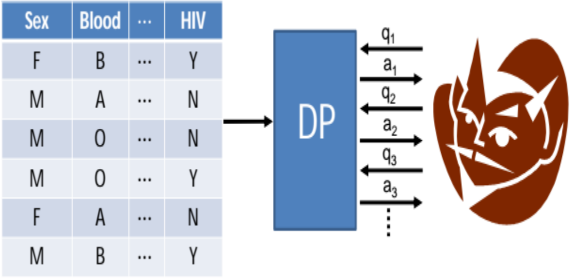
We are making hard but data-driven decisions to balance the level of detail we can provide in our published statistics, especially for smaller

Privacy definition (informal)

An analysis on the dataset D is private if Thibault knows almost no more about Marco after the analysis than what he would have known had he conducted the same analysis on an identical database with Marco's data removed.

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

Definition

Dwork et al. (2006)

Definition. Let $X_{1:n} = (X_1, \dots, X_n) \in \mathcal{X}^n$ be some dataset. A randomized function $h : \mathcal{X}^n \to \mathbb{R}^d$ is (ε, δ)-*differentially private* if for all pairs of datasets $(X_{1:n}, \tilde{X}_{1:n})$ with $d_H(X_{1:n}, \tilde{X}_{1:n}) = \sum_{i=1}^{n} \mathbb{1}(X_i \neq X_i') = 1$ and all measurable subsets of outputs $\mathcal{E}$:

$$\mathbb{P}(h(X_{1:n}) \in \mathcal{E}) \leq e^{\varepsilon} \mathbb{P}(h(\tilde{X}_{1:n}) \in \mathcal{E}) + \delta.$$

Definition

Dwork et al. (2006)

Definition. Let $X_{1:n} = (X_1, \ldots, X_n) \in \mathcal{X}^n$ be some dataset. A randomized function $h : \mathcal{X}^n \to \mathbb{R}^d$ is (ε, δ)-*differentially private* if for all pairs of datasets $(X_{1:n}, \tilde{X}_{1:n})$ with $d_H(X_{1:n}, \tilde{X}_{1:n}) = \sum_{i=1}^{n} \mathbb{1}(X_i \neq X_i') = 1$ and all measurable subsets of outputs $\mathcal{E}$:

$$\mathbb{P}(h(X_{1:n}) \in \mathcal{E}) \leq e^{\varepsilon} \mathbb{P}(h(\tilde{X}_{1:n}) \in \mathcal{E}) + \delta.$$

Remarks :

- Probabilities computed over randomness of the algorithms
- Ensures that one's participation in a survey will not in itself be disclosed, nor will participation lead to disclosure of any specifics that one has contributed to the survey.

Example : Bernoulli sample

Let $x_{1:n} = (x_1, \ldots, x_n) \in \{0,1\}^n$. In this case we can construct a simple randomized mean estimator

$$m(x_{1:n}) = \bar{x} + \frac{1}{\varepsilon n} Z,$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and Z is a r.v. with density function $f(z) = \frac{1}{2} e^{-\frac{1}{2}|z|}$, $z \in \mathbb{R}$. Indeed, $m(x_{1:n})$ is $(\varepsilon, 0)$-DP.

Remarks :

- If $x_{1:n}$ and $x'_{1:n}$ differ only at one entry, then $|\bar{x} - \bar{x}'| = \frac{1}{n}$.
- For fixed x, the density function of $m(x)$ is

$$f_{m(x)}(z) = \frac{\varepsilon n}{2} e^{-\frac{\varepsilon n}{2}|z - \bar{x}|}$$

Proposition (Laplace mechanism)

For $f : \mathcal{X}^n \to R^p$ the global sensitivity of a deterministic function f as

$$GS_1(f) := \sup_{X_{1:n}, \tilde{X}_{1:n}, d_H(X_{1:n}, \tilde{X}'_{1:n})=1} \|f(X_{1:n}) - f(\tilde{X}_{1:n})\|_1.$$

Then, the following output is is $(\varepsilon, 0)$-DP :

$$A_f(X_{1:n}) = f(X_{1:n}) + \frac{GS_1(f)}{\varepsilon} Lap_p(1).$$

Proposition (Laplace mechanism)

For $f : \mathcal{X}^n \to R^p$ the global sensitivity of a deterministic function f as

$$GS_1(f) := \sup_{X_{1:n}, \tilde{X}_{1:n}, d_H(X_{1:n}, \tilde{X}'_{1:n})=1} \|f(X_{1:n}) - f(\tilde{X}_{1:n})\|_1.$$

Then, the following output is is $(\varepsilon, 0)$-DP :

$$A_f(X_{1:n}) = f(X_{1:n}) + \frac{GS_1(f)}{\varepsilon} Lap_p(1).$$

Proof. Denote the densities of $A_f(X_{1:n})$ and $A_f(\tilde{X}_{1:n})$ by $g_{A_f(X)}$ and $g_{A_f(\tilde{X})}$. Note that as long as $d_H(X_{1:n}, \tilde{X}_{1:n}) = 1$,

$$\begin{aligned}
\frac{g_{A_f(X)}(y)}{g_{A_f(\tilde{X})}(y)} &= \frac{\exp(-\|y - f(X_{1:n})\|_1 \varepsilon / GS_1(f))}{\exp(-\|y - f(\tilde{X}_{1:n})\|_1 \varepsilon / GS_1(f))} \\
&= \exp(\frac{\varepsilon}{GS_1(f)}(\|y - f(\tilde{X}_{1:n})\|_1 - \|y - f(X_{1:n})\|_1) \\
&\leq \exp(\frac{\varepsilon}{GS_1(f)}(\|f(X_{1:n}) - f(\tilde{X}_{1:n})\|_1) \leq e^{\varepsilon}.
\end{aligned}$$

Hence, $\int_O g_{A_f(X)}(y)dy \leq e^{\varepsilon} \int_O g_{A_f(\tilde{X})}(y)dy$. □

Two important properties

Proposition (Postprocessing)

Let $A : \mathcal{X}^n \mapsto \mathbb{R}^m$ be a randomized algorithm that is (ε, δ)-differentially private. Let $f : \mathbb{R}^m \mapsto \mathbb{R}^d$ be an arbitrary mapping. Then $f \circ A : \mathcal{X}^n \mapsto \mathbb{R}^d$ is (ε, δ)-differentially private.

Two important properties

Proposition (Postprocessing)

Let $A : \mathcal{X}^n \mapsto \mathbb{R}^m$ be a randomized algorithm that is (ε, δ)-differentially private. Let $f : \mathbb{R}^m \mapsto \mathbb{R}^d$ be an arbitrary mapping. Then $f \circ A : \mathcal{X}^n \mapsto \mathbb{R}^d$ is (ε, δ)-differentially private.

Theorem (Composition)

Let $A_1 : \mathcal{X}^n \mapsto \mathbb{R}^m$ be a $(\varepsilon_1, \delta_1)$-differentially private algorithm, and let $A_2 : \mathcal{X}^n \mapsto \mathbb{R}^d$ be a $(\varepsilon_2, \delta_2)$-differentially private algorithm. Then their combination, defined to be $A_{1,2} : \mathcal{X}^n \mapsto \mathbb{R}^m \times \mathbb{R}^d$ by the mapping $A_{1,2}(x) = (A_1(x), A_2(x))$ is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-differentially private.

Example : Bounded data

Let $x_{1:n} = (x_1, \ldots, x_n) \in [0, B]^n$ and suppose we want to release private estimators of the mean and the variance. Consider the noisy empirical moments :

$$\hat{m}_1(x_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{B}{\varepsilon n} Z_1 \quad \text{and} \quad \hat{m}_2(x_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 + \frac{B^2}{\varepsilon n} Z_2,$$

where Z_1 and Z_2 are standard independent Laplace.

- Both $\hat{m}_1(x_{1:n})$ and $\hat{m}_2(x_{1:n})$ are $(\varepsilon, 0)$-DP by the Laplace mechanism.
- Releasing $(\hat{m}_1(x_{1:n}), \hat{m}_2(x_{1:n}))$ is $(2\varepsilon, 0)$-DP by composition.
- Hence, releasing $(\hat{m}_1(x_{1:n}), \hat{m}_2(x_{1:n}) - n\hat{m}_1(x_{1:n})^2)$ is $(2\varepsilon, 0)$-DP by post-processing.

Example : noisy histogram

Let $x_{1:n} = (x_1, \dots, x_n) \in [0,1]^n$. Partition $[0,1]$ in bins of equal bandwidth h_n s.t. $k_n = h_n^{-1}$ is an integer. Let $B_j := [(j-1)h_n, jh_n)$, $j = 1, 2, \dots, k_n$. The histogram density estimator is

$$f_{h_n}(x) = \frac{1}{h_n} \sum_{j=1}^{k_n} \frac{n_j}{n} \mathbb{1}(x \in B_j),$$

where $n_j := \sum_{i=1}^{n} \mathbb{1}(x_i \in B_j)$.

Example : noisy histogram

Let $x_{1:n} = (x_1, \ldots, x_n) \in [0,1]^n$. Partition $[0,1]$ in bins of equal bandwidth h_n s.t. $k_n = h_n^{-1}$ is an integer. Let $B_j := [(j-1)h_n, jh_n)$, $j = 1, 2, \ldots, k_n$. The histogram density estimator is

$$f_{h_n}(x) = \frac{1}{h_n} \sum_{j=1}^{k_n} \frac{n_j}{n} \mathbb{1}(x \in B_j),$$

where $n_j := \sum_{i=1}^{n} \mathbb{1}(x_i \in B_j)$.

- Noisy counts :

  $$\hat{n}_j = n_j + z_j, \ \forall j \in \{1, \ldots, k_n\},$$

  where the z_j's are i.i.d. $\frac{1}{\varepsilon} Lap(1)$. k_n is the total number of cells. Note that $GS(n_j) = 1$. So by Laplace mechanism $\hat{n}_j$ is $(\varepsilon, 0)$-DP.

Example : noisy histogram

Let $x_{1:n} = (x_1, \dots, x_n) \in [0,1]^n$. Partition $[0,1]$ in bins of equal bandwidth h_n s.t. $k_n = h_n^{-1}$ is an integer. Let $B_j := [(j-1)h_n, jh_n)$, $j = 1, 2, \dots, k_n$. The histogram density estimator is

$$f_{h_n}(x) = \frac{1}{h_n} \sum_{j=1}^{k_n} \frac{n_j}{n} \mathbb{1}(x \in B_j),$$

where $n_j := \sum_{i=1}^{n} \mathbb{1}(x_i \in B_j)$.

- Noisy counts :

  $$\hat{n}_j = n_j + z_j, \ \forall j \in \{1, \dots, k_n\},$$

  where the z_j's are i.i.d. $\frac{1}{\varepsilon} Lap(1)$. k_n is the total number of cells. Note that $GS(n_j) = 1$. So by Laplace mechanism $\hat{n}_j$ is $(\varepsilon, 0)$-DP.
- Noisy histogram :

  $$\hat{f}_{h_n}(x) = \frac{1}{h_n} \sum_{j=1}^{k_n} \frac{\hat{n}_j}{n} \mathbb{1}(x \in B_j)$$

  is $(\varepsilon, 0)$-DP by post-processing. Releasing all $\tilde{f}_{h_n}$ is $(k_n\varepsilon, 0)$-DP.

Example : mean estimation with unbounded data

Suppose $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Perhaps the most popular approach truncating the data as $y_i = \operatorname{sign}(x_i) \max(|x_i|, B_n)$ for $i = 1, \ldots, n$ and some known bound $B_n < \infty$. Compute the $(\varepsilon, 0)$-DP mean estimator

$$\hat{m}_{B_n}(y_{1:n}) = \bar{y} + \frac{2B_n}{\varepsilon n} Z, \tag{1}$$

where Z is a standard Laplace random variable.

Example : mean estimation with unbounded data

Suppose $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Perhaps the most popular approach truncating the data as $y_i = \operatorname{sign}(x_i) \max(|x_i|, B_n)$ for $i = 1, \dots, n$ and some known bound $B_n < \infty$. Compute the $(\varepsilon, 0)$-DP mean estimator

$$\hat{m}_{B_n}(y_{1:n}) = \bar{y} + \frac{2B_n}{\varepsilon n} Z, \tag{1}$$

where Z is a standard Laplace random variable.

- If $\{X_i\} \overset{iid}{\sim} N(\mu, \sigma^2)$, a natural bound is $B_n = C\sigma\sqrt{\log n}$ for some $C > 0$ since the maximum of n normal random variables will lie in the interval $[\mu - 2\sigma\sqrt{K \log n}, \mu + 2\sigma\sqrt{K \log n}]$, with probability at least $1 - \frac{2}{n^K}$ for a positive constant K
- $\hat{m}_{B_n}(y_{1:n})$ is nice theoretically but in practice depends on the choice of the unknown constants σ and C. One could have some initial DP estimates of these parameters or have an idea of the range of these parameters (e.g. a known upper bound on σ and a known interval containing μ).

DP mechanisms

Proposition (Laplace mechanism)

For $f : \mathcal{X}^n \to R^p$ the global sensitivity of f as

$$GS_1(f) := \sup_{X_{1:n},\tilde{X}_{1:n}D',d_H(X_{1:n},\tilde{X}'_{1:n})=1} \|f(X_{1:n}) - f(\tilde{X}_{1:n})\|_1 .$$

Then, the following output is is $(\varepsilon, 0)$-DP :

$$A_f(X_{1:n}) = f(X_{1:n}) + \frac{GS_1(f)}{\varepsilon} Lap_p(1).$$

Proposition (Laplace mechanism)

For $f : \mathcal{X}^n \to R^p$ the global sensitivity of f as

$$GS_1(f) := \sup_{X_{1:n},\tilde{X}_{1:n}D', d_H(X_{1:n},\tilde{X}'_{1:n})=1} \|f(X_{1:n}) - f(\tilde{X}_{1:n})\|_1.$$

Then, the following output is is $(\varepsilon, 0)$-DP :

$$A_f(X_{1:n}) = f(X_{1:n}) + \frac{GS_1(f)}{\varepsilon} Lap_p(1).$$

Proof. Denote the densities of $A_f(X_{1:n})$ and $A_f(\tilde{X}_{1:n})$ by $g_{A_f(X)}$ and $g_{A_f(\tilde{X})}$. Note that as long as $d_H(X_{1:n}, \tilde{X}_{1:n}) = 1$,

$$\begin{aligned}\frac{g_{A_f(X)}(y)}{g_{A_f(\tilde{X})}(y)} &= \frac{\exp(-\|y - f(X_{1:n})\|_1 \varepsilon / GS_1(f))}{\exp(-\|y - f(\tilde{X}_{1:n})\|_1 \varepsilon / GS_1(f))} \\ &= \exp(\frac{\varepsilon}{GS_1(f)}(\|y - f(\tilde{X}_{1:n})\|_1 - \|y - f(X_{1:n})\|_1) \\ &\leq \exp(\frac{\varepsilon}{GS_1(f)}(\|f(X_{1:n}) - f(\tilde{X}_{1:n})\|_1) \leq e^{\varepsilon}.\end{aligned}$$

Hence, $\int_O g_{A_f(X)}(y)dy \leq e^{\varepsilon} \int_O g_{A_f(\tilde{X})}(y)dy$. □

Proposition (Gaussian mechanism)

For $f : \mathcal{X}^n \to R^p$ the global sensitivity of f as

$$GS_2(f) := \sup_{X_{1:n},\tilde{X}_{1:n},d_H(X_{1:n},\tilde{X}'_{1:n})=1} \|f(X_{1:n}) - f(\tilde{X}_{1:n})\|_2.$$

Then, for $\varepsilon \in (0,1)$ and $\sigma = \frac{\sqrt{2\ln(1.25/\delta)}GS_2(f)}{\varepsilon}$ the following output is is (ε, δ)-DP :

$$A_f(X_{1:n}) = f(X_{1:n}) + \; N(0, \sigma^2 I).$$

Proposition (Gaussian mechanism)

For $f : \mathcal{X}^n \to R^p$ the global sensitivity of f as

$$GS_2(f) := \sup_{X_{1:n},\tilde{X}_{1:n},d_H(X_{1:n},\tilde{X}'_{1:n})=1} \|f(X_{1:n}) - f(\tilde{X}_{1:n})\|_2.$$

Then, for $\varepsilon \in (0,1)$ and $\sigma = \frac{\sqrt{2\ln(1.25/\delta)}GS_2(f)}{\varepsilon}$ the following output is is (ε,δ)-DP :

$$A_f(X_{1:n}) = f(X_{1:n}) + \ N(0,\sigma^2 I).$$

Proof. Denote the densities of $A_f(X_{1:n})$ and $A_f(\tilde{X}_{1:n})$ by $g_{A_f(X)}$ and $g_{A_f(\tilde{X})}$ respectively. Note that as long as $d_H(X_{1:n},\tilde{X}'_{1:n}) = 1$,

$$\begin{aligned}\frac{g_{A_f(X)}(y)}{g_{A_f(\tilde{X})}(y)} &= \frac{\exp(-\frac{\varepsilon^2}{2GS_2(f)^2)}\|y - f(X_{1:n})\|_2^2)}{\exp(-\frac{\varepsilon^2}{2GS_2(f)^2)}\|y - f(\tilde{X}_{1:n})\|_2^2)} \\ &\leq \dots \\ &\leq e^{\varepsilon} + \delta.\end{aligned}$$

Hence, $\int_O g_{A_f(X)}(y)dy \leq e^{\varepsilon} \int_O g_{A_f(\tilde{X})}(y)dy + \delta$. (See Appendix A of book by Dwork and Roth)

Exponential Mechanism

The following sampling procedure leads to $(\varepsilon, 0)$-DP releases :

- Define a utility function $u : \mathcal{X}^n \times \Theta \mapsto \mathbb{R}$, mapping database/output pairs to utility scores. The higher the better for the user.
- $$\Delta u = \max_{\theta \in \Theta} \max_{d_H(x_{1:n}, \tilde{x}_{1:n})=1} |u(x_{1:n}, \theta) - u(\tilde{x}_{1:n}, \theta)|$$
- Select and output an element $\theta \in \Theta$ with probability proportional to $\exp(\frac{\varepsilon u(x_{1:n},\theta)}{2\Delta u})$.

Remark : in a statistical context, the utility function could be the likelihood function or the square of the score function.

Weaker notions of sensitivity

For $f : \mathcal{X}^n \to R^p$ the global sensitivity of f is

$$GS_2(f) := \sup_{x_{1:n}, \tilde{x}_{1:n}, d_H(x_{1:n}, \tilde{x}_{1:n})=1} \|f(x_{1:n}) - f(\tilde{x}_{1:n})\|_2.$$

The Laplace mechanism idea can be adapted for a notion of local sensitivity. For $\xi > 0$, the *ξ-smooth sensitivity* of f at $x_{1:n}$ is

$$SS_\xi(f, x_{1:n}) := \sup_{\tilde{x}_{1:n}} \left\{ e^{-\xi d_H(x_{1:n}, \tilde{x}_{1:n})} LS_2(f, \tilde{x}_{1:n}) \right\},$$

where

$$LS_2(f, x_{1:n}) := \sup_{\tilde{x}_{1:n}, d_H(x_{1:n}, \tilde{x}_{1:n})=1} \|f(x_{1:n}) - f(\tilde{x}_{1:n})\|_2$$

Weaker notions of sensitivity

For $f : \mathcal{X}^n \to R^p$ the global sensitivity of f is

$$GS_2(f) := \sup_{x_{1:n}, \tilde{x}_{1:n}, d_H(x_{1:n}, \tilde{x}_{1:n})=1} \|f(x_{1:n}) - f(\tilde{x}_{1:n})\|_2.$$

The Laplace mechanism idea can be adapted for a notion of local sensitivity. For $\xi > 0$, the *ξ-smooth sensitivity* of f at $x_{1:n}$ is

$$SS_\xi(f, x_{1:n}) := \sup_{\tilde{x}_{1:n}} \left\{ e^{-\xi d_H(x_{1:n}, \tilde{x}_{1:n})} LS_2(f, \tilde{x}_{1:n}) \right\},$$

where

$$LS_2(f, x_{1:n}) := \sup_{\tilde{x}_{1:n}, d_H(x_{1:n}, \tilde{x}_{1:n})=1} \|f(x_{1:n}) - f(\tilde{x}_{1:n})\|_2$$

Choosing $\xi \leq \frac{\varepsilon}{4(p+2\log(2/\delta))}$, scaling the additive Gaussian noise with the ξ-smooth sensitivity (in Euclidean norm) instead of $GS_2(f)$ guarantees (ε, δ)-DP.

Smooth sensitivity : median

Suppose $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and let $\hat{m}(x_{1:n}) = \text{med}_{1 \leq i \leq n}(x_i)$. Work with truncated data $y_i = \text{sign}(x_i) \max(|x_i|, B_n)$ for $i = 1, \ldots, n$ and some known bound $B_n < \infty$. Note that $GS(\hat{m}) = 2B_n$

Smooth sensitivity : median

Suppose $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and let $\hat{m}(x_{1:n}) = \text{med}_{1 \leq i \leq n}(x_i)$. Work with truncated data $y_i = \text{sign}(x_i) \max(|x_i|, B_n)$ for $i = 1, \ldots, n$ and some known bound $B_n < \infty$. Note that $GS(\hat{m}) = 2B_n$ ...so the standard Laplace mechanism would add constant noise ...consider instead

$$\tilde{m}(y_{1:n}) = \hat{m}(y_{1:n}) + \frac{2}{\varepsilon} SS_{\hat{m}_T}^{(\beta)}(y_{1:n}) Z, x \in \mathbb{R}^n,$$

where $\beta = \frac{\varepsilon}{4\{1+\log(2/\delta)\}}$ and Z is a standard normal random variable.

Smooth sensitivity : median

Suppose $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and let $\hat{m}(x_{1:n}) = \mathrm{med}_{1 \leq i \leq n}(x_i)$. Work with truncated data $y_i = \mathrm{sign}(x_i) \max(|x_i|, B_n)$ for $i = 1, \ldots, n$ and some known bound $B_n < \infty$. Note that $GS(\hat{m}) = 2B_n$ . . .so the standard Laplace mechanism would add constant noise . . .consider instead

$$\tilde{m}(y_{1:n}) = \hat{m}(y_{1:n}) + \frac{2}{\varepsilon} SS^{(\beta)}_{\hat{m}_T}(y_{1:n}) Z, x \in \mathbb{R}^n,$$

where $\beta = \frac{\varepsilon}{4\{1+\log(2/\delta)\}}$ and Z is a standard normal random variable.

Under mild conditions $SS^{(\beta)}_{\hat{m}_T}(y_{1:n})$ is small !

Cond. M Assume X_1 has a unique median m and a density f in $[m-r, m+r]$. Moreover, $\exists r, L$ such that $f(u) \geq L$, for all $u \in [m-r, m+r]$.

Smooth sensitivity : median

Suppose $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ and let $\hat{m}(x_{1:n}) = \mathrm{med}_{1 \leq i \leq n}(x_i)$. Work with truncated data $y_i = \mathrm{sign}(x_i) \max(|x_i|, B_n)$ for $i = 1, \dots, n$ and some known bound $B_n < \infty$. Note that $GS(\hat{m}) = 2B_n$ . . .so the standard Laplace mechanism would add constant noise . . .consider instead

$$\tilde{m}(y_{1:n}) = \hat{m}(y_{1:n}) + \frac{2}{\varepsilon} SS^{(\beta)}_{\hat{m}_T}(y_{1:n}) Z, x \in \mathbb{R}^n,$$

where $\beta = \frac{\varepsilon}{4\{1+\log(2/\delta)\}}$ and Z is a standard normal random variable.

Under mild conditions $SS^{(\beta)}_{\hat{m}_T}(y_{1:n})$ is small !

Cond. M Assume X_1 has a unique median m and a density f in $[m-r, m+r]$. Moreover, $\exists r, L$ such that $f(u) \geq L$, for all $u \in [m-r, m+r]$.

Lemma. Let Cond. M holds. Then, for $\tau \in (0, 1]$. With probability at least $1 - 2\tau - 2e^{-n(q_2-q_1)^2/8}$,

$$SS^{(\beta)}_{\hat{m}}(Y_{1:n}) \leq \frac{2r}{eL\beta(n-1)} \left(\log(\sqrt{n}) + \log(1/\tau)\right) + 2Te^{-\beta\sqrt{n}},$$

where$q_1 = F(m-r)$ and $q_2 = F(m+r)$.

Propose-test-release : median

Idea : *if the data is not in a bad configuration we could add less noise.*

Propose-test-release : median

Idea : *if the data is not in a bad configuration we could add less noise.*

Suppose $x_{1:n} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and let $\hat{m}(x_{1:n}) = \mathrm{med}_{1 \leq i \leq n}(x_i)$.

$$\begin{aligned} \hat{A}_\eta : \mathbb{R}^n &\to \{0, 1, 2, \ldots, n\} \\ x_{1:n} &\mapsto \min\{k \geq 0 : \exists x' \; s.t. \; d_H(x_{1:n}, x'_{1:n}) \leq k, \; |\hat{m}(x'_{1:n}) - \hat{m}(x_{1:n})| > \eta\}. \end{aligned}$$

Propose-test-release : median

Idea : *if the data is not in a bad configuration we could add less noise.*

Suppose $x_{1:n} = (x_1, \dots, x_n) \in \mathbb{R}^n$ and let $\hat{m}(x_{1:n}) = \mathrm{med}_{1 \leq i \leq n}(x_i)$.

$$\hat{A}_\eta : \mathbb{R}^n \to \{0, 1, 2, \dots, n\}$$

$$x_{1:n} \mapsto \min\{k \geq 0 : \exists x' \; s.t. \; d_H(x_{1:n}, x'_{1:n}) \leq k, \; |\hat{m}(x'_{1:n}) - \hat{m}(x_{1:n})| > \eta\}.$$

Let Z_1, Z_2 be i.i.d. standard Laplace and define the randomized functions

$$\tilde{A}_\eta(x_{1:n}) = \hat{A}_\eta(x_{1:n}) + \frac{1}{\varepsilon} Z_1$$

and

$$\tilde{m}_\eta(x_{1:n}) = \begin{cases} \perp & \text{if } \tilde{A}_\eta(x_{1:n}) \leq 1 + \frac{\log(2/\delta)}{\varepsilon} \\ \hat{m}(x_{1:n}) + \frac{\eta}{\varepsilon} Z_2 & \text{otherwise.} \end{cases}$$

Propose-test-release : median

Idea : *if the data is not in a bad configuration we could add less noise.*

Suppose $x_{1:n} = (x_1, \dots, x_n) \in \mathbb{R}^n$ and let $\hat{m}(x_{1:n}) = \text{med}_{1 \leq i \leq n}(x_i)$.

$$\begin{aligned} \hat{A}_\eta : \mathbb{R}^n &\to \{0, 1, 2, \dots, n\} \\ x_{1:n} &\mapsto \min\{k \geq 0 : \exists x' \; s.t. \; d_H(x_{1:n}, x'_{1:n}) \leq k, \; |\hat{m}(x'_{1:n}) - \hat{m}(x_{1:n})| > \eta\}. \end{aligned}$$

Let Z_1, Z_2 be i.i.d. standard Laplace and define the randomized functions

$$\tilde{A}_\eta(x_{1:n}) = \hat{A}_\eta(x_{1:n}) + \frac{1}{\varepsilon} Z_1$$

and

$$\tilde{m}_\eta(x_{1:n}) = \begin{cases} \perp & \text{if } \tilde{A}_\eta(x_{1:n}) \leq 1 + \frac{\log(2/\delta)}{\varepsilon} \\ \hat{m}(x_{1:n}) + \frac{\eta}{\varepsilon} Z_2 & \text{otherwise.} \end{cases}$$

Prop. The estimator $\tilde{m}_\eta(x_{1:n})$ is $(2\varepsilon, \delta)$-DP and can be computed in $O(n \log n)$ time. Furthermore, if Cond. M holds and we choose $\eta \leq \frac{4 C_{\tau,\varepsilon,\delta}}{Ln} + \frac{4 \log(2/\tau)}{3Ln}$, then w. p. at least $1 - \tau$,

$$\tilde{\theta}(x_{1:n}) = \hat{\theta}(x_{1:n}) + \frac{\eta}{\varepsilon} Z_2.$$

Some references

- Seminal paper : Dwork, McSherry, Nissim and Smith (TCC 2006) introduced the definition of DP as well as Laplace and exponential mechanism.
- Smooth sensitivity : introduced in Nissim, Raskhonikova and Smith (STOC 2007). They give a fast linear time algorithm for computing the smooth sensitivity of the median.
- Propose-test-release : Dwork and Lei (STOC 2009) introduced this framework in the context of making robust statistics differentially private.

Some references

- Seminal paper : Dwork, McSherry, Nissim and Smith (TCC 2006) introduced the definition of DP as well as Laplace and exponential mechanism.
- Smooth sensitivity : introduced in Nissim, Raskhonikova and Smith (STOC 2007). They give a fast linear time algorithm for computing the smooth sensitivity of the median.
- Propose-test-release : Dwork and Lei (STOC 2009) introduced this framework in the context of making robust statistics differentially private.
- Some recent work on refinements and variants of Propose-Test-Release includes Avella-Medina and Brunel (2019), Bun and Steinke (NeurIPS 2019), Liu, Kong and Oh (COLT 2022), Wang et. al (NeurIPS 2022).
- Asi and Duchi (NeurIPS 2020) propose another local form of noise calibration they call inverse sensitivity mechanism and has some optimality properties.

Hypothesis testing view of DP mechanisms

Differential privacy as hypothesis testing

Given two neighboring datasets X and X', and a randomized algorithm A test $H_0 : P = A(X) \quad vs \quad H_1 : P = A(X')$.

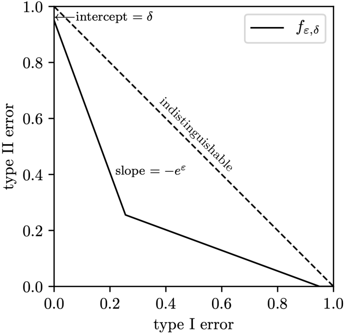
Differential privacy as hypothesis testing

Given two neighboring datasets X and X', and a randomized algorithm A test $H_0 : P = A(X) \quad vs \quad H_1 : P = A(X')$.

Theorem (Wasserman and Zhou (2010, JASA))

A is (ε, δ)-DP iff for all neighboring data sets and $s, t \in \mathcal{X}$, any α-level test for $H_0 : x_i = s \;\; vs \;\; H_1 : x_i = t$ has power function bounded by $\beta(\alpha) \leq 1 - \max\{0, 1 - \delta - e^{\varepsilon}\alpha, e^{-\varepsilon}(1 - \delta - \alpha)\}$.

Figure – Trade-off function plot for (ε, δ)-DP from Dong, Roth and Su (2021, JRSS B).



Gaussian differential privacy

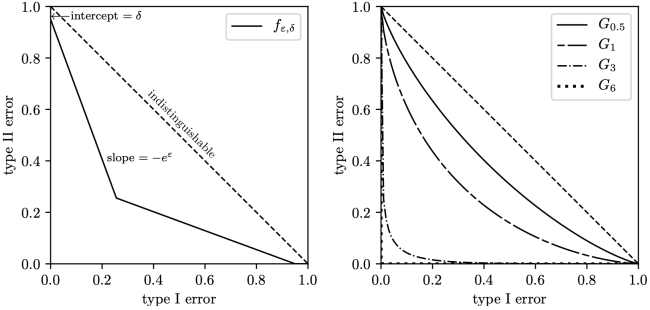
Interpretation : telling whether someone is in the dataset is harder than telling apart $N(0, 1)$ and $N(\mu, 1)$

Gaussian differential privacy

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

Definition (Dong, Roth and Su (2021, JRSS B))

A is μ-GDP iff for all neighboring data sets and $s,t \in \mathcal{X}$, any α-level test for $H_0 : x_i = t$ vs $H_1 : x_i = s$ has power $\beta(\alpha) \leq 1 - \Phi(\Phi^{-1}(1-\alpha) - \mu)$.



Gaussian differential privacy

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

Gaussian differential privacy

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

Definition (Dong, Roth and Su (2021, JRSS B))

A is μ-GDP iff for all neighboring data sets and $s,t \in \mathcal{X}$, any α-level test for $H_0 : x_i = t$ vs $H_1 : x_i = s$ has power $\beta(\alpha) \leq 1 - \Phi(\Phi^{-1}(1-\alpha) - \mu)$.

- A mechanism is μ-GDP if and only if it is $(\varepsilon, \delta(\varepsilon))$-DP for all $\varepsilon \geq 0$, where
$$\delta(\varepsilon) = \Phi(-\frac{\varepsilon}{\mu} + \frac{\mu}{2}) - e^{\varepsilon}\Phi(-\frac{\varepsilon}{\mu} - \frac{\mu}{2})$$

Neyman-Pearson and Gaussian mechanism revisited

Let $m : \mathcal{X}^n \mapsto \mathbb{R}$ has a finite global sensitivity $GS(m)$. Set $\sigma = \frac{1}{\mu}\mathrm{GS}(m)$ and consider the Gaussian mechanism

$$\tilde{m}(x_{1:n}) = m(x_{1:n}) + N(0, \sigma^2).$$

Neyman-Pearson and Gaussian mechanism revisited

Let $m : \mathcal{X}^n \mapsto \mathbb{R}$ has a finite global sensitivity $GS(m)$. Set $\sigma = \frac{1}{\mu}\mathrm{GS}(m)$ and consider the Gaussian mechanism

$$\tilde{m}(x_{1:n}) = m(x_{1:n}) + N(0, \sigma^2).$$

- If $X \sim N(\mu', 1)$, Neyman-Pearson Lemma shows that Likelihood Ratio Test is the most powerful for testing $H_0 : \mu' = 0$ versus $H_1 : \mu' = \mu$. It has power $\beta(\alpha) \leq 1 - \Phi(\Phi^{-1}(1 - \alpha) - \mu)$

Neyman-Pearson and Gaussian mechanism revisited

Let $m : \mathcal{X}^n \mapsto \mathbb{R}$ has a finite global sensitivity $GS(m)$. Set $\sigma = \frac{1}{\mu}\mathrm{GS}(m)$ and consider the Gaussian mechanism

$$\tilde{m}(x_{1:n}) = m(x_{1:n}) + N(0, \sigma^2).$$

- If $X \sim N(\mu', 1)$, Neyman-Pearson Lemma shows that Likelihood Ratio Test is the most powerful for testing $H_0 : \mu' = 0$ versus $H_1 : \mu' = \mu$. It has power $\beta(\alpha) \leq 1 - \Phi(\Phi^{-1}(1-\alpha) - \mu)$
- Let $x'_{1:n}$ be such that $\sum_{i=1}^{n} \mathbb{1}\{x_i \neq x'_i\} = 1$. Given $x_{1:n}$ and $x'_{1:n}$,

$$\tilde{m}(x_{1:n}) \sim P = N(m(x_{1:n}), \sigma^2) \quad \text{and} \quad \tilde{m}(x'_{1:n}) \sim Q = N(m(x'_{1:n}), \sigma^2).$$

Neyman-Pearson and Gaussian mechanism revisited

Let $m : \mathcal{X}^n \mapsto \mathbb{R}$ has a finite global sensitivity $GS(m)$. Set $\sigma = \frac{1}{\mu}\mathrm{GS}(m)$ and consider the Gaussian mechanism

$$\tilde{m}(x_{1:n}) = m(x_{1:n}) + N(0, \sigma^2).$$

- If $X \sim N(\mu', 1)$, Neyman-Pearson Lemma shows that Likelihood Ratio Test is the most powerful for testing $H_0 : \mu' = 0$ versus $H_1 : \mu' = \mu$. It has power $\beta(\alpha) \leq 1 - \Phi(\Phi^{-1}(1-\alpha) - \mu)$
- Let $x'_{1:n}$ be such that $\sum_{i=1}^n \mathbb{1}\{x_i \neq x'_i\} = 1$. Given $x_{1:n}$ and $x'_{1:n}$,
$$\tilde{m}(x_{1:n}) \sim P = N(m(x_{1:n}), \sigma^2) \quad \text{and} \quad \tilde{m}(x'_{1:n}) \sim Q = N(m(x'_{1:n}), \sigma^2).$$
- Testing $P = Q$ is equivalent to $H_0 : \mu = m(x_{1:n})$ versus $H_1 : \mu = m(x'_{1:n})$. For $GS(m) \geq m(x'_{1:n}) - m(x_{1:n}) \geq 0$, the LRT has power
$$\begin{aligned}\beta(\alpha) &= 1 - \Phi(z_{1-\alpha} - \frac{m(x'_{1:n}) - m(x_{1:n})}{\sigma}) \\ &= 1 - \Phi(z_{1-\alpha} - \frac{m(x_{1:n})' - m(x_{1:n})}{GS(m)/\mu}) \leq 1 - \Phi(z_{1-\alpha} - \mu).\end{aligned}$$

Neyman-Pearson and Gaussian mechanism revisited

Indeed, consider $H_0 : X \sim N(\theta, \sigma^2)$ versus $H_1 : X \sim N(\theta', \sigma^2)$. When $\theta' - \theta \geq 0$, the likelihood ratio of $N(\theta, \sigma^2)$ and $N(\theta', \sigma^2)$ is

$$\frac{\varphi(\frac{x-\theta'}{\sigma})/\sigma}{\varphi(\frac{x-\theta}{\sigma})/\sigma} = \frac{e^{-\frac{(x-\theta')^2}{2\sigma^2}}}{e^{-\frac{(x-\theta)^2}{2\sigma^2}}} = e^{\frac{\theta'-\theta}{\sigma}x - \frac{1}{2\sigma^2}(\theta'^2-\theta^2)},$$

a monotone increasing function in x. So the LRT rejects if $X > t$ and reject otherwise.

Neyman-Pearson and Gaussian mechanism revisited

Indeed, consider $H_0 : X \sim N(\theta, \sigma^2)$ versus $H_1 : X \sim N(\theta', \sigma^2)$. When $\theta' - \theta \geq 0$, the likelihood ratio of $N(\theta, \sigma^2)$ and $N(\theta', \sigma^2)$ is

$$\frac{\varphi(\frac{x-\theta'}{\sigma})/\sigma}{\varphi(\frac{x-\theta}{\sigma})/\sigma} = \frac{e^{-\frac{(x-\theta')^2}{2\sigma^2}}}{e^{-\frac{(x-\theta)^2}{2\sigma^2}}} = e^{\frac{\theta'-\theta}{\sigma}x - \frac{1}{2\sigma^2}(\theta'^2 - \theta^2)},$$

a monotone increasing function in x. So the LRT rejects if $X > t$ and reject otherwise. The corresponding type I and type II errors are

$$\alpha(t) = \mathbb{P}_{H_0}(X > t) = 1 - \Phi(\frac{t-\theta}{\sigma}), \quad \beta(t) = \mathbb{P}_{H_1}(X \leq t) = \Phi(\frac{t-\theta'}{\sigma}).$$

Solving α for t gives $t = \sigma\Phi^{-1}(1-\alpha) + \theta$.

Neyman-Pearson and Gaussian mechanism revisited

Indeed, consider $H_0 : X \sim N(\theta, \sigma^2)$ versus $H_1 : X \sim N(\theta', \sigma^2)$. When $\theta' - \theta \geq 0$, the likelihood ratio of $N(\theta, \sigma^2)$ and $N(\theta', \sigma^2)$ is

$$\frac{\varphi(\frac{x-\theta'}{\sigma})/\sigma}{\varphi(\frac{x-\theta}{\sigma})/\sigma} = \frac{e^{-\frac{(x-\theta')^2}{2\sigma^2}}}{e^{-\frac{(x-\theta)^2}{2\sigma^2}}} = e^{\frac{\theta'-\theta}{\sigma}x - \frac{1}{2\sigma^2}(\theta'^2-\theta^2)},$$

a monotone increasing function in x. So the LRT rejects if $X > t$ and reject otherwise. The corresponding type I and type II errors are

$$\alpha(t) = \mathbb{P}_{H_0}(X > t) = 1 - \Phi(\frac{t-\theta}{\sigma}), \quad \beta(t) = \mathbb{P}_{H_1}(X \leq t) = \Phi(\frac{t-\theta'}{\sigma}).$$

Solving α for t gives $t = \sigma\Phi^{-1}(1-\alpha) + \theta$. Therefore

$$\beta(\alpha) = \Phi(z_{1-\alpha} - \frac{(\theta'-\theta)}{\sigma})$$

Composition revisited

- Exact composition for Gaussian DP : suppose you K outputs, each of them is μ_k-GDP for $k = 1, \ldots, K$. The combined privacy guarantee is $\sqrt{\sum_{k=1} \mu_k^2}$-GDP.

Composition revisited

- Exact composition for Gaussian DP : suppose you K outputs, each of them is μ_k-GDP for $k = 1, \ldots, K$. The combined privacy guarantee is $\sqrt{\sum_{k=1} \mu_k^2}$-GDP.
- Compositions comes from

$$H_0 : P \sim \prod_{k=1}^{K} N(0,1) \quad \text{versus} \quad H_1 : P \sim \prod_{k=1}^{K} N(\mu_k, 1)$$

Composition revisited

- Exact composition for Gaussian DP : suppose you K outputs, each of them is μ_k-GDP for $k = 1, \ldots, K$. The combined privacy guarantee is $\sqrt{\sum_{k=1} \mu_k^2}$-GDP.
- Compositions comes from

$$H_0 : P \sim \prod_{k=1}^{K} N(0,1) \quad \text{versus} \quad H_1 : P \sim \prod_{k=1}^{K} N(\mu_k, 1)$$

- The analogous statement we saw was that K-fold composition of $(\varepsilon_k, \delta_k)$-DP outputs is $(\sum_{k=1}^{K} \varepsilon_k, \sum_{k=1}^{K} \delta_k)$-DP. This is in fact a loose bound...

Composition revisited

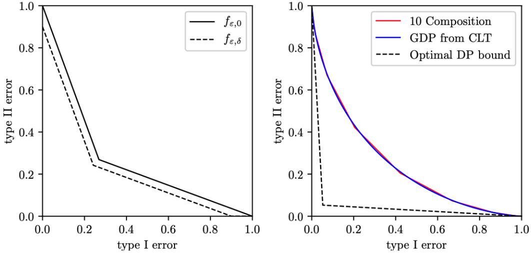
- Exact composition for Gaussian DP : suppose you K outputs, each of them is μ_k-GDP for $k = 1, \dots, K$. The combined privacy guarantee is $\sqrt{\sum_{k=1} \mu_k^2}$-GDP.
- Compositions comes from

$$H_0 : P \sim \prod_{k=1}^{K} N(0,1) \quad \text{versus} \quad H_1 : P \sim \prod_{k=1}^{K} N(\mu_k, 1)$$

- The analogous statement we saw was that K-fold composition of $(\varepsilon_k, \delta_k)$-DP outputs is $(\sum_{k=1}^{K} \varepsilon_k, \sum_{k=1}^{K} \delta_k)$-DP. This is in fact a loose bound...
- Advanced composition : to ensure $(\varepsilon', K\delta + \delta')$ after K-fold composition of (ε, δ)-DP mechanisms, it suffices to take $\varepsilon = \frac{\varepsilon'}{2\sqrt{2K \ln(1/\delta')}}$

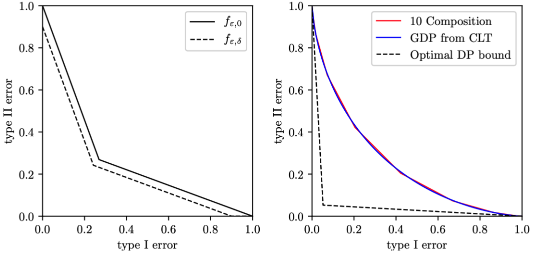
GDP and CLT for composition

The plot below shows the 1-GDP approximation to 10-fold composition of $(1/\sqrt{10}, 0)$-DP mechanisms :



GDP and CLT for composition

The plot below shows the 1-GDP approximation to 10-fold composition of $(1/\sqrt{10}, 0)$-DP mechanisms :



Thm. Fix $\mu > 0$ and let $\varepsilon = \mu/\sqrt{K}$. There is a constant $c > 0$ that only depends on μ satisfying

$$\Phi(z_{1-\alpha-\frac{c}{K}} - \mu) - \frac{c}{K} \leq f_{\varepsilon,0}^{\otimes K}(\alpha) \leq \Phi(z_{1-\alpha-\frac{c}{K}} - \mu) + \frac{c}{K},$$

for all $K \geq 1$ and $c/K \leq \alpha \leq 1 - c/K$.

GDP and CLT for composition

Dong, Roth and Su (2021, JRSS B)

Idea : the composition of many mechanisms that can be interpreted as tests of hypothesis can be approximated by μ-GDP.

GDP and CLT for composition

Dong, Roth and Su (2021, JRSS B)

Idea : the composition of many mechanisms that can be interpreted as tests of hypothesis can be approximated by μ-GDP.

Def. Let $h : \mathcal{X}^n \to \mathbb{R}^p$ be a randomized function. We say that h is f-DP if any α-level test between simple hypotheses of the form $H_0 : x_i = t$ vs. $H_1 : x_i = s$ has power function $\beta(\alpha) \leq 1 - f(\alpha)$, where f is a convex, continuous, non-increasing function satisfying $f(\alpha) \leq 1 - \alpha$ for all $\alpha \in [0, 1]$.

Thm. Consider the K-fold composition of f_k-DP mechanisms for $k = 1, \dots, K$. Let $\kappa_1(f_k) = -\int_0^1 \log |f_k'(x)| dx$ and $\kappa_2(f_k) = -\int_0^1 \log^2 |f_k'(x)| dx$ and suppose $\sum_{k=1}^K \kappa_1(f_k) \to \mu$, $\max_{1 \leq k \leq K} \kappa_1(f_k) \to 0$ and $\sum_{k=1}^K \kappa_2(f_k) \to s^2$. Then,

$$\lim_{K \to \infty} f_1 \otimes f_2 \otimes \cdots \otimes f_K(\alpha) = \Phi(z_{1-\alpha} - 2\mu/s) = G_{2\mu/s}(\alpha),$$

uniformly for all $\alpha \in [0, 1]$.

Some connections to robust statistics

Using robust statistics ideas in differential privacy

- Influence Function : calibrating privacy inducing noise via the IF in Avella-Medina (2021, JASA), builds particularly on ideas from Chaudhuri and Hsu (2012, ICML)
- Finite sample breakdown point : non-asymptotic deviations analysis for location estimators explored in Avella-Medina & Brunel (2020, ArXiv), motivated by approach in Dwork and Lei (STOC, 2009)
- Convex optimization : noisy gradient descent and noisy Newton algorithms in Avella-Medina, Bradshaw and Loh (2023, Ann. Statist.). Builds on large literature on noisy optimization and work by many people (Bassily, Chaudhuri, Duchi, Feldman, Jain, Smith, Talwar, Thakurta . . .)

Ingredient I

M-estimators

An M-estimator (Huber, 1964) is an estimate $\hat{\theta} = T(F_n)$ defined by

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

or by an implicit equation as

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

Ingredient II

Robust statistics tools

- The IF of a functional $T(F)$ is a special (Gâteaux) derivative given by

  $$IF(z; F, T) = \lim_{\epsilon \to 0} \frac{T(F_\epsilon) - T(F)}{\epsilon},$$

  where $F_\epsilon = (1-\epsilon)F + \epsilon\Delta_z$ and Δ_z is a mass point.
- It can be interpreted as limit of the sensitivity curve of the statistic $T_n = T(F_n)$

  $$SC(z; z_1, \ldots, z_{n-1}) = n(T_n(z_1, \ldots, z_{n-1}, z) - T_{n-1}(z_1, \ldots, z_{n-1}))$$

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

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

Private M-estimation

Avella-Medina (JASA, 2021)

For an M-estimator

$$T(F_n) = \hat{\theta} = \mathrm{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(z_i, \theta)$$

defined through a bounded function Ψ, one can simply return

$$A_T(F_n) = T(F_n) + \gamma(T, F_n) \frac{5\sqrt{2 \log(n) \log(1/\delta)}}{\varepsilon n} N_p(0, I)$$

where $\gamma(T, F_n) = \sup_x \|\mathrm{IF}(x; T, F_n)\|$.

Private M-estimation

Avella-Medina (JASA, 2021)

For an M-estimator

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

defined through a bounded function Ψ, one can simply return

$$A_T(F_n) = T(F_n) + \gamma(T, F_n) \frac{5\sqrt{2 \log(n) \log(1/\delta)}}{\varepsilon n} N_p(0, I)$$

where $\gamma(T, F_n) = \sup_x \|\text{IF}(x; T, F_n)\|$.

Theorem A_T is (ε, δ)-differentially private

Private M-estimation

Avella-Medina (JASA, 2021)

For an M-estimator

$$T(F_n) = \hat{\theta} = \mathrm{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(z_i, \theta)$$

defined through a bounded function Ψ, one can simply return

$$A_T(F_n) = T(F_n) + \gamma(T, F_n) \frac{5\sqrt{2\log(n)\log(1/\delta)}}{\varepsilon n} N_p(0, I)$$

where $\gamma(T, F_n) = \sup_x \|\mathrm{IF}(x; T, F_n)\|$.

Theorem A_T is (ε, δ)-differentially private . . . for large n.

Proof idea

Noise calibration with smooth sensitivity

For $f : \mathcal{X}^n \to R^p$ the global sensitivity of f is

$$GS_2(f) := \sup_{x_{1:n}, \tilde{x}_{1:n}, d_H(x_{1:n}, \tilde{x}_{1:n}) = 1} \| f(x_{1:n}) - f(\tilde{x}_{1:n}) \|_2 .$$

The Laplace mechanism idea can be adapted for a notion of local sensitivity. For $\xi > 0$, the *ξ-smooth sensitivity* of f at $x_{1:n}$ is

$$SS_\xi(f, x_{1:n}) := \sup_{\tilde{x}_{1:n}} \left\{ e^{-\xi d_H(x_{1:n}, \tilde{x}_{1:n})} LS_2(f, \tilde{x}_{1:n}) \right\},$$

where

$$LS_2(f, x_{1:n}) := \sup_{\tilde{x}_{1:n}, d_H(x_{1:n}, \tilde{x}_{1:n}) = 1} \| f(x_{1:n}) - f(\tilde{x}_{1:n}) \|_2$$

Choosing $\xi \leq \frac{\varepsilon}{4(p+2\log(2/\delta))}$, scaling the additive Gaussian noise with the ξ-smooth sensitivity (in Euclidean norm) instead of $GS_2(f)$ guarantees (ε, δ)-DP.

Proof idea

Our key insight

Let $C_0 > 0$, $\|\Psi\| \leq K_n$ and $\|\dot{\Psi}\| \leq L_n$.

Lemma 1

$$SS_\xi(T, D(F_n)) \leq \max\left\{\frac{2\Gamma_n}{n}, C_0 K_n \exp\left(-\xi\sqrt{\frac{n\log(1/\delta)}{p}} + \xi\right)\right\}.$$

Lemma 2

$$\Gamma_n \leq 2\gamma(T, F_n)\left(1 + C_0\sqrt{\frac{p\log(1/\delta)}{n}}\left(C_1 + L_n + C_2\gamma(T, F_n)\right)\right).$$

Corollary. $SS_\xi(T, D(F_n)) \lesssim \frac{1}{n}\sup_x \|IF(x; T, F_n)\| = \frac{1}{n}\gamma(T, F_n)$

Remarks

A few technical points

- The smooth sensitivity is hard to compute !
- The above lemmas used

  $$\Gamma_n := \sup \left\{ \gamma_{1/n}(T, G) : d_\infty(F_n, G) \leq C\sqrt{\frac{p \log(1/\delta)}{n}} \right\}$$

  and hence a fixed scale versions of the influence function, i.e. for a fixed $\rho > 0$

  $$IF_\epsilon(x; T(F)) := \frac{T((1-\epsilon)F + \epsilon\Delta_x) - T(F)}{\epsilon}$$

  and

  $$\gamma_\epsilon(T, F) := \sup_x \|IF_\epsilon(x; T, F)\|$$

Remarks

A few technical points

- Our parameter estimates attain near minimax rates of convergence under DP according to Cai, Wang and Zhang (2021, Ann. Statist.)

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

- Any (ε,δ)-DP estimator has to be robust as there is a lower bound depending on the gross-error sensitivity [Chaudhuri and Hsu (2012, ICML) ; Avella-Medina (2021, JASA)]...

$$\mathbb{E}_{F_n}\mathbb{E}_A\Big[\|A(F_n))-T(F)\|\Big] \gtrsim \frac{1}{\varepsilon}\gamma(T,F)$$

Remarks

A few technical points

- Our parameter estimates attain near minimax rates of convergence under DP according to Cai, Wang and Zhang (2021, Ann. Statist.)

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

- Any (ε,δ)-DP estimator has to be robust as there is a lower bound depending on the gross-error sensitivity [Chaudhuri and Hsu (2012, ICML) ; Avella-Medina (2021, JASA)]...

$$\mathbb{E}_{F_n}\mathbb{E}_A\Big[\|A(F_n)) - T(F)\|\Big] \gtrsim \frac{1}{\varepsilon}\gamma(T,F)$$

- Our DP estimator is asymptotically normally distributed but inference is not immediate as we are not releasing variance estimates at this point !

Related work

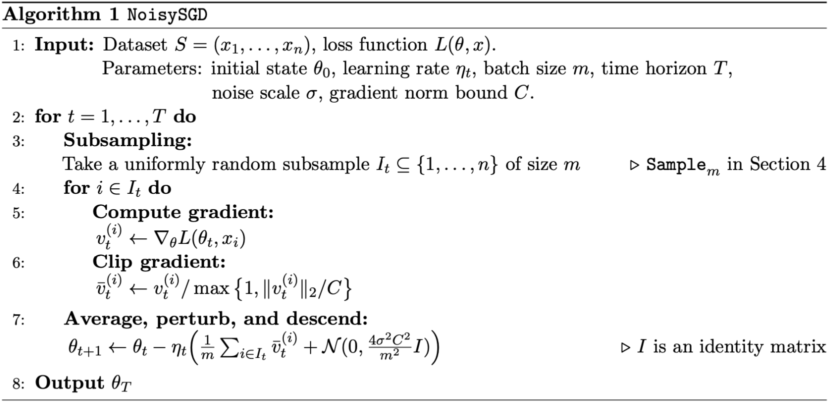
- Large CS/ML literature on differentially private estimation following the seminal work of Dwork et al. (2006, TCC)
- Previous work related to M-estimation and robust statistics : Dwork and Lei (2007, STOC), Lei (2011, NeurIPS), Smith (2011, STOC), Chaudhuri and Hsu (2012, ICML)
- Statistical minimax rates : Wasserman and Zhou (2010, JASA), Duchi et al. (2018, JASA), Cai et al. (2021, Ann. Statist.)
- Not much on inference on 2018 when I had a first draft on this : Sheffet (2017, ICML), Barrientos et al. (2019, JCGS), Awan and Slavkovic (2018, NeurIPS), Canone et al. (2019, STOC)

Private inference via noisy optimization

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

Private Stochastic Gradient Descent Algorithm

The canonical choice in practice



Algorithm 1 `NoisySGD`

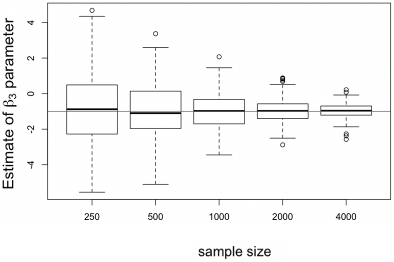
```
Input: Dataset $S = (x_1, \dots, x_n)$, loss function $L(\theta, x)$.
       Parameters: initial state $\theta_0$, learning rate $\eta_t$, batch size $m$, time horizon $T$,
                   noise scale $\sigma$, gradient norm bound $C$.
for $t = 1, \dots, T$ do
    Subsampling:
    Take a uniformly random subsample $I_t \subseteq \{1, \dots, n\}$ of size $m$        ▷ $\mathtt{Sample}_m$ in Section 4
    for $i \in I_t$ do
        Compute gradient:
        $v_t^{(i)} \leftarrow \nabla_\theta L(\theta_t, x_i)$
        Clip gradient:
        $\bar{v}_t^{(i)} \leftarrow v_t^{(i)} / \max\left\{1, \|v_t^{(i)}\|_2 / C\right\}$
    Average, perturb, and descend:
    $\theta_{t+1} \leftarrow \theta_t - \eta_t \left( \frac{1}{m} \sum_{i \in I_t} \bar{v}_t^{(i)} + \mathcal{N}(0, \frac{4\sigma^2 C^2}{m^2} I) \right)$        ▷ $I$ is an identity matrix
Output $\theta_T$
```

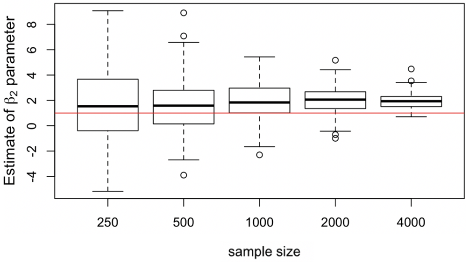
A remark on clipping

- Clipped likelihood as M-estimator

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

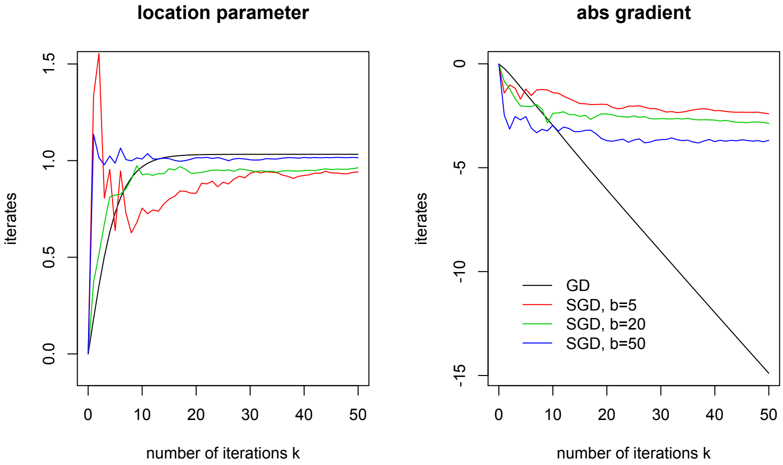
where $h_c(z) = z \min\{1, \frac{c}{\|z\|_2}\}$ is the multivariate Huber function.


Linear Regression


Logistic Regression

GD V. SGD

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ and consider the Huber estimator.



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...
- Self-concordance and statistics : Bach (2010), Karimireddy et al. (2019), Sun and Tran-Dinh (2020), Ostrovskii and Bach (2021)
- 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)...

M-estimators

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

$$\hat{\theta} = \text{argmin}_{\theta \in \mathbb{R}^p} \mathcal{L}_n(\theta) = \text{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(z_i, \theta),$$

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 :

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

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

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

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

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

Conditions for convergence analysis of NGD

Cond. 1 The gradient of the loss function is such that

$$\sup_{x\in\mathcal{X},\theta\in\Theta} \|\Psi(x,\theta)\|_2 \leq B < \infty.$$

Cond. 2 The loss $\mathcal{L}_n$ is locally τ_1-strongly convex and τ_2-smooth, i.e.,

$$\mathcal{L}_n(\theta_1) - \mathcal{L}_n(\theta_2) \geq \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \tau_1 \|\theta_1 - \theta_2\|_2^2, \quad \forall \theta_1, \theta_2 \in \mathcal{B}_r(\theta^{(0)}),$$

and

$$\mathcal{L}_n(\theta_1) - \mathcal{L}_n(\theta_2) \leq \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \tau_2 \|\theta_1 - \theta_2\|_2^2, \quad \forall \theta_1, \theta_2 \in \Theta \subseteq \mathbb{R}^p.$$

Conditions for convergence analysis of NGD

Cond. 1 The gradient of the loss function is such that

$$\sup_{x \in \mathcal{X}, \theta \in \Theta} \|\Psi(x, \theta)\|_2 \leq B < \infty.$$

Cond. 2 The loss $\mathcal{L}_n$ is locally τ_1-strongly convex and τ_2-smooth, i.e.,

$$\mathcal{L}_n(\theta_1) - \mathcal{L}_n(\theta_2) \geq \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \tau_1 \|\theta_1 - \theta_2\|_2^2, \quad \forall \theta_1, \theta_2 \in \mathcal{B}_r(\theta^{(0)}),$$

and

$$\mathcal{L}_n(\theta_1) - \mathcal{L}_n(\theta_2) \leq \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \tau_2 \|\theta_1 - \theta_2\|_2^2, \quad \forall \theta_1, \theta_2 \in \Theta \subseteq \mathbb{R}^p.$$

Key Lemma. With probability at least $1 - \xi$,

$$\|Z_k\|_2 \leq \{4\sqrt{p} + 2\sqrt{2\log(K/\xi)}\},$$

for all $k < K$.

Convergence analysis of NGD I

For simplicity, assume τ_1-strict convexity (not just local) and τ_2-smoothness. Consider

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla \mathcal{L}_n(\theta^{(k)}) + N_k,$$

where $\|N_k\| \leq r_n$ for $k = 1, \ldots, K$ (holds with high probability by Key Lemma). By Cauchy-Schwarz

$$\begin{aligned}
&\|\theta^{(t+1)} - \hat{\theta}\|_2^2 = \|\theta^{(k)} - \eta \nabla \mathcal{L}_n(\theta^{(k)}) - \hat{\theta} + N_k\|_2^2 \\
&\leq \|\theta^{(k)} - \eta \nabla \mathcal{L}_n(\theta^{(k)}) - \hat{\theta}\|_2^2 + (\|\theta^{(k)} - \hat{\theta}\|_2 + 2\eta \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2) r_n + r_n^2.
\end{aligned}$$

Leveraging the inequality $xy \leq \frac{1}{2\alpha}x^2 + \frac{\alpha}{2}y^2$ for arbitrary $\alpha > 0$, we can take arbitrary $\alpha > 0$ and $\beta > 0$ to further upper bound the second term on the right hand side of the last display to obtain

$$\begin{aligned}
\|\theta^{(t+1)} - \hat{\theta}\|_2^2 \leq \|\theta^{(k)} - \eta \nabla \mathcal{L}_n(\theta^{(k)}) - \hat{\theta}\|_2^2 + \alpha \|\theta^{(k)} - \hat{\theta}\|_2^2 \\
+ 2\eta\beta \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + \left(\frac{1}{\alpha} + \frac{2\eta}{\beta} + 1\right) r_n^2.
\end{aligned}$$

Convergence analysis of NGD II

Note that τ_2-smoothness implies $\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 \le 4\tau_2(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}))$. That and τ_1-strong convexity gives

$$\begin{aligned}
&\|\theta^{(k)} - \eta\nabla\mathcal{L}_n(\theta^{(k)}) - \hat{\theta}\|_2^2 \\
&= \|\theta^{(k)} - \hat{\theta}\|_2^2 - 2\eta\langle\nabla\mathcal{L}_n(\theta^{(k)}), \theta^{(k)} - \hat{\theta}\rangle + \eta^2\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 \\
&\le (1 - 2\eta\tau_1)\|\theta^{(k)} - \hat{\theta}\|_2^2 - 2\eta(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})) + \eta^2\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 \\
&\le (1 - 2\eta\tau_1)\|\theta^{(k)} - \hat{\theta}\|_2^2 - 2\eta(1 - 2\eta\tau_2)(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}))
\end{aligned}$$

Note that τ_2-smoothness implies $\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 \le 4\tau_2(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}))$, we obtain

$$\begin{aligned}
\|\theta^{(t+1)} - \hat{\theta}\|_2^2 &\le (1 - 2\eta\tau_1 + \alpha)\|\theta^{(k)} - \hat{\theta}\|_2^2 + \left(\frac{1}{\alpha} + \frac{2\eta}{\beta} + 1\right) r_n^2 \\
&\qquad - 2\eta(1 - 2\eta\tau_2 + 4\beta\tau_2)(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})) \\
&\le (1 - 2\eta\tau_1 + \alpha)\|\theta^{(k)} - \hat{\theta}\|_2^2 + \left(\frac{1}{\alpha} + \frac{2\eta}{\beta} + 1\right) r_n^2.
\end{aligned}$$

The last inequality follows from the optimality of $\hat{\theta}$ and $\eta \le \frac{1}{2\tau_2}$.

Convergence analysis of NGD III

Therefore taking $\alpha = \eta\tau_1$ and $\beta = 2\eta$ we get

$$
\begin{aligned}
\|\theta^{(k+1)} - \hat{\theta}\|_2^2 &\leq (1 - \eta\tau_1)\|\theta^{(k)} - \hat{\theta}\|_2^2 + \left(\frac{1}{\eta\tau_1} + 2\right) r_n^2 \\
&\leq (1 - \eta\tau_1)^{k+1}\|\theta^{(0)} - \hat{\theta}\|_2^2 + \frac{1}{\eta\tau_1}\left(\frac{1}{\eta\tau_1} + 2\right) r_n^2 \\
&\leq \frac{2}{\eta\tau_1}\left(\frac{1}{\eta\tau_1} + 2\right) r_n^2,
\end{aligned}
$$

where the last inequality holds as long as

$$
k \geq 1 + \frac{\log(1/\|\theta^{(0)} - \hat{\theta}\|_2^2) + \log\left(\frac{1}{\eta\tau_1}(\frac{1}{\eta\tau_1} + 2) r_n^2\right)}{\log(1 - \eta\tau_1)}.
$$

Conclusion : We need $k \geq C \log(r_n^2/\|\theta^{(0)} - \hat{\theta}\|_2^2)$ for some $C > 0$ to get

$$
\|\theta^{(k+1)} - \hat{\theta}\|_2^2 \leq O(r_n^2).
$$

Remark

Optimal minimax rates of convergence : under (ε, δ)-DP the optimal rates of convergence are according to Cai, Wang and Zhang (2021, AoS)

$$\inf_{A \in \mathcal{A}_{\varepsilon,\delta}} \sup_{P \in \mathcal{P}(\sigma,p)} \mathbb{E}\|A(F_n) - \theta^{(0)}\| \gtrsim \sigma \left(\sqrt{\frac{p}{n}} + \frac{p\sqrt{\log(1/\delta)}}{n\varepsilon} \right)$$

Example : linear regression

- Consider a linear regression model

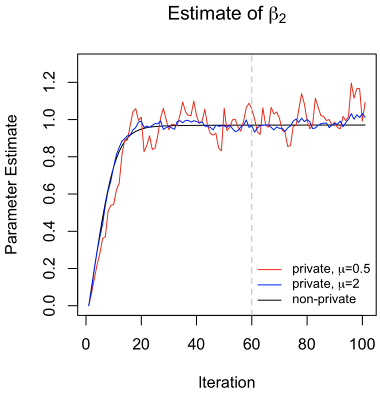
$$y_i = x_i^T \beta + u_i \text{ for } i = 1, \dots, n$$
$$x_i \in \mathbb{R}^p$$
$$u_i \sim N(0, \sigma^2)$$

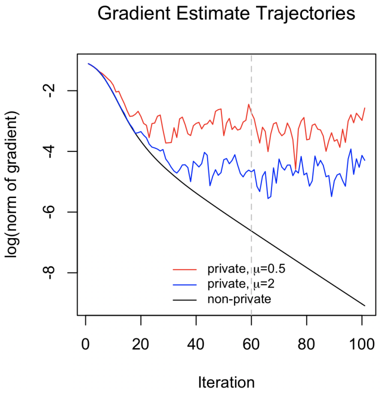
- We want to solve

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

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

Example : linear regression


Estimate of β_2


Gradient Estimate Trajectories

Optimization : gradient descent and Newton's method

- Gradient descent iterations :

$$\theta^{(k+1)} = \theta^{(k)} - \eta \frac{1}{n} \sum_{i=1}^{n} \Psi(x_i, \theta^{(k)})$$

- Newton iterations :

$$\theta^{(k+1)} = \theta^{(k)} - \left(\sum_{i=1}^{n} \dot{\Psi}(x_i, \theta^{(k)}) \right)^{-1} \sum_{i=1}^{n} \Psi(x_i, \theta^{(k)})$$

Optimization : gradient descent and Newton's method

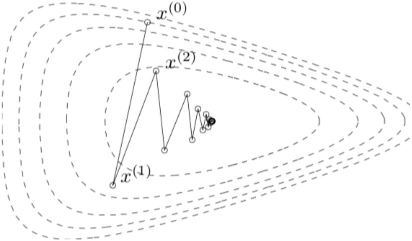


Figure – Gradient descent iterates

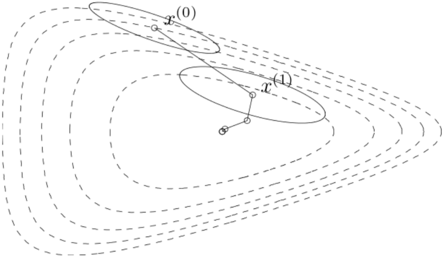
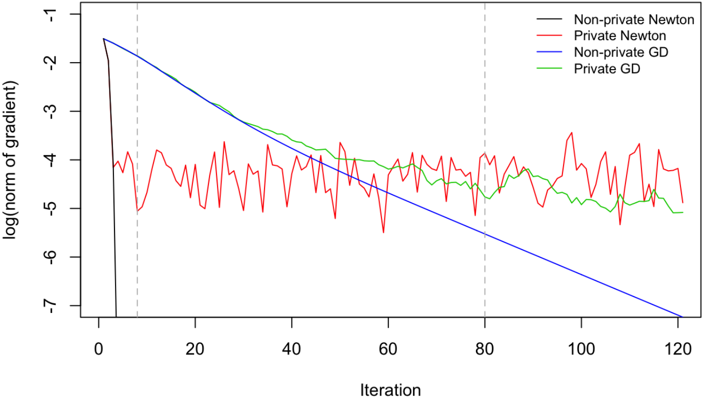


Figure – Newton's iterates

Noisy optimization : private iterations



Noisy Newton

- Noisy Newton :

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

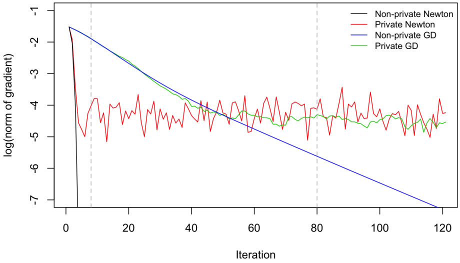
  where $\{N_k\}$ and $\{W_k\}$ are i.i.d. sequences of vectors and symmetric matrices with i.i.d. standard normal components.

- Condition. Hessian of the form

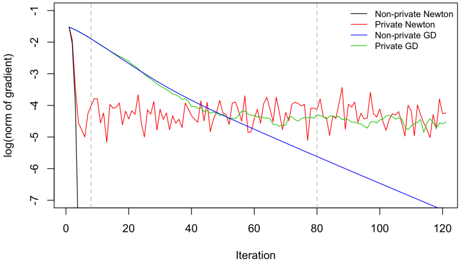
$$\nabla^2\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} a(x_i,\theta)a(x_i,\theta)^\top,$$

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

Noisy Newton theory



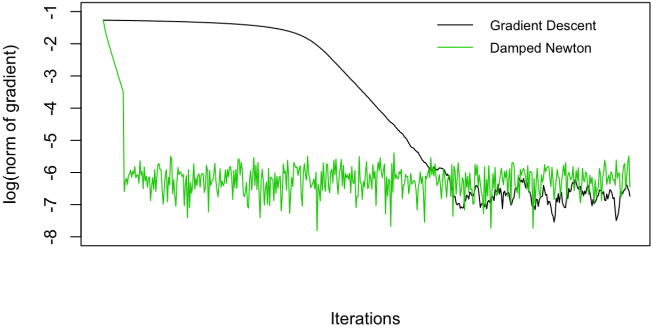
Noisy Newton theory



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

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.

Self-concordance

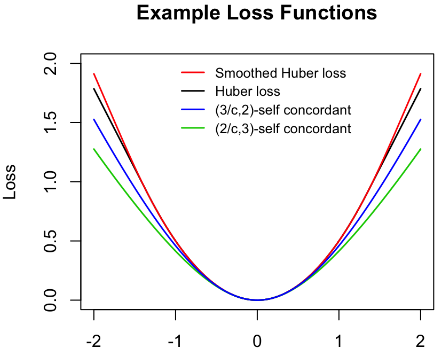
A univariate function $f : \mathbb{R} \to \mathbb{R}$ is (γ, ν)-*self-concordant* if

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

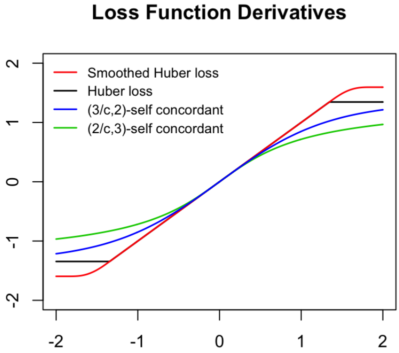
for all x. A multivariate function $f : \mathbb{R}^p \to \mathbb{R}$ is (γ, ν)-self-concordant if

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

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



Example Loss Functions



Loss Function Derivatives

Asymptotic variance

Let's go back to our robust regression example

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

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

Asymptotic variance

Let's go back to our robust regression example

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

where $w(x_i) = \min\left(1, \frac{1}{\|x_i\|_2^2}\right)$ and κ is a Fisher consistency constant. The formulas needed for estimating the variance of $\hat{\beta}$ are :

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi_c^2 \Big(\frac{y_i - x_i^T \theta}{\sigma} \Big) w(x_i)^2 x_i x_i^\top = \frac{1}{n} \sum_{i=1}^{n} z_i z_i^\top$$

$$M_n(\theta) = \frac{1}{n\sigma} \sum_{i=1}^{n} \dot{\psi}_c \Big(\frac{y_i - x_i^\top \theta}{\sigma} \Big) w(x_i) x_i x_i^\top = \frac{1}{n} \sum_{i=1}^{n} \tilde{z}_i \tilde{z}_i^\top$$

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

Private sandwich formula

1. Plug-in estimators $M_n(\theta^{(K)})$ and $Q_n(\theta^{(K)})$, where $\theta^{(K)} = (\beta^{(K)}, \sigma^{(K)})$ are not yet private.

Private sandwich formula

1. Plug-in estimators $M_n(\theta^{(K)})$ and $Q_n(\theta^{(K)})$, where $\theta^{(K)} = (\beta^{(K)}, \sigma^{(K)})$ are not yet private.

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

Private sandwich formula

1. Plug-in estimators $M_n(\theta^{(K)})$ and $Q_n(\theta^{(K)})$, where $\theta^{(K)} = (\beta^{(K)}, \sigma^{(K)})$ are not yet private.

2. Matrix Gaussian mechanism : add symmetric matrix with i.i.d. Gaussians in upper triangular part of the matrix. (Dwork et al. 2014, STOC)

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

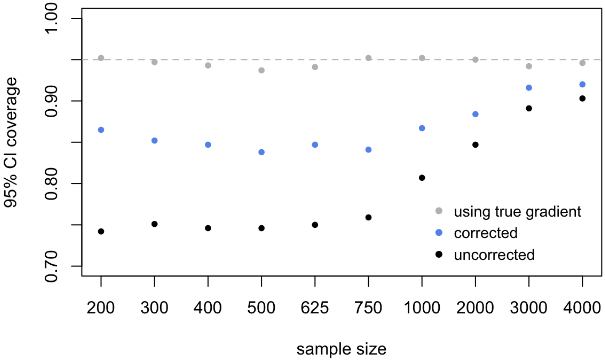
3. Compute $V_n(\theta^{(K)}) = \tilde{M}_n(\theta^{(K)})^{-1} \tilde{Q}_n(\theta^{(K)}) \tilde{M}_n(\theta^{(K)})^{-1}$

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 :

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



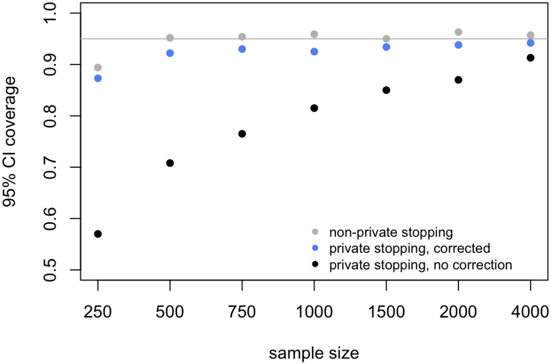
GDP Confidence Interval Coverage

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 ?

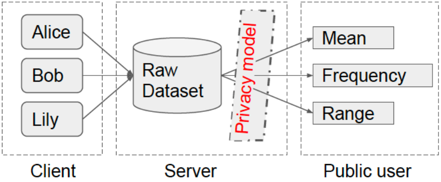
1. Algorithms are easy to implement and computationally efficient !

2. Importance of (local) strong convexity for optimal parametric rates of convergence

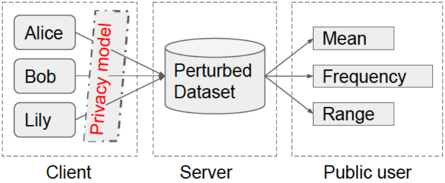
3. General framework for differentially private parametric inference

4. Connections between optimization, differential privacy and robust statistics

Extension to local differential privacy



(a) Centralized differential privacy



(b) Local differential privacy

Some key randomization ideas in local DP go back to Warner (JASA, 1965) in the official statistics literature !

Extension to local differential privacy

- Interactive noisy gradient descent :

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

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

Extension to local differential privacy

- Interactive noisy gradient descent :

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

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

- Under local strong convexity and smoothness, after K iterations of NGD iterations, with probability at least $1 - \tau$,

$$\|\theta^{(k+1)} - \hat{\theta}\|_2 \leq O\left(\frac{\log(n) \sup \|\Psi\|_2 \sqrt{p + \log(n/\tau)}}{\mu \sqrt{n}} \right).$$

Extension to local differential privacy

- Interactive noisy gradient descent :

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

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

- Under local strong convexity and smoothness, after K iterations of NGD iterations, with probability at least $1 - \tau$,

$$\|\theta^{(k+1)} - \hat{\theta}\|_2 \leq O\left(\frac{\log(n) \sup \|\Psi\|_2 \sqrt{p + \log(n/\tau)}}{\mu \sqrt{n}} \right).$$

- See Duchi, Jordan and Wainwright (JASA, 2019) for minimax analysis of various models under local DP.

Future work

There are many open problems in DP. In my opinion the following ones are among the most obvious and perhaps urgent for practical data analysis :

1. High dimensional statistical inference
2. Hyperparameter tuning
3. DP methodology for longitudinal data
4. Model diagnostics and visualization tool
5. Better inference for local DP model