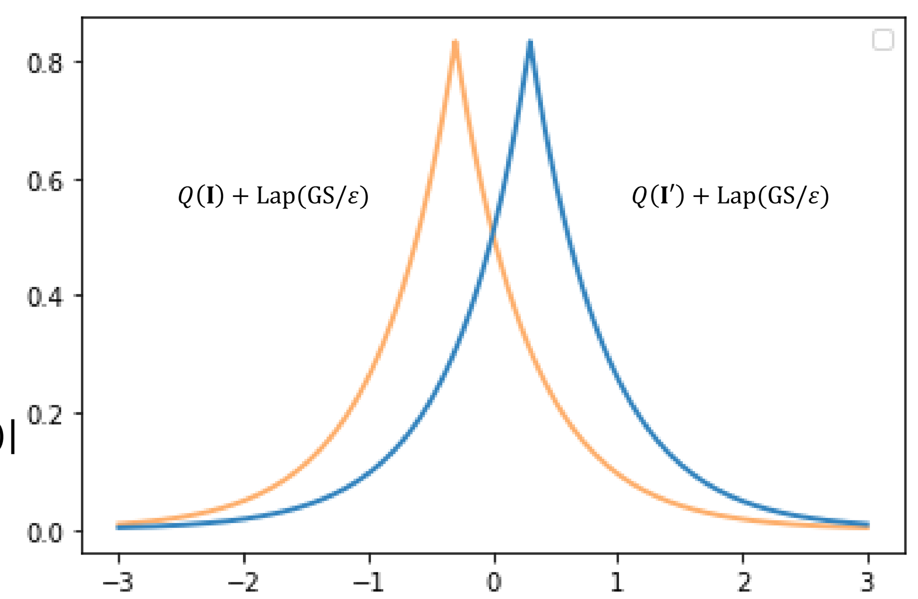
Sum Estimation under Personalized Local Differential Privacy

Dajun Sun, Wei Dong, Yuan Qiu, Ke Yi, Graham Cormode

HKUST, NTU, Southeast University, University of Warwick

Background

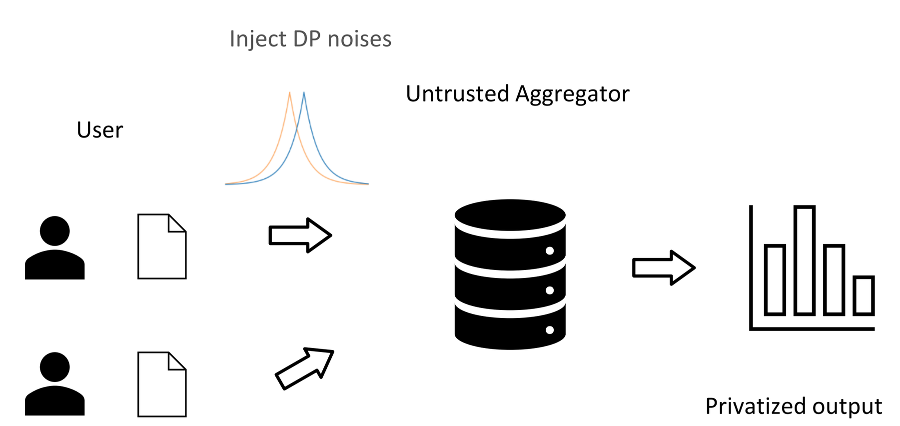
- Differential Privacy
  - Database instance $\mathbf{I}$
  - For any $\mathbf{I}, \mathbf{I}'$, they are neighbors ($\mathbf{I} \sim \mathbf{I}'$) if they differ by one individual's information
  - For $\varepsilon, \delta > 0$, a mechanism M is (ε, δ)-DP if for any $\mathbf{I} \sim \mathbf{I}'$, any subset of outputs Y
    $\Pr[M(\mathbf{I}) \in Y] \leq e^{\varepsilon} \cdot \Pr[M(\mathbf{I}') \in Y]$.
    - ε: controls privacy level



- Laplace mechanism
  - Denote the query as Q
  - Global sensitivity (GS):
    - $\mathrm{GS} = \max_{\mathbf{I}} \max_{\mathbf{I}', d(\mathbf{I}, \mathbf{I}')=1} |Q(\mathbf{I}) - Q(\mathbf{I}')|$
- Standard DP has uniform privacy parameter ε for all users
- Different users may have different requirements
  - Rich people may be more concerned about their privacy
  - People with some diseases may need stronger privacy protection

We need a more flexible DP model

- **Personalized Differential Privacy (PDP)**[Jorgensen et al. 2015]
  - Each user u specifies his own privacy parameter $\Phi(u)$ (analog to ε)
- Definition: A mechanism M is Φ-PDP if for any $\mathbf{I} \sim_u \mathbf{I}'$ that differ by user $u's$ information, M is $\Phi(u)$-DP, namely
  $\Pr[M(\mathbf{I}) \in Y] \leq e^{\Phi(u)} \cdot \Pr[M(\mathbf{I}') \in Y]$
- The PDP framework is meaningful in the local DP setting
  - Each user privatizes his data by himself using a local randomizer M
- **We want to extend our study of the PDP model to the local setting**



Problem Definition

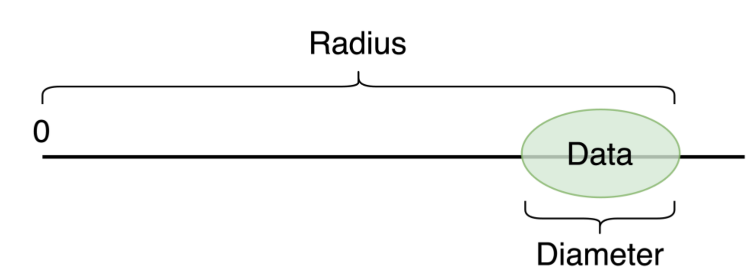
- We study the high-dimensional sum estimation problem under local PDP
- Assume n users, each user u holds:
  - An integer valued d-dimensional vector $\mathbf{I}(u) \in \{0,1,\dots,B\}^d$ (Private)
  - His privacy parameter $\Phi(u)$ (Public)
- Privacy requirement: For each user u and any pair of values $\mathbf{I}(u)$ and $\mathbf{I}'(u)$, his output through the local randomizer M should satisfy:
  $$D_\alpha\big(M(\mathbf{I}(u)) \,||\, M(\mathbf{I}'(u))\big) \leq \alpha \cdot \Phi(u)$$
  For any $\alpha > 1$, where $D_\alpha(\cdot \,||\, \cdot)$ denotes the α-Rényi divergence.
  - Known as Concentrated Differential Privacy (CDP) which has better composition properties in high dimensions
- Want to produce a privatized estimation for Sum($\mathbf{I}$)=$\sum_u \mathbf{I}(u)$

Local PDP Sum: First Attempt

- Consider a set of noise scales s
  - For each value of s, define user u's personalized truncation threshold $\tau(u) = s\sqrt{2\Phi(u)}$
    - Each user performs a truncation on the ℓ_2 norm of their data and obtains $\mathbf{I}_\tau(u) = \mathbf{I}(u)$ *min(1, $\frac{\tau(u)}{||\mathbf{I}(u)||_2}$)
    - Each user adds a Gaussian noise with scale proportional to s on $\mathbf{I}_\tau(u)$ and sends the result to the aggregator
  - The aggregator computes noisy sums for different s, and determine which value of s leads to the best result
- It's enough to examine $s = \frac{2^i}{\sqrt{2\rho_{\max}}}$ for $i = 1,2,\dots,O(\log B)$
- Our result: We achieve a **ℓ_2 error guarantee**
  $$\tilde{O}\left(\min_{s\in\mathbb{R}_{\geq 0}}\left(\left\|\sum_u \left(\|\mathbf{I}(u)\|_2 - s\sqrt{2\Phi(u)}\right)^+ \frac{\mathbf{I}(u)}{\|\mathbf{I}(u)\|_2}\right\|_2 + s\sqrt{nd}\right)\right)$$

Radius vs Diameter

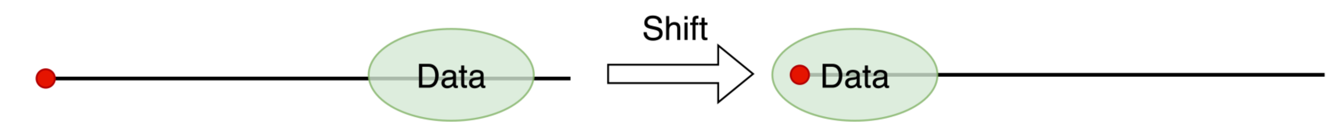
- When $\Phi(\cdot) \equiv \rho$, the previous bound degenerates to $\tilde{O}(\max_u ||\mathbf{I}(u)||_2 \sqrt{\frac{nd}{\rho}})$
  - rad($\mathbf{I}$)= $\max_u ||\mathbf{I}(u)||_2$ is known as the radius of the dataset
  - **Good for arbitrary dataset, but real data is usually concentrated**



- Consider the dataset where $\mathbf{I}(u_i) = \frac{B}{2} + i$ for i=1,2,…, n
  - **rad(I)=$\Theta(B)$** (assume $B \gg n$)
    - A noise proportional to B is too large
  - However, diameter of $\mathbf{I}$, denoted as $\boldsymbol{\omega(\mathbf{I}) = \max_{u_i,u_j} ||\mathbf{I}(u_i)\text{-}\mathbf{I}(u_j)||_2 = \Theta(n)}$
    - It will be better if we can achieve a noise proportional to n

Diameter Sum

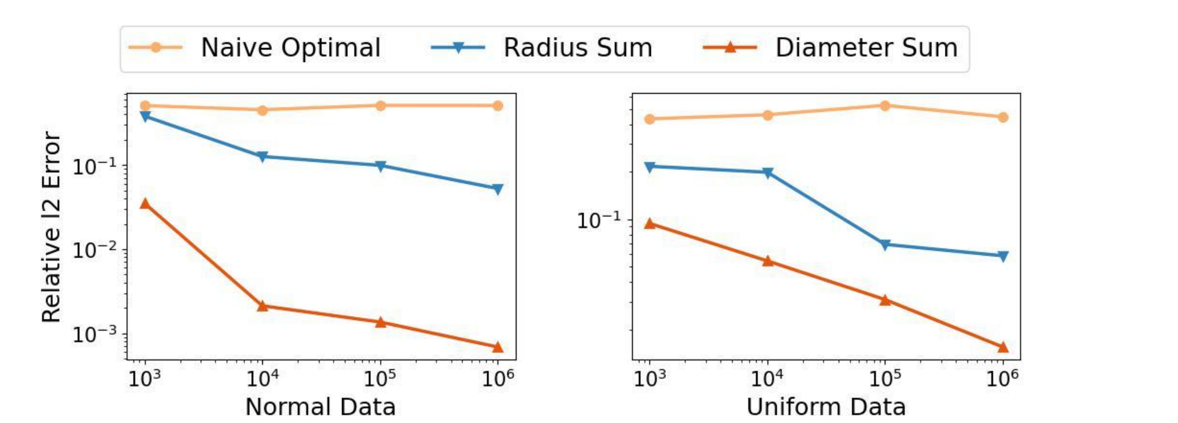
- Intuition: First shift the dataset toward the origin



  - Then applying the radius sum algorithm leads to a diameter bound
- The diameter sum algorithm provides a **ℓ_2 error guarantee**:
  - Error is $\tilde{O}\left(\min_{s\in\mathbb{R}_{\geq 0}}\left(\sqrt{\sum_u \mathbb{I}(s\sqrt{\Phi(u)/2} < \omega(\mathbf{I}))\omega(\mathbf{I})} + s\sqrt{nd}\right)\right)$
  - When $\Phi(\cdot) \equiv \rho$, it degenerates to $\tilde{O}(\omega(\mathbf{I})\sqrt{\frac{nd}{\rho}})$

Experiments

Data	Result ℓ_2 Norm	Technique	Relative ℓ_2 Error(%)	Time(s)
Normal Data	9.04×10^8	Naive	51.24	0.02
		Radius Sum	9.75	0.93
		Diameter Sum	**0.14**	16.10
Uniform Data	5.65×10^8	Naive	52.67	0.02
		Radius Sum	7.39	1.03
		Diameter Sum	**3.10**	16.50
MNIST Digit 0	4.39×10^7	Naive	85.77	0.02
		Radius Sum	41.84	0.92
		Diameter Sum	**6.54**	25.40



Varying number of users