Location privacy and random walk

March Boedihardjo (MSU)
Joint work with Thomas Strohmer and Roman Vershynin

Karlstad Applied Analysis Seminar
April 2024

Problem

Publish the locations of n individuals in a private way.

Examples

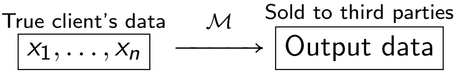
- Mobile phone location
- IP address location
- Covid patient location

Problem

Publish the locations of n individuals in a private way.

- Add noise to the locations.
- More noise $\Longrightarrow$ More privacy, less accuracy.
- Find the optimal trade off.

Definition of differential privacy



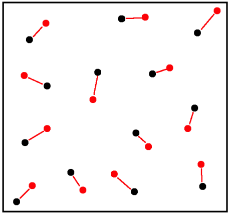
Definition

$\mathcal{M}$ is ϵ-differentially private if $\mathcal{M}$ is a randomized algorithm s.t.

$$e^{-\epsilon} \leq \frac{\mathbb{P}(\mathcal{M}(x_1, \ldots, \widetilde{x}_i, \ldots, x_n) \in S)}{\mathbb{P}(\mathcal{M}(x_1, \ldots, x_n) \in S)} \leq e^{\epsilon} \quad \forall i \; \forall \widetilde{x}_i \; \forall S.$$

Wasserstein distance

$$W(\{x_i\}_{1\le i\le n}, \{y_i\}_{1\le i\le n}) = \inf_{\sigma} \frac{1}{n}\sum_{i=1}^{n} \|x_i - y_{\sigma(i)}\|.$$



Can also define $W(\{x_i\}_{1\le i\le m}, \{y_i\}_{1\le i\le n})$ and $W(\mu_1,\mu_2)$ for probability measures μ_1, μ_2 on $\mathbb{R}^d$.

Problem

Publish the location of n individuals in a private way.

- Add noise to the locations.
- More noise $\Longrightarrow$ More privacy, less accuracy.
- Find the optimal trade off.

Problem

Publish the location of n individuals in a private way.

- Add noise to the locations.
- More noise $\Longrightarrow$ More privacy, less accuracy.
- Find the optimal trade off.

More precisely, design the noise such that

(1) it's ϵ-differentially private

(2) the error in the Wasserstein distance is minimized

One-dimensional locations

All locations in $[0, 1]$.

If μ_1 and μ_2 are probability measures on $[0, 1]$, then

$$W(\mu_1, \mu_2) = \int_0^1 |\mu_1([0, t]) - \mu_2([0, t])| \, dt.$$

One-dimensional locations

Problem

Design the noise such that

1. (1) it's ϵ-differentially private
2. (2) the error in the Wasserstein distance is minimized

This is equivalent to

Problem

Design the probability density $f(z) = \frac{1}{\beta} e^{V(z)}$ on $\mathbb{R}^n$ such that

1. (1) $|V(x) - V(y)| \leq \|x - y\|_1 \;\; \forall x, y \in \mathbb{R}^n$
2. (2) if $(Z_1, \ldots, Z_n) \sim f$, then

$$\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}|Z_1 + \ldots + Z_k| \quad \text{is minimized.}$$

Note: Add noise to the weights, not to the location.

One-dimensional locations

Problem

Design the probability density $f(z) = \frac{1}{\beta} e^{V(z)}$ on $\mathbb{R}^n$ such that

(1) $|V(x) - V(y)| \leq \|x - y\|_1 \ \forall x, y \in \mathbb{R}^n$

(2) if $(Z_1, \ldots, Z_n) \sim f$, then

$$\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}|Z_1 + \ldots + Z_k| \quad \text{is minimized.}$$

If $V(z) = -\|z\|_1$, then $Z_1, \ldots, Z_n$ are i.i.d.,
(1) is satisfied ✓
(2):

$$c\sqrt{n} \leq \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}|Z_1 + \ldots + Z_k| \leq C\sqrt{n}.$$

This is the classical random walk.

One-dimensional locations

Problem

Design the probability density $f(z) = \frac{1}{\beta} e^{V(z)}$ on $\mathbb{R}^n$ such that

(1) $|V(x) - V(y)| \le \|x - y\|_1 \ \forall x, y \in \mathbb{R}^n$

(2) if $(Z_1, \dots, Z_n) \sim f$, then

$$\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}|Z_1 + \dots + Z_k| \quad \text{is minimized.}$$

If $V(z) = -\|z\|_1$, then $Z_1, \dots, Z_n$ are i.i.d.,
(1) is satisfied ✓
(2):

$$c\sqrt{n} \le \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}|Z_1 + \dots + Z_k| \le C\sqrt{n}.$$

This is the classical random walk.

Question: Can we do better than $\sqrt{n}$?

One-dimensional locations

Problem

Design the probability density $f(z) = \frac{1}{\beta} e^{V(z)}$ on $\mathbb{R}^n$ such that

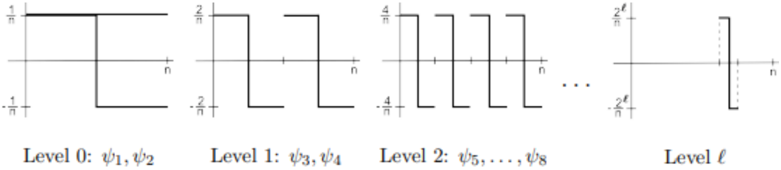
(1) $|V(x) - V(y)| \leq \|x - y\|_1 \ \forall x, y \in \mathbb{R}^n$

(2) if $(Z_1, \ldots, Z_n) \sim f$, then

$$\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}|Z_1 + \ldots + Z_k| \quad \text{is minimized.}$$

- Need mean reversion
- Tried stochastic differential equation
- But Brownian motion is not suitable, because it's ℓ^1 norm.

Haar basis



These functions serve as mean reversion functions.

$$k \mapsto \psi_j(1) + \ldots + \psi_j(k)$$

has a bump and then returns to 0.

One-dimensional locations

Problem

Design the probability density $f(z) = \frac{1}{\beta} e^{V(z)}$ on $\mathbb{R}^n$ such that

(1) $|V(x) - V(y)| \leq \|x - y\|_1 \ \forall x, y \in \mathbb{R}^n$

(2) if $(Z_1, \ldots, Z_n) \sim f$, then

$$\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}|Z_1 + \ldots + Z_k| \quad \text{is minimized.}$$

Take

$$Z_k = \sum_{j=1}^{n} \Lambda_j \psi_j(k) \quad \text{for } k = 1, \ldots, n,$$

where $\Lambda_1, \ldots, \Lambda_n$ are i.i.d. Laplace random variables, i.e., $\frac{1}{2b} e^{-\frac{|x|}{b}}$.

One-dimensional locations

Problem

Design the probability density $f(z) = \frac{1}{\beta} e^{V(z)}$ on $\mathbb{R}^n$ such that

(1) $|V(x) - V(y)| \leq \|x - y\|_1 \ \forall x, y \in \mathbb{R}^n$

(2) if $(Z_1, \ldots, Z_n) \sim f$, then

$$\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}|Z_1 + \ldots + Z_k| \quad \text{is minimized.}$$

Take

$$Z_k = \sum_{j=1}^{n} \Lambda_j \psi_j(k) \quad \text{for } k = 1, \ldots, n,$$

where $\Lambda_1, \ldots, \Lambda_n$ are i.i.d. Laplace random variables, i.e., $\frac{1}{2b} e^{-\frac{|x|}{b}}$.

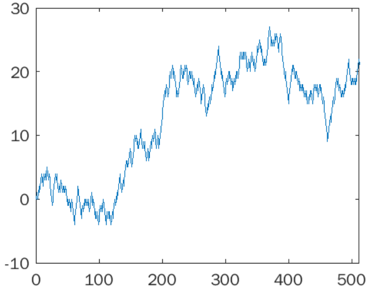
(1) is satisfied ✓

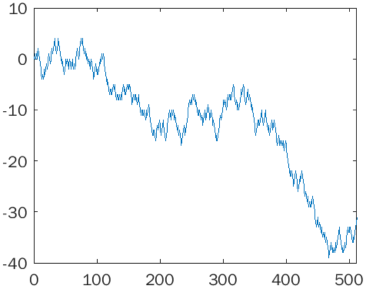
(2):

$$\max_{1 \leq k \leq n} \mathbb{E}|Z_1 + \ldots + Z_k| \leq C \log^{\frac{3}{2}} n.$$

Random walk

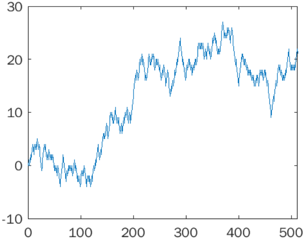
Classical random walk:

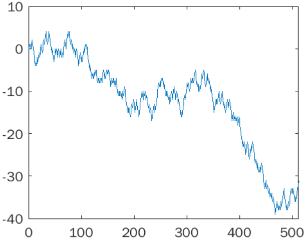




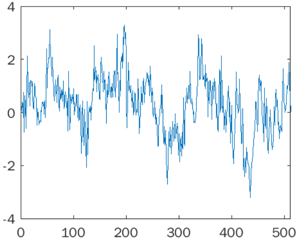
Random walk

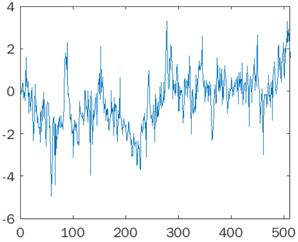
Classical random walk:





Super-regular random walk:





Main theorem (One-dimension)

Theorem (B., Strohmer, Vershynin, PTRF to appear)

There is an ϵ-differentially private algorithm for locations in $[0,1]$ such that the expected error in the Wasserstein distance is at most

$$\frac{C \log^{\frac{3}{2}} \epsilon n}{\epsilon n},$$

where n is the number of individuals.

Lower bound: For ϵ-DP algorithms, it's impossible to do better than $O(\frac{1}{n})$.

Main theorem (Higher-dimension)

Theorem

There is an ϵ-differentially private algorithm for locations in $[0,1]^d$ such that the expected error in the Wasserstein distance is at most

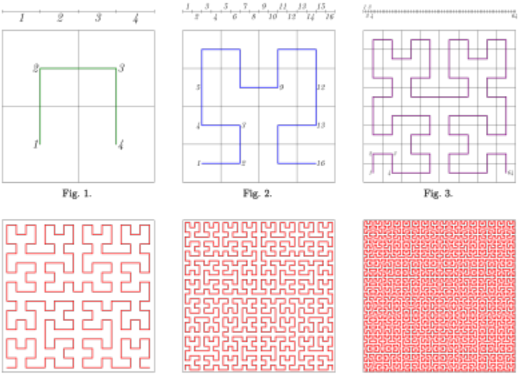
$$\left(\frac{C \log^{\frac{3}{2}} \epsilon n}{\epsilon n}\right)^{\frac{1}{d}},$$

where n is the number of individuals.

Lower bound: For ϵ-DP algorithms, it's impossible to do better than $O(n^{-1/d})$.

Proof of main theorem (Higher dimension)

Use a space-filling curve and apply the main result in 1D.

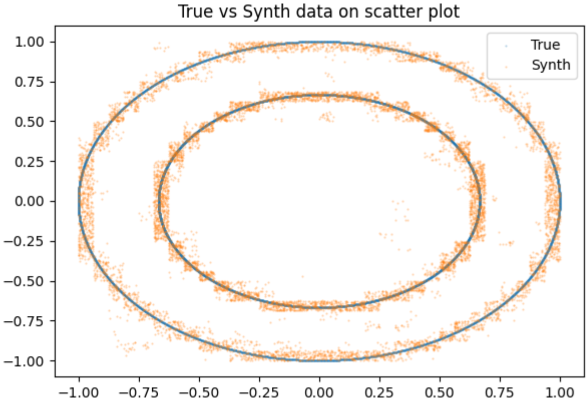


Source: Wiki

$n = 10,000$

$x_1, \ldots, x_n$: Points on the blue line

Private measure ν: Uniformly distributed on the orange points



Note: The 2 clusters are preserved.

Wasserstein distance

If $W(\mu_1, \mu_2)$ is small, then

(1) All Lipschitz queries are uniformly preserved:

$$W(\mu_1, \mu_2) = \sup_f \left| \int f \, d\mu_1 - \int f \, d\mu_2 \right|,$$

where the sup is over all 1-Lipschitz f.

Often times algorithms generating synthetic data require users to specify the queries f.

(2) Clusters are preserved (even non-convex clusters), since for any set S,

$$f_S(y) = \mathrm{dist}(y, S) = \inf_{x \in S} \|y - x\|$$

is a 1-Lipschitz function.

Prior result

Theorem (Wang et al 2016 JMLR)

There is an ϵ-differentially private algorithm for locations in $[0,1]^d$ such that the expected error

$$\sup_f \left| \int f \, d\mu_1 - \int f \, d\mu_2 \right|,$$

where the sup is over all K-smooth f, is at most

$$\frac{C}{\epsilon} n^{-\frac{K}{2d+K}}.$$

Our result: $K = 1$ with error $O(n^{-1/d} \cdot \mathrm{polylog}(n))$.
Optimal up to the polylog factor.

References

(1) *Private measures, random walks, and synthetic data*
M. Boedihardjo, T. Strohmer, R. Vershynin
Probability Theory and Related Fields, to appear

(2) *Differentially private data releasing for smooth queries*
Z. Wang, C. Jin, K. Fan, J. Zhang, J. Huang, Y. Zhong, L. Wang
Journal of Machine Learning Research (2016)