Differential privacy in distributed learning

Yi Yu
Department of Statistics, University of Warwick

A privacy mechanism is a randomised algorithm taking an input dataset $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ and producing publishable data Z. Formally, it is a collection of conditional distributions $\mathcal{Q} = \{Q(\cdot|x) : x \in \mathcal{X}^n\}$ such that

$$Z|\{X = x\} \sim Q(\cdot|x).$$

Privacy mechanism Q is called α-(central) differentially private (Dwork et al., 2006) if

$$\sup_A \frac{Q(A|x)}{Q(A|x')} = \sup_A \frac{\mathbb{P}(Z \in A|X = x)}{\mathbb{P}(Z \in A|X = x')} \leq e^{\alpha},$$

for all $x, x \ \in \mathcal{X}^n$ such that $\sum_{i=1}^n \mathbf{1}\{x_i \neq x_i'\} \leq 1$. We focus on the regime $\alpha \in (0, 1]$.

For the central differential privacy (CDP), where there is a trusted central data curator having access to all the raw data. For example, when estimating a univariate mean, we can have

$$\widehat{\theta} = Z = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n\alpha} W, \quad \text{with } W \sim \mathrm{Lap}(1).$$

Total added noise is of order $(n^2\alpha^2)^{-1}$.

A stronger notion of differential privacy is the local differential privacy (LDP), where data are randomised before collection, that is

$$\sup_{A} \sup_{x,x' \in \mathcal{X}} \frac{\mathbb{P}(Z_i \in A | X_i = x)}{\mathbb{P}(Z_i \in A | X_i = x')} \leq e^{\alpha}, \quad i \in \{1, \ldots, n\}.$$

For example, when estimating a univariate mean, we can have

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i=1}^{n} \left(X_i + \frac{1}{\alpha} W_i \right), \quad \text{with } \{W_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \mathrm{Lap}(1).$$

Total added noise is of order $(n\alpha^2)^{-1}$.

Remarks

- Non-interactive, sequentially interactive and fully-interactive LDP mechanisms.
- Pure and approximate DP.

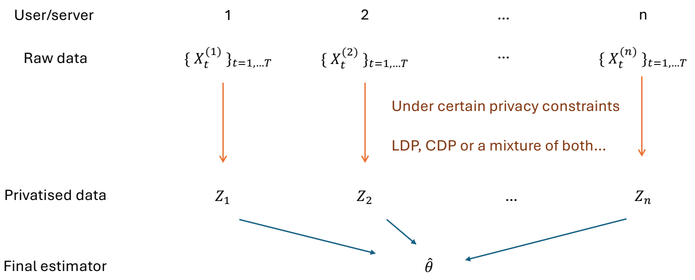
  Pure DP: $Q(A|x) \leq e^{\alpha} Q(A|x)$ and Approximate DP: $Q(A|x) \leq e^{\alpha} Q(A|x) + \beta$.
- Similarity: both CDP and LDP assume that each user possesses one unit of data.
- Difference: all raw data can be used before privatisation in CDP, but every unit of raw data needs to be privatised before any statistical inference in LDP.
- Question: do we have something in between when each user possesses multiple units of data?

Remarks

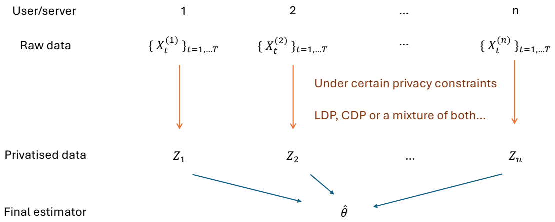
- Non-interactive, sequentially interactive and fully-interactive LDP mechanisms.
- Pure and approximate DP.

  Pure DP: $Q(A|x) \leq e^{\alpha} Q(A|x)$ and Approximate DP: $Q(A|x) \leq e^{\alpha} Q(A|x) + \beta$.
- Similarity: both CDP and LDP assume that each user possesses one unit of data.
- Difference: all raw data can be used before privatisation in CDP, but every unit of raw data needs to be privatised before any statistical inference in LDP.
- Question: do we have something in between when each user possesses multiple units of data?

User-level LDP & federated DP



User-level LDP & federated DP



- LDP: Rate optimality and phase transition for user-level local differential privacy (arXiv: 2405.11923, Alexander Kent, Thomas B. Berrett and **Y.**)
- CDP: Federated transfer learning with differential privacy (arXiv: 2403.11343, Mengchu Li, Ye Tian, Yang Feng and **Y.**)
- A mixture of both: Private distributed learning in functional data (ongoing work, Gengyu Xue, Zhenhua Lin and **Y.**)

A simple example: univariate mean estimation measured in squared loss, with n users/sites and T units of data per user.

Setting	Minimax rates	References
No privacy	$1/(nT)$	Very easy to show
Local item-level	$1/(nT\alpha^2)$	Duchi et al. (2018)
Local user-level (small T)	$1/(nT\alpha^2)$	Our result
Local user-level (large T)	$e^{-n\alpha^2}$	Our result
Central item-level	$1/(nT) \vee 1/(n^2T^2\alpha^2)$	Levy et al. (2021)
Central user-level (small T)	$1/(nT) \vee 1/(n^2T\alpha^2)$	Levy et al. (2021)
Federated	$1/(nT) \vee 1/(nT^2\alpha^2)$	Our result

Extensions

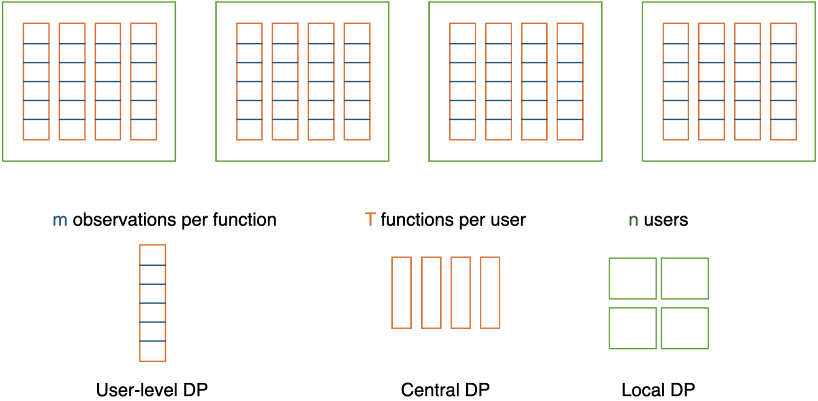
- Hierarchy
- Heterogeneity

A simple example: univariate mean estimation measured in squared loss, with n users/sites and T units of data per user.

Setting	Minimax rates	References
No privacy	$1/(nT)$	Very easy to show
Local item-level	$1/(nT\alpha^2)$	Duchi et al. (2018)
Local user-level (small T)	$1/(nT\alpha^2)$	Our result
Local user-level (large T)	$e^{-n\alpha^2}$	Our result
Central item-level	$1/(nT) \vee 1/(n^2T^2\alpha^2)$	Levy et al. (2021)
Central user-level (small T)	$1/(nT) \vee 1/(n^2T\alpha^2)$	Levy et al. (2021)
Federated	$1/(nT) \vee 1/(nT^2\alpha^2)$	Our result

Extensions

- Hierarchy
- Heterogeneity


m observations per function
T functions per user
n users
User-level DP
Central DP
Local DP

Sparse functional mean estimation: Sobolev class $\mathcal{W}(\gamma, C)$ mean function estimation measured in functional L_2-norm squared loss, with n users/sites, T functions data per user and m observational points per function.
Imposing central user-level for within each user and federated across users, we have

$$\frac{1}{nT} \vee \frac{1}{nT^2\alpha^2} \vee (nTm)^{-\frac{2\gamma}{2\gamma+1}} \vee (nT^2 m\alpha^2)^{-\frac{\gamma}{\gamma+1}}.$$

Private distributed learning in functional data (ongoing work, Gengyu Xue, Zhenhua Lin and **Y.**)

In general, we have that

$$\text{Minimax rate} \asymp \text{target-only minimax rate} \wedge \text{transfer-learning minimax rate},$$

where

$$\text{target-only rate} \asymp \text{non-private rate} \vee \text{central DP rate}$$

and

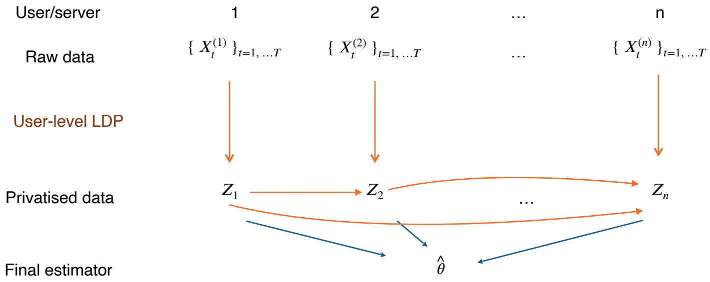
$$\begin{aligned}&\text{transfer-learning rate}\\ \asymp\ &\text{upper bound on source-target diff} \vee \text{non-private rate} \vee \text{federated DP rate}\end{aligned}$$

Problem	Target only	Transfer learning
Univariate mean estimation	$\frac{1}{T} + \frac{1}{T^2\alpha^2}$	$h^2 + \frac{1}{nT} + \frac{1}{nT^2\alpha^2}$
Low-dim regression	$\frac{d}{T} + \frac{d^2}{T^2\alpha^2}$	$h^2 + \frac{d}{nT} + \frac{d}{nT^2\alpha^2}$
High-dim regression	$\frac{s}{T} + \frac{s^2}{T^2\alpha^2}$	$h^2 + \frac{s}{nT} + \frac{sd}{nT^2\alpha^2}$

Federated transfer learning with differential privacy (arXiv: 2403.11343, Mengchu Li, Ye Tian, Yang Feng and **Y.**)

User-level local differential privacy

(with Alexander Kent and Thomas B. Berrett, arXiv: 2405.11923)


User/server
1
2
…
n
Raw data
$\{ X_t^{(1)} \}_{t=1, \ldots T}$
$\{ X_t^{(2)} \}_{t=1, \ldots T}$
…
$\{ X_t^{(n)} \}_{t=1, \ldots T}$
User-level LDP
Privatised data
Z_1
Z_2
…
Z_n
Final estimator
$\hat{\theta}$

- A minimax framework
- Infinite-T scenario with general minimax upper and lower bounds
- Finite-T scenario
  - Multivariate mean estimation (omitted in the talk)
  - Sparse, high-dimensional mean estimation
  - Nonparametric density estimation (omitted in the talk)

For $\alpha > 0$, a collection of conditional distributions $\{Q_i\}_{i=1}^n$ constitutes a user-level α-LDP mechanism if, for all $i \in \{1, \ldots, n\}$, all $x_{1:T}^{(i)}, {x'}_{1:T}^{(i)} \in \mathcal{X}^T$ and all $z_{1:(i-1)} \in \mathcal{Z}^{i-1}$,

$$\sup_S \frac{Q_i(Z_i \in S | X_{1:T}^{(i)} = x_{1:T}^{(i)}, Z_{1:(i-1)} = z_{1:(i-1)})}{Q_i(Z_i \in S | X_{1:T}^{(i)} = {x'}_{1:T}^{(i)}, Z_{1:(i-1)} = z_{1:(i-1)})} \leq e^{\alpha}.$$

We consider the user-level α-LDP minimax risk

$$\mathcal{R}_{n,T,\alpha}(\theta(P), \Phi \circ \rho) = \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}\{\Phi \circ \rho(\hat{\theta}, \theta(P))\}.$$

A motivating example

Estimating the mean of a distribution from the family $\mathcal{P} = \{P : \mathbb{E}_P(X) \in [-1, 1]\}$, we can show that the user-level LDP minimax risk is lower bounded

$$\mathcal{R}_{n,T,\alpha}\left(\theta(\mathcal{P}),\, (\cdot)^2\right) \gtrsim 1 \wedge \frac{1}{nT\alpha^2}.$$

This coincides with the item-level minimax rate (Duchi et al., 2018).

Question: When $T \to \infty$, will $\mathcal{R}_{n,T,\alpha}\left(\theta(\mathcal{P}),\, (\cdot)^2\right)$ vanish?

Answer: Up to logarithmic factor

$$\mathcal{R}_{n,\infty,\alpha}\left(\theta(\mathcal{P}),\, (\cdot)^2\right) \asymp e^{-cn\alpha^2},$$

where

- $\mathcal{R}_{n,\infty,\alpha}\left(\theta(\mathcal{P}),\, (\cdot)^2\right) = \mathcal{R}_{n,1,\alpha}\left(\theta(\mathcal{P}^\infty),\, (\cdot)^2\right)$ and
- $\mathcal{P}^\infty = \{\delta_\theta : \theta \in \theta(\mathcal{P})\}$ - collection of Dirac distributions.

A motivating example

Estimating the mean of a distribution from the family $\mathcal{P} = \{P : \mathbb{E}_P(X) \in [-1, 1]\}$, we can show that the user-level LDP minimax risk is lower bounded

$$\mathcal{R}_{n,T,\alpha}\big(\theta(\mathcal{P}),\ (\cdot)^2\big) \gtrsim 1 \wedge \frac{1}{nT\alpha^2}.$$

This coincides with the item-level minimax rate (Duchi et al., 2018).

Question: When $T \to \infty$, will $\mathcal{R}_{n,T,\alpha}\big(\theta(\mathcal{P}),\ (\cdot)^2\big)$ vanish?

Answer: Up to logarithmic factor

$$\mathcal{R}_{n,\infty,\alpha}\big(\theta(\mathcal{P}),\ (\cdot)^2\big) \asymp e^{-cn\alpha^2},$$

where

- $\mathcal{R}_{n,\infty,\alpha}\big(\theta(\mathcal{P}),\ (\cdot)^2\big) = \mathcal{R}_{n,1,\alpha}\big(\theta(\mathcal{P}^\infty),\ (\cdot)^2\big)$ and
- $\mathcal{P}^\infty = \{\delta_\theta : \theta \in \theta(\mathcal{P})\}$ - collection of Dirac distributions.

A motivating example

Estimating the mean of a distribution from the family $\mathcal{P} = \{P : \mathbb{E}_P(X) \in [-1, 1]\}$, we can show that the user-level LDP minimax risk is lower bounded

$$\mathcal{R}_{n,T,\alpha}\big(\theta(\mathcal{P}),\, (\cdot)^2\big) \gtrsim 1 \wedge \frac{1}{nT\alpha^2}.$$

This coincides with the item-level minimax rate (Duchi et al., 2018).

Question: When $T \to \infty$, will $\mathcal{R}_{n,T,\alpha}\big(\theta(\mathcal{P}),\, (\cdot)^2\big)$ vanish?

Answer: Up to logarithmic factor

$$\mathcal{R}_{n,\infty,\alpha}\big(\theta(\mathcal{P}),\, (\cdot)^2\big) \asymp e^{-cn\alpha^2},$$

where

- $\mathcal{R}_{n,\infty,\alpha}\big(\theta(\mathcal{P}),\, (\cdot)^2\big) = \mathcal{R}_{n,1,\alpha}\big(\theta(\mathcal{P}^\infty),\, (\cdot)^2\big)$ and
- $\mathcal{P}^\infty = \{\delta_\theta : \theta \in \theta(\mathcal{P})\}$ - collection of Dirac distributions.

General infinite-T rates

Given $\delta > 0$, let $N(\delta)$ be the δ-covering number of the metric space (Θ, ρ) with $\Theta = \theta(\mathcal{P})$ and suppose that $N(2\delta) > 1$. For $\alpha \in (0, 1]$ and with $\mathrm{diam}(\Theta) = \sup_{\theta,\theta' \in \Theta} \rho(\theta, \theta')$, it holds that

$$\frac{\Phi(\delta)}{2}\left\{1 - \frac{12n\alpha^2 + \log(2)}{\log\big(N(2\delta)\big)}\right\} \leq \mathcal{R}_{n,\infty,\alpha}\big(\theta(\mathcal{P}),\ \Phi \circ \rho\big)$$

$$\leq \Phi(\delta) + \Phi\big(\mathrm{diam}(\Theta)\big) N(\delta) e^{-n\alpha^2/20}.$$

$$\frac{\Phi(\delta)}{2}\left\{1-\frac{12n\alpha^2+\log(2)}{\log\left(N(2\delta)\right)}\right\}\leq\mathcal{R}_{n,\infty,\alpha}\left(\theta(\mathcal{P}),\,\Phi\circ\rho\right)$$
$$\leq\Phi(\delta)+\Phi\left(\mathrm{diam}(\Theta)\right)N(\delta)e^{-n\alpha^2/20}$$

Remarks

- For all $T\in\mathbb{N}$, it holds that

$$\mathcal{R}_{n,T,\alpha}\left(\theta(\mathcal{P}),\,\Phi\circ\rho\right)\gtrsim\frac{\Phi(\delta)}{2}\left\{1-\frac{12n\alpha^2+\log(2)}{\log\left(N(2\delta)\right)}\right\}.$$

- Choosing

$$N(2\delta_{\mathrm{LB}})\geq\exp\left(\lceil 24n\alpha^2+2\log(2)\rceil\right)\text{ and }\Phi(\delta_{\mathrm{UB}})\geq\Phi\left(\mathrm{diam}(\Theta)\right)N(\delta_{\mathrm{UB}})e^{-n\alpha^2/20},$$

we have that

$$\Phi(\delta_{\mathrm{LB}})\lesssim\mathcal{R}_{n,\infty,\alpha}\left(\theta(\mathcal{P}),\,\Phi\circ\rho\right)\lesssim\Phi(\delta_{\mathrm{UB}}).$$

$$\frac{\Phi(\delta)}{2}\left\{1-\frac{12n\alpha^2+\log(2)}{\log\left(N(2\delta)\right)}\right\} \leq \mathcal{R}_{n,\infty,\alpha}\left(\theta(\mathcal{P}),\, \Phi\circ\rho\right)$$
$$\leq \Phi(\delta)+\Phi\left(\mathrm{diam}(\Theta)\right)N(\delta)e^{-n\alpha^2/20}$$

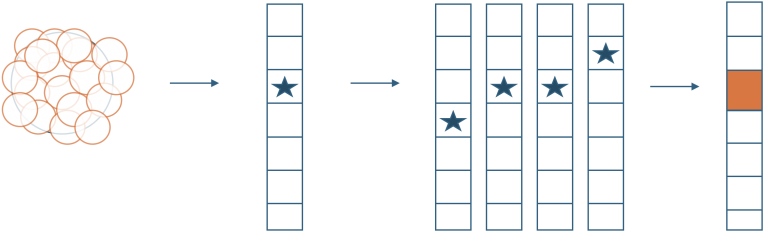
The lower bound is due to an application of Fano's inequality and an upper bound on the private Kullback–Leibler divergence (Duchi et al., 2018).

The upper bound is obtained via a non-interactive mechanism with a voting procedure.

$$\mathcal{R}_{n,\infty,\alpha}\big(\theta(\mathcal{P}),\, \Phi \circ \rho\big) \leq \Phi(\delta) + \Phi\big(\mathrm{diam}(\Theta)\big) N(\delta) e^{-n\alpha^2/20}$$

An upper bound procedure

- Step 1. Construct a δ-cover of (Θ, ρ) and make it non-overlapping.
- Step 2. Each user publicises a private vote for which ball their data lie in.
- Step 3. Output the centre of the majority-vote ball.



$$\mathcal{R}_{n,\infty,\alpha}\big(\theta(\mathcal{P}),\ \Phi \circ \rho\big) \leq \Phi(\delta) + \Phi\big(\mathrm{diam}(\Theta)\big) N(\delta) e^{-n\alpha^2/20}$$

An upper bound procedure

- Step 1. Construct a δ-cover of (Θ, ρ) and make it non-overlapping.
- Step 2. Each user publicises a private vote for which ball their data lie in.
- Step 3. Output the centre of the majority-vote ball.

Interpretation of the upper bound

- $\Phi(\delta)$ - the error occurred when the correct ball is chosen.
- $\Phi\big(\mathrm{diam}(\Theta)\big)$ - the error occurred when the correct ball is not chosen.
- $N(\delta)e^{-n\alpha^2/20}$ - the probability upper bound of the correct ball is not chosen.

Applications of the general bounds

	d-dim. mean ($\mathbb{B}_2(1)$)	Sparse mean	Density (Sobolev β-smooth)
No privacy	d/n	$s\log(d/s)/n$	$n^{-2\beta/(2\beta+1)}$
$\mathcal{P}$	$d/(n\alpha^2)$	$sd/(n\alpha^2)$	$(n\alpha^2)^{-2\beta/(2\beta+2)}$
$\mathcal{P}^\infty$	$e^{-n\alpha^2/d}$	$e^{-n\alpha^2/s}$	$(n\alpha^2)^{-2\beta}$

Consider the family of distributions

$$\mathcal{P}_{d,s} = \left\{P : \text{supp}(P) \subset \mathbb{B}_\infty(1) \subset \mathbb{R}^d, \ \|\mathbb{E}_P(X)\|_0 \leq s\right\}$$

and the functional $\theta(P) = \mathbb{E}_P(X)$.

Theorem For s satisfying $16\log(d)/3 \leq s \leq d$, assume that $n\alpha^2 \gtrsim s\log(ed)$. We have that

$$s\left[\frac{1}{T} \wedge \left\{\left(1 + \frac{d}{n\alpha^2}\right)^{1/T} - 1\right\}\right] \vee e^{-Cn\alpha^2/s} \lesssim \mathcal{R}_{n,T,\alpha}\left(\theta(\mathcal{P}_{d,s}), \|\cdot\|_2^2\right)$$
$$\left\{\frac{s\log(nT\alpha^2 d)}{T} \vee e^{-cn\alpha^2/s}\right\} \wedge \left\{\frac{sd\log^2(nT\alpha^2)}{nT\alpha^2} \vee e^{-cn\alpha^2/d}\right\}.$$

Consider the family of distributions

$$\mathcal{P}_{d,s} = \left\{P : \operatorname{supp}(P) \subset \mathbb{B}_\infty(1) \subset \mathbb{R}^d, \ \|\mathbb{E}_P(X)\|_0 \leq s\right\}$$

and the functional $\theta(P) = \mathbb{E}_P(X)$.

THEOREM For s satisfying $16\log(d)/3 \leq s \leq d$, assume that $n\alpha^2 \gtrsim s\log(ed)$. We have that

$$s\left[\frac{1}{T} \wedge \left\{\left(1 + \frac{d}{n\alpha^2}\right)^{1/T} - 1\right\}\right] \vee e^{-Cn\alpha^2/s} \lesssim \mathcal{R}_{n,T,\alpha}\left(\theta(\mathcal{P}_{d,s}), \|\cdot\|_2^2\right)$$
$$\left\{\frac{s\log(nT\alpha^2 d)}{T} \vee e^{-cn\alpha^2/s}\right\} \wedge \left\{\frac{sd\log^2(nT\alpha^2)}{nT\alpha^2} \vee e^{-cn\alpha^2/d}\right\}.$$

$$s\left[\frac{1}{T} \wedge \left\{\left(1+\frac{d}{n\alpha^2}\right)^{1/T} - 1\right\}\right] \vee e^{-Cn\alpha^2/s} \lesssim \mathcal{R}_{n,T,\alpha}\left(\theta(\mathcal{P}_{d,s}),\ \|\cdot\|_2^2\right)$$
$$\left\{\frac{s\log(nT\alpha^2 d)}{T} \vee e^{-cn\alpha^2/s}\right\} \wedge \left\{\frac{sd\log^2(nT\alpha^2)}{nT\alpha^2} \vee e^{-cn\alpha^2/d}\right\}.$$

Remarks

Roughly speaking, under the condition that $T \gtrsim \log\{d/(n\alpha^2)\}$, we consider two regimes.

- If $n\alpha^2 \lesssim d^\gamma$, for some $0 < \gamma < 1$, then up to logarithmic factors
$$\mathcal{R}_{n,T,\alpha}\left(\theta(\mathcal{P}_{d,s}),\ \|\cdot\|_2^2\right) \asymp s/T \vee e^{-Cn\alpha^2/s}.$$
- If $n\alpha^2 \gtrsim d\log(nT\alpha^2)$, then up to logarithmic factors
$$\mathcal{R}_{n,T,\alpha}\left(\theta(\mathcal{P}_{d,s}),\ \|\cdot\|_2^2\right) \asymp sd/(nT\alpha^2).$$

Roughly speaking, we say the rate is

$$\mathcal{R}_{n,T,\alpha}\left(\theta(\mathcal{P}_{d,s}),\ \|\cdot\|_2^2\right) \asymp \frac{s}{T} \vee \frac{s}{T}\frac{d}{n\alpha^2} \vee e^{-Cn\alpha^2/s}.$$

$$s\left[\frac{1}{T} \wedge \left\{\left(1+\frac{d}{n\alpha^2}\right)^{1/T} - 1\right\}\right] \vee e^{-Cn\alpha^2/s} \lesssim \mathcal{R}_{n,T,\alpha}(\theta(\mathcal{P}_{d,s}), \|\cdot\|_2^2)$$
$$\left\{\frac{s\log(nT\alpha^2 d)}{T} \vee e^{-cn\alpha^2/s}\right\} \wedge \left\{\frac{sd\log^2(nT\alpha^2)}{nT\alpha^2} \vee e^{-cn\alpha^2/d}\right\}.$$

The lower bound is due to an application of Assouad's method and an upper bound on the private total-variation distance (Acharya et al., 2023).

The upper bound is obtained by a two-component procedure depending on the value of T.

- Large T. If $n\alpha^2 \lesssim d\log(nT\alpha^2)$, then we summon a hashing-type voting procedure. Half of the users voting for the non-zero coordinates and the other half conduct an s-dimensional mean estimation.
- Small T. If $n\alpha^2 \gtrsim d\log(nT\alpha^2)$, then we summon a thresholding step after initial estimation.

In the large T scenario, the intuition is that T data points are enough to obtain a good enough coordinate selection.

With a pre-specified threshold ε, which is also used to select entries to be non-zero as long as the T-sample average exceeds ε, let

$$S_1 = \{j : |\theta_j| > 2\varepsilon\}, \quad S_2 = \{j : 0 < |\theta_j| \le 2\varepsilon\} \quad \text{and} \quad S_0 = \{j : \theta_j = 0\}.$$

Let $\mathcal{A}$ be the event that S_1 are all chosen and S_0 are all non-chosen.

the estimation error follows

$$\begin{aligned} & \mathbb{E}\{\|\hat{\theta} - \theta\|_2^2\} \\ \lesssim & \sum_{j:\hat{\theta}_j=0,\theta_j=0} 0 + \sum_{j:\hat{\theta}_j=0,\theta_j\neq 0} [\varepsilon^2 \mathbb{P}\{\mathcal{A}\} + 1\mathbb{P}\{\mathcal{A}^c\}] + \sum_{j:\hat{\theta}_j\neq 0} \text{error} \\ \lesssim & 0 + s\varepsilon^2 + s\mathbb{P}\{\mathcal{A}^c\} + s\text{-dim vector est error rate} \\ \lesssim & \frac{s\log(dT\alpha^2)}{T} + \frac{s^2\log(nT\alpha^2/s)}{nT\alpha^2} \vee e^{-Cn\alpha^2/s} \end{aligned}$$

Finite-T scenario: Sparse, high-dim. mean estimation

In the large T scenario, the intuition is that T data points are enough to obtain a good enough coordinate selection.

With a pre-specified threshold ε, which is also used to select entries to be non-zero as long as the T-sample average exceeds ε, let

$$S_1 = \{j:\ |\theta_j| > 2\varepsilon\}, \quad S_2 = \{j:\ 0 < |\theta_j| \leq 2\varepsilon\} \quad \text{and} \quad S_0 = \{j:\ \theta_j = 0\}.$$

Let $\mathcal{A}$ be the event that S_1 are all chosen and S_0 are all non-chosen.

the estimation error follows

$$\begin{aligned}
&\mathbb{E}\{\|\hat{\theta} - \theta\|_2^2\} \\
\lesssim &\sum_{j:\hat{\theta}_j=0,\theta_j=0} 0 + \sum_{j:\hat{\theta}_j=0,\theta_j\neq 0} \left[\varepsilon^2\mathbb{P}\{\mathcal{A}\} + 1\mathbb{P}\{\mathcal{A}^c\}\right] + \sum_{j:\hat{\theta}_j\neq 0} \text{error} \\
\lesssim &0 + s\varepsilon^2 + s\mathbb{P}\{\mathcal{A}^c\} + s\text{-dim vector est error rate} \\
\lesssim &\frac{s\log(dT\alpha^2)}{T} + \frac{s^2\log(nT\alpha^2/s)}{nT\alpha^2} \vee e^{-Cn\alpha^2/s}
\end{aligned}$$

In the large T scenario, the intuition is that T data points are enough to obtain a good enough coordinate selection.

With a pre-specified threshold ε, which is also used to select entries to be non-zero as long as the T-sample average exceeds ε, let

$$S_1 = \{j:\ |\theta_j| > 2\varepsilon\}, \quad S_2 = \{j:\ 0 < |\theta_j| \leq 2\varepsilon\} \quad \text{and} \quad S_0 = \{j:\ \theta_j = 0\}.$$

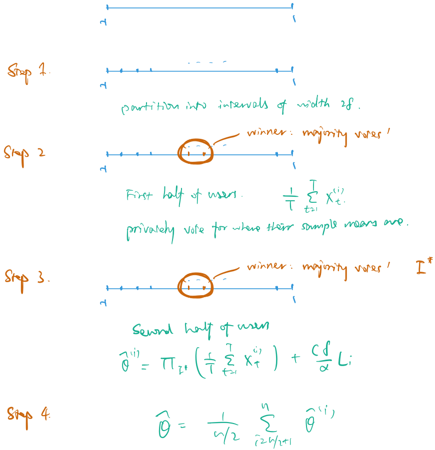
Let $\mathcal{A}$ be the event that S_1 are all chosen and S_0 are all non-chosen.

the estimation error follows

$$\begin{aligned}
&\mathbb{E}\{\|\hat{\theta} - \theta\|_2^2\} \\
\lesssim &\sum_{j:\hat{\theta}_j=0,\theta_j=0} 0 + \sum_{j:\hat{\theta}_j=0,\theta_j\neq 0} \left[\varepsilon^2\mathbb{P}\{\mathcal{A}\} + 1\mathbb{P}\{\mathcal{A}^c\}\right] + \sum_{j:\hat{\theta}_j\neq 0} \text{error} \\
\lesssim &0 + s\varepsilon^2 + s\mathbb{P}\{\mathcal{A}^c\} + s\text{-dim vector est error rate} \\
\lesssim &\frac{s\log(dT\alpha^2)}{T} + \frac{s^2\log(nT\alpha^2/s)}{nT\alpha^2} \vee e^{-Cn\alpha^2/s}
\end{aligned}$$

Lying in the core of the sparse, high-dimensional mean estimation procedures is a multivariate mean estimation procedure (with dist. supported on $\mathbb{B}_\infty(1)$).

Lying in the core of the multivariate ($\mathbb{B}_\infty(1)$) mean estimation procedure is a univariate mean estimation procedure (with dist. supported on $[-1, 1]$).

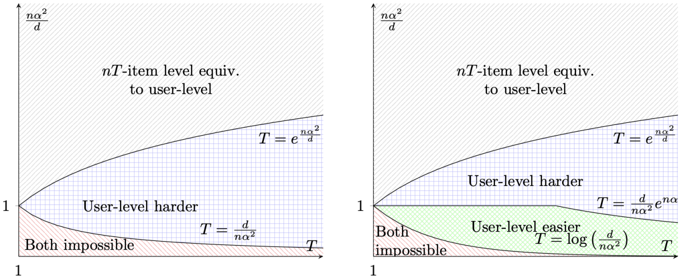


$$s\left[\frac{1}{T} \wedge \left\{\left(1+\frac{d}{n\alpha^2}\right)^{1/T} - 1\right\}\right] \vee e^{-Cn\alpha^2/s} \lesssim \mathcal{R}_{n,T,\alpha}\left(\theta(\mathcal{P}_{d,s}), \|\cdot\|_2^2\right)$$
$$\left\{\frac{s\log(nT\alpha^2 d)}{T} \vee e^{-cn\alpha^2/s}\right\} \wedge \left\{\frac{sd\log^2(nT\alpha^2)}{nT\alpha^2} \vee e^{-cn\alpha^2/d}\right\}.$$

Discussions

- Comparisons with item-level LDP rates.
- The exponential terms in upper and lower bounds: Where are they from?
- What if we do not have the knowledge of s?

FINITE-T SCENARIO



(a) Estimation on ℓ_2-ball

(b) Sparse mean estimation ($s = 1$)

	d-dim. mean ($\mathbb{B}_2(1)$)	s-sparse d-dim. mean	Density (Sobolev β-smooth)
Small T	$d/(nT\alpha^2)$	$s/T \wedge sd/(nT\alpha^2)$	$(nT\alpha^2)^{-2\beta/(2\beta+2)}$
Large T	$e^{-n\alpha^2/d}$	$e^{-n\alpha^2/s}$	$(n\alpha^2)^{-2\beta}$
Boundary	$e^{n\alpha^2/d}$	$\begin{cases} s^{n\alpha^2/s}, & d/(n\alpha^2) \gtrsim 1 \\ e^{n\alpha^2/d}, & d/(n\alpha^2) \lesssim 1 \end{cases}$	$(n\alpha^2)^{2\beta+1}$

- User-level LDP in other statistical tasks, e.g. testing.
- Mixture of different notions of DP, including use of public data in distributed learning.
- Phase transition regarding T in FDP.
- Large ε.
- Adaptivity.
- Dependent data.

- User-level LDP in other statistical tasks, e.g. testing.
- Mixture of different notions of DP, including use of public data in distributed learning.
- Phase transition regarding T in FDP.
- Large ε.
- Adaptivity.
- Dependent data.