Optimal federated learning under differential privacy constraints

Yi Yu
Department of Statistics, University of Warwick

The New York Times

GIVE THE

The 2020 Census Suggests That People Live Underwater. There's a Reason.

Technology advances forced the Census Bureau to use sweeping measures to ensure privacy for respondents. The ensuing debate goes to the heart of what a census is.

Share full article

The Census Bureau says that 14 people live in this bend in the Chicago River. It's one of thousands of bits of incorrect data in the 2020 census meant to protect the privacy of census respondents. Jamie Kelter Davis for The New York Times

https://www.theverge.com/2015/3/10/8177683/apple-research-kit-app-ethics-medical-research

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 ε-(central) differentially private (Dwork et al., 2006) if

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

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

At a high level, this quantifies how similar the private outcomes are in terms of total variation distance, by changing one out of n samples.

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 ε-(central) differentially private (Dwork et al., 2006) if

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

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

At a high level, this quantifies how similar the private outcomes are in terms of total variation distance, by changing one out of n samples.

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 ε-(central) differentially private (Dwork et al., 2006) if

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

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

At a high level, this quantifies how similar the private outcomes are in terms of total variation distance, by changing one out of n samples.

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

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

The variance of total added noise is of order $(n^2\varepsilon^2)^{-1}$.

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

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

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

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

The variance of total added noise is of order $(n\varepsilon^2)^{-1}$.

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

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

The variance of total added noise is of order $(n^2\varepsilon^2)^{-1}$.

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

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

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

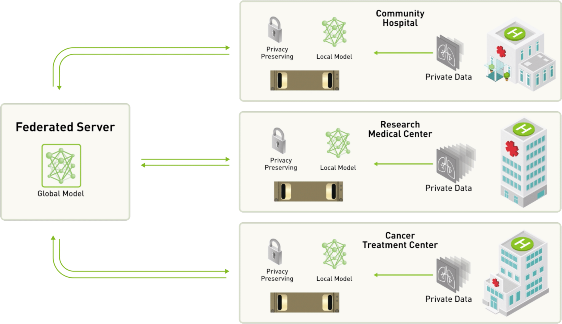
$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i=1}^{n}\left(X_i + \frac{1}{\varepsilon} W_i\right), \quad \text{with } \{W_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \mathrm{Lap}(1).$$

The variance of total added noise is of order $(n\varepsilon^2)^{-1}$.

Remarks

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

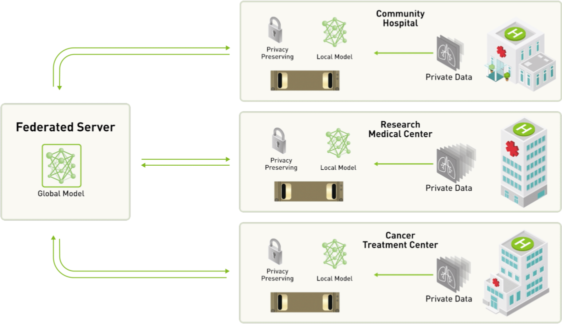
  Pure DP: $Q(A|x) \leq e^{\varepsilon} Q(A|x)$ and Approximate DP: $Q(A|x) \leq e^{\varepsilon} Q(A|x) + \delta$.



https://blogs.nvidia.com/blog/what-is-federated-learning/

Challenges

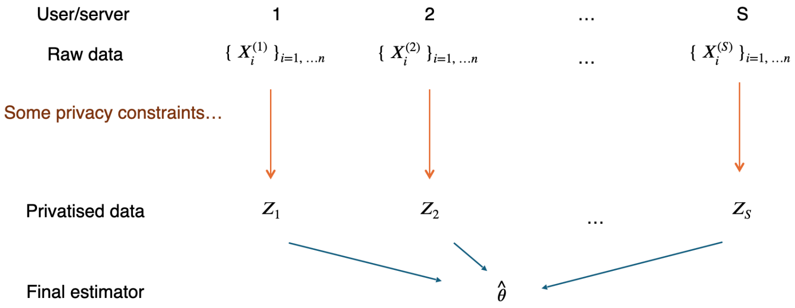
- Heterogeneity: distributions, privacy requirement types, privacy budgets.
- Communications: efficiency in aggregating and communicating siloed information.

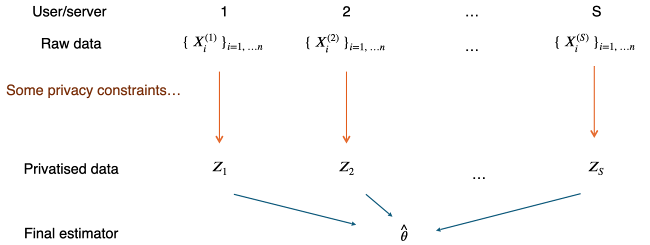


https://blogs.nvidia.com/blog/what-is-federated-learning/

Challenges

- Heterogeneity: distributions, privacy requirement types, privacy budgets.
- Communications: efficiency in aggregating and communicating siloed information.


User/server
1
2
…
S
Raw data
$\{ X_i^{(1)} \}_{i=1,\ldots n}$
$\{ X_i^{(2)} \}_{i=1,\ldots n}$
…
$\{ X_i^{(S)} \}_{i=1,\ldots n}$
Some privacy constraints…
Privatised data
Z_1
Z_2
…
Z_S
Final estimator
$\hat{\theta}$

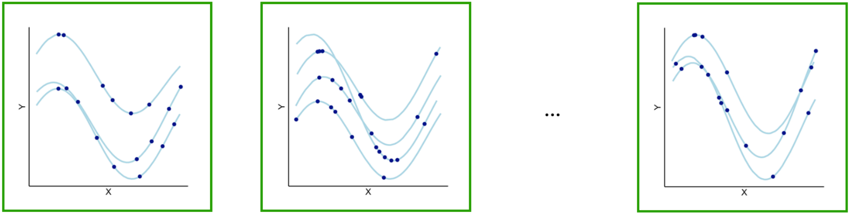


- User-level DP: Rate optimality and phase transition for user-level local differential privacy (arXiv: 2405.11923, Alexander Kent, Thomas B. Berrett and **Y.**)
- CDP: Federated transfer learning with differential privacy (arXiv: 2403.11343, Mengchu Li, Ye Tian, Yang Feng and **Y.**)
- A mixture of both: Private distributed learning in functional data (arXiv:2412.06582, Gengyu Xue, Zhenhua Lin and **Y.**)

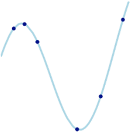
A simple example: univariate mean estimation measured in squared loss, with S users/sites and n units of data per user.

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

Server 1 | Server 2 | ... | Server S

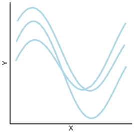


m observations per function



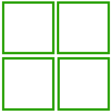
User-level DP

n_s functions per site

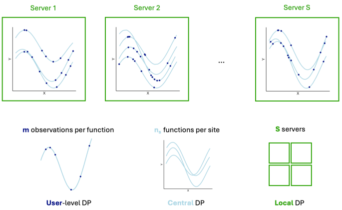


Central DP

S servers



Local DP

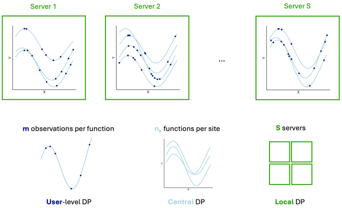


- Optimal estimation in private distributed functional data analysis (arXiv: 2412.06582)

Gengyu Xue
(Univ. of Warwick)

Zhenhua Lin
(National Univ. of Singapore)

Cai, T., Chakraborty, A., & Vuursteen, L. (2024). Optimal Federated Learning for Functional Mean Estimation under Heterogeneous Privacy Constraints. *arXiv preprint arXiv:2412.18992.*



- Optimal estimation in private distributed functional data analysis (arXiv: 2412.06582)

Gengyu Xue
(Univ. of Warwick)

Zhenhua Lin
(National Univ. of Singapore)

Cai, T., Chakraborty, A., & Vuursteen, L. (2024). Optimal Federated Learning for Functional Mean Estimation under Heterogeneous Privacy Constraints. *arXiv preprint arXiv:2412.18992.*

- Data: we have

$$\left\{\left(X_j^{(s,i)},\ Y_j^{(s,i)}\right)\right\}_{i,j,s=1}^{n,m,S},$$

i.e. m observations per function, n functions per server and S servers.

- Model:

$$Y_j^{(s,i)} = \mu^*(X_j^{(s,i)}) + U^{(s,i)}(X_j^{(s,i)}) + \xi_{s,ij},$$

where

  - $\{X_j^{(s,i)}\}$ are observation grids on $[0, 1]$,
  - $\mu^*(\cdot)$ is a deterministic function and is the goal of estimation,
  - $\{U^{(s,i)}(\cdot)\}$ are random mean-zero functions, and
  - $\{\xi_{s,ij}\}$ are measurement error random variables.

- Data: we have
$$\Big\{\big(X_j^{(s,i)},\ Y_j^{(s,i)}\big)\Big\}_{i,j,s=1}^{n,m,S},$$
i.e. m observations per function, n functions per server and S servers.

- Model:
$$Y_j^{(s,i)} = \mu^*(X_j^{(s,i)}) + U^{(s,i)}(X_j^{(s,i)}) + \xi_{s,ij},$$
where
  - $\{X_j^{(s,i)}\}$ are observation grids on $[0, 1]$,
  - $\mu^*(\cdot)$ is a deterministic function and is the goal of estimation,
  - $\{U^{(s,i)}(\cdot)\}$ are random mean-zero functions, and
  - $\{\xi_{s,ij}\}$ are measurement error random variables.

- Independent sampling grids
- α-Sobolev functions: mean and noise functions
- Sub-Gaussianity: noise functions norms and measurement error

(For notational simplicity, we only present an easy case.)

Assume that

$$n_s = n, \quad m_s = m, \quad \varepsilon_s = \varepsilon \quad \text{and} \quad \delta_s = \delta.$$

We have that,

$$\inf_{Q \in \mathcal{Q}_{\varepsilon,\delta}} \inf_{\widetilde{\mu}} \sup_{P_X, P_Y} \mathbb{E}_{P_X, P_Y, Q} \left\| \widetilde{\mu} - \mu^* \right\|_{L_2}^2 \asymp \frac{1}{Sn} \vee \left(\frac{1}{Snm} \right)^{\frac{2\alpha}{2\alpha+1}} \vee \frac{1}{Sn^2\varepsilon^2} \vee \left(\frac{1}{Sn^2 m\varepsilon^2} \right)^{\frac{2\alpha}{2\alpha+2}}.$$

$$\inf_{Q \in \mathcal{Q}_{\varepsilon,\delta}} \inf_{\widetilde{\mu}} \sup_{P_X, P_Y} \mathbb{E}_{P_X, P_Y, Q} \|\widetilde{\mu} - \mu^*\|_{L_2}^2 \asymp \frac{1}{Sn} \vee \left(\frac{1}{Snm}\right)^{\frac{2\alpha}{2\alpha+1}} \vee \frac{1}{Sn^2\varepsilon^2} \vee \left(\frac{1}{Sn^2 m\varepsilon^2}\right)^{\frac{2\alpha}{2\alpha+2}}$$

- Privacy vs. no privacy

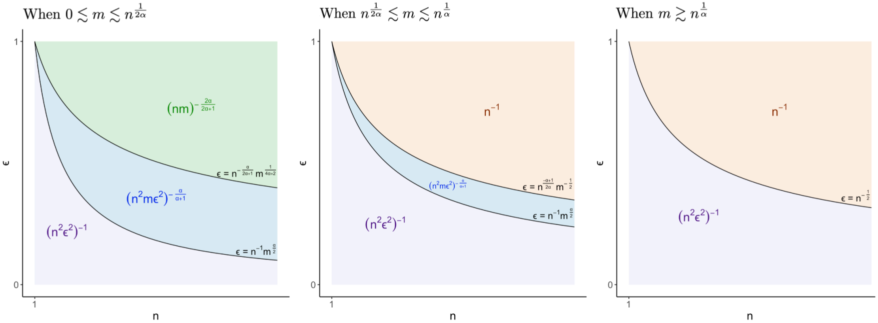
$$\frac{1}{Sn^2\varepsilon^2} \vee \left(\frac{1}{Sn^2 m\varepsilon^2}\right)^{\frac{2\alpha}{2\alpha+2}} \quad \text{vs.} \quad \frac{1}{Sn} \vee \left(\frac{1}{Snm}\right)^{\frac{2\alpha}{2\alpha+1}}$$

- Sparse vs. dense

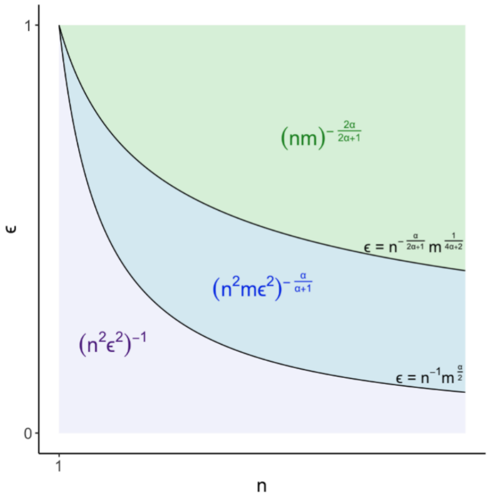
$$\left(\frac{1}{Sn^2 m\varepsilon^2}\right)^{\frac{2\alpha}{2\alpha+2}} \vee \left(\frac{1}{Snm}\right)^{\frac{2\alpha}{2\alpha+1}} \quad \text{vs.} \quad \frac{1}{Sn^2\varepsilon^2} \vee \frac{1}{Sn}$$

To understand the phase transition, let's focus on a simple case that

$$S = 1.$$


When $0 \lesssim m \lesssim n^{\frac{1}{2\alpha}}$
When $n^{\frac{1}{2\alpha}} \lesssim m \lesssim n^{\frac{1}{\alpha}}$
When $m \gtrsim n^{\frac{1}{\alpha}}$
$(nm)^{-\frac{2\alpha}{2\alpha+1}}$
$\epsilon = n^{-\frac{\alpha}{2\alpha+1}} m^{\frac{1}{4\alpha+2}}$
$(n^2 m\epsilon^2)^{-\frac{\alpha}{\alpha+1}}$
$(n^2\epsilon^2)^{-1}$
$\epsilon = n^{-1} m^{\frac{\alpha}{2}}$
n^{-1}
$\epsilon = n^{\frac{-\alpha+1}{2\alpha}} m^{-\frac{1}{2}}$
$\epsilon = n^{-\frac{1}{2}}$
ϵ
n
1
0

When $0 \lesssim m \lesssim n^{\frac{1}{2\alpha}}$



$$\left\{\left(X_j^{(s,i)},\ Y_j^{(s,i)}\right)\right\}_{i,j,s=1}^{n,m,S}$$

$$Y_j^{(s,i)} = \mu^*(X_j^{(s,i)}) + U^{(s,i)}(X_j^{(s,i)}) + \xi_{s,ij}$$

In a nutshell, the algorithm we adopt is Gaussian perturbed stochastic gradient descent based on basis expansion, with r basis functions.

A non-private estimator is

$$\widehat{\mu}(\cdot) = \Phi_r^\top(\cdot)a,$$

with

$$\widehat{a} = \underset{a \in \mathbb{R}^r}{\operatorname{argmin}} \left[\frac{1}{nmS} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{s=1}^{S} \{Y_j^{(s,i)} - a^\top \Phi_r(X_j^{(s,i)})\}^2 \right]$$

and $\Phi_r(\cdot) = (\phi_1(\cdot), \ldots, \phi_r(\cdot))^\top$ being the leading r basis functions.

A private SGD estimator is obtained by updating gradients with added Gaussian noise.

In a nutshell, the algorithm we adopt is Gaussian perturbed stochastic gradient descent based on basis expansion, with r basis functions.

A non-private estimator is

$$\widehat{\mu}(\cdot) = \Phi_r^\top(\cdot)a,$$

with

$$\widehat{a} = \operatorname*{argmin}_{a \in \mathbb{R}^r} \left[\frac{1}{nmS} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{s=1}^{S} \{ Y_j^{(s,i)} - a^\top \Phi_r(X_j^{(s,i)}) \}^2 \right]$$

and $\Phi_r(\cdot) = (\phi_1(\cdot), \ldots, \phi_r(\cdot))^\top$ being the leading r basis functions.

A private SGD estimator is obtained by updating gradients with added Gaussian noise.

In a nutshell, the algorithm we adopt is Gaussian perturbed stochastic gradient descent based on basis expansion, with r basis functions.

A non-private estimator is

$$\widehat{\mu}(\cdot) = \Phi_r^\top(\cdot)a,$$

with

$$\widehat{a} = \operatorname*{argmin}_{a \in \mathbb{R}^r} \left[\frac{1}{nmS} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{s=1}^{S} \{ Y_j^{(s,i)} - a^\top \Phi_r(X_j^{(s,i)}) \}^2 \right]$$

and $\Phi_r(\cdot) = (\phi_1(\cdot), \ldots, \phi_r(\cdot))^\top$ being the leading r basis functions.

A private SGD estimator is obtained by updating gradients with added Gaussian noise.

The ℓ_2-sensitivity of a function $f : \mathcal{D} \to \mathbb{R}^r$ is defined as

$$\Delta_2(f) = \sup_{D \sim D'} \|f(D) - f(D')\|_2.$$

The standard Gaussian mechanism states that

$$M(D) = f(D) + Z, \quad \text{with } Z \sim \mathcal{N}(0, 2(\Delta_2(f)/\varepsilon)^2 \log(1.25/\delta)),$$

satisfies (ε, δ)-DP.

Let $\Delta(f) = (\Delta f_1, \ldots, \Delta f_r)^\top$ with $\Delta f_\ell = \sup_{D \sim D'} |f_\ell(D) - f_\ell(D')|, \ell \in \{1, \ldots, r\}$. We propose the anisotropic Gaussian mechanism that

$$M(D) = f(D) + Z, \quad \text{with } Z \sim \mathcal{N}(0, \Sigma),$$

where $\Sigma = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_r^2\}$ and $\sigma_\ell^2 = 4 \log(2/\delta) \Delta f_\ell \|\Delta(f)\|_1 / \varepsilon^2$. We have that $M(\cdot)$ is (ε, δ)-DP.

The ℓ_2-sensitivity of a function $f : \mathcal{D} \to \mathbb{R}^r$ is defined as

$$\Delta_2(f) = \sup_{D \sim D'} \|f(D) - f(D')\|_2.$$

The standard Gaussian mechanism states that

$$M(D) = f(D) + Z, \quad \text{with } Z \sim \mathcal{N}(0, 2(\Delta_2(f)/\varepsilon)^2 \log(1.25/\delta)),$$

satisfies (ε, δ)-DP.

Let $\Delta(f) = (\Delta f_1, \ldots, \Delta f_r)^\top$ with $\Delta f_\ell = \sup_{D \sim D'} |f_\ell(D) - f_\ell(D')|$, $\ell \in \{1, \ldots, r\}$. We propose the anisotropic Gaussian mechanism that

$$M(D) = f(D) + Z, \quad \text{with } Z \sim \mathcal{N}(0, \Sigma),$$

where $\Sigma = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_r^2\}$ and $\sigma_\ell^2 = 4\log(2/\delta)\Delta f_\ell \|\Delta(f)\|_1/\varepsilon^2$. We have that $M(\cdot)$ is (ε, δ)-DP.

The *de facto* upper bound is

$$\|\widetilde{\mu} - \mu^*\|_{L^2}^2 \lesssim \frac{r^2}{\sum_{s=1}^{S}(r^2 n_s \wedge r^2 n_s^2 \varepsilon_s^2 \wedge r n_s m \wedge n_s^2 m \varepsilon_s^2)} + r^{-2\alpha}.$$

Assuming homogeneity across servers, one can choose

$$r \asymp (Sn)^{\frac{1}{2\alpha}} \wedge (Snm)^{\frac{1}{2\alpha+1}} \wedge (Sn^2\varepsilon^2)^{\frac{1}{2\alpha}} \wedge (Sn^2 m\varepsilon^2)^{\frac{1}{2\alpha+2}}$$

and lead to

$$\|\widetilde{\mu} - \mu^*\|_{L^2}^2 \lesssim \frac{1}{Sn} \vee \left(\frac{1}{Snm}\right)^{\frac{2\alpha}{2\alpha+1}} \vee \frac{1}{Sn^2\varepsilon^2} \vee \left(\frac{1}{Sn^2 m\varepsilon^2}\right)^{\frac{2\alpha}{2\alpha+2}}.$$

An intermediate result

$$\inf_{Q \in \mathcal{Q}_{\varepsilon,\delta,T}} \inf_{\widetilde{\mu}} \sup_{P_X, P_Y} \mathbb{E}_{P_X, P_Y, Q} \left\| \widetilde{\mu} - \mu^* \right\|_{L_2}^2$$
$$\gtrsim \frac{1}{ST(b \wedge b^2 \varepsilon^2)} \vee \frac{r_0^2}{ST(b^2 m \varepsilon^2 \wedge r_0 bm)},$$

where r_0 is the solution to

$$r^{2\alpha+2} = ST(b^2 m \varepsilon^2 \wedge rbm).$$

Proof ingredients

- Solve optimal b - the batch size.
- Case 1, constant functions for the mean function and the noise functions.
- Case 2, r-dimensional vector estimation.
- The van-Trees inequality.

An intermediate result

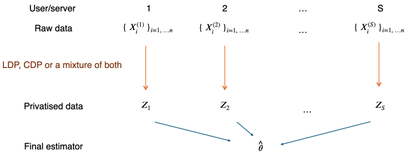
$$\inf_{Q\in\mathcal{Q}_{\varepsilon,\delta,T}} \inf_{\widetilde{\mu}} \sup_{P_X,P_Y} \mathbb{E}_{P_X,P_Y,Q}\left\|\widetilde{\mu}-\mu^*\right\|_{L_2}^2 \gtrsim \frac{1}{ST(b\wedge b^2\varepsilon^2)} \vee \frac{r_0^2}{ST(b^2 m\varepsilon^2 \wedge r_0 bm)},$$

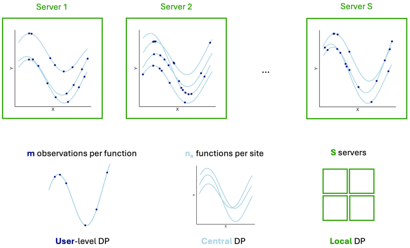
where r_0 is the solution to

$$r^{2\alpha+2} = ST(b^2 m\varepsilon^2 \wedge rbm).$$

Proof ingredients

- Solve optimal b - the batch size.
- Case 1, constant functions for the mean function and the noise functions.
- Case 2, r-dimensional vector estimation.
- The van-Trees inequality.





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