Decision Tree for Locally Private Estimation with Public Data

Oct 10, 2023

Yuheng Ma, Han Zhang, Yuchao Cai, Hanfang Yang. NeurIPS 2023.

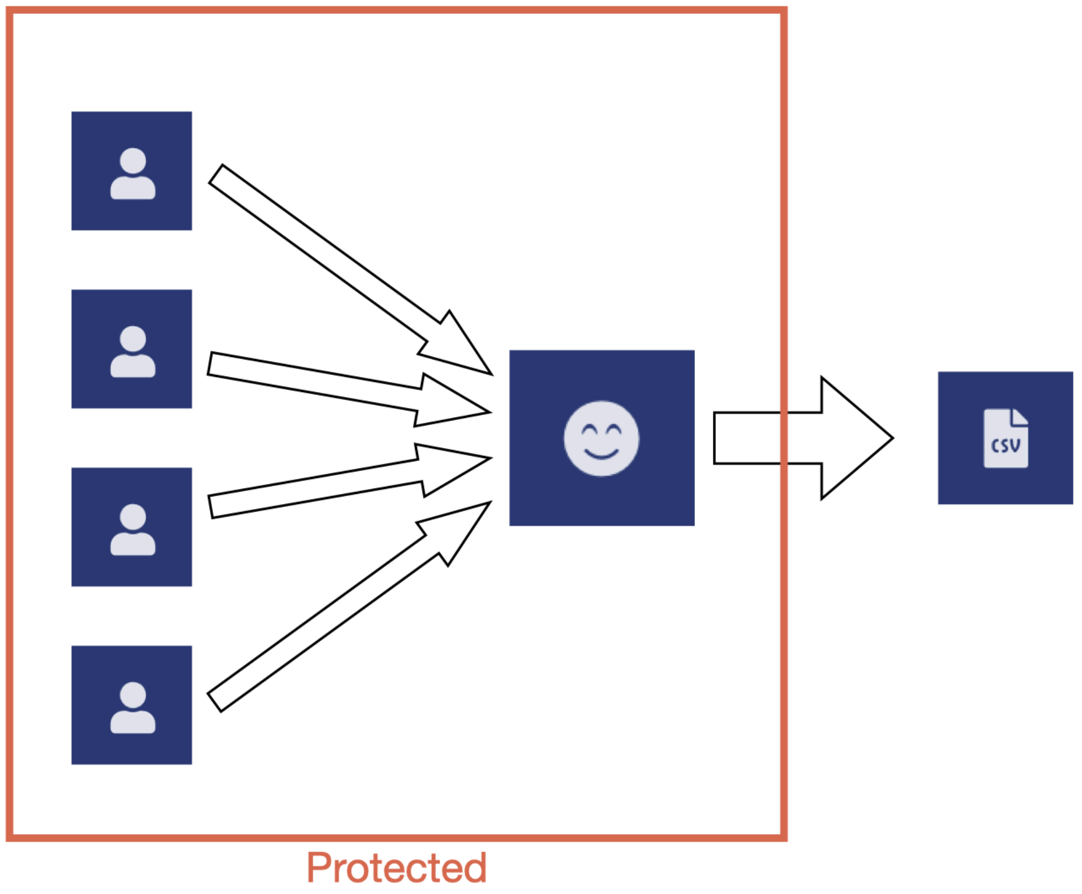
Motivation

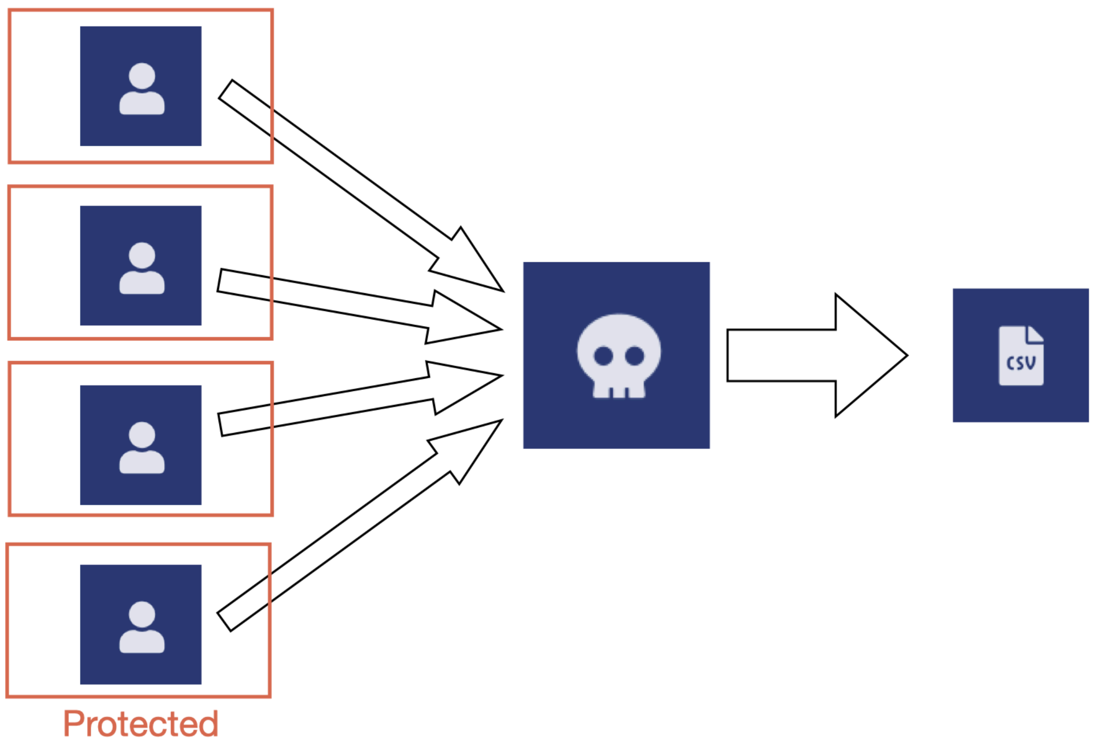
Differential Privacy (DP)

- “the” mathematical definition of privacy leakage
- Involving noise adding to
    - Gradient: gradient perturbation
    - Loss function: objective perturbation
    - Essential statistics: output perturbation
- Generally, more privacy, more noise, less accuracy

Local DP

- Opposite to central DP, local DP pose more strict privacy constraints
    - Trusted curator (for instance, group leader)
    - Untrusted curator (for instance, large tech company)





Fundamental Problem of LDP

Slow convergence

- More amount of noise added in LDP, see [1][2] for instance

Resource demanding

- Computation & memory & communication capacity of terminal machine

Basic operations are prohibited

- PCA, SVD, standardization, decision tree partition

[1] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. Journal of the American Statistical Association, 113(521): 182–201, 2018.
[2] Cai T T, Wang Y, Zhang L. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy[J]. The Annals of Statistics, 2021, 49(5): 2825-2850.

Public Data Helps!

Slow convergence

- Improve utility by public pretraining; public gradient preconditioning [1][2]

Resource demanding

- Allow designing of non-interactive methods [3]

Basic operations are prohibited

- Public standardization [4], covariance matrix estimation [3]
- Decision tree partition (ours)

[1] Da Yu et al. Differentially private fine-tuning of language models. ICLR 2022.
[2] Da Yu et al. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. ICLR 2021.
[3] Di Wang, Lijie Hu, Huanyu Zhang, Marco Gaboardi, and Jinhui Xu. Generalized linear models in non-interactive local differential privacy with public data. Journal of Machine Learning Research, 24(132):1–57, 2023.
[4] Bie A, Kamath G, Singhal V. Private estimation with public data. NeurIPS 2022.

Why is decision tree important?

- Previous work on nonparametric regression [1][2][3] show the theoretical superiority of histogram over other attempts
- Empirical evidence: histogram is inefficient!
    - Curse of dimensionality
    - Effected by marginal (density variation & useless feature)
    - Ignore information in data
- Decision tree has: higher accuracy than histogram; interpretability; efficiency; stability, extensiveness to multiple feature types; resistance to the curse of dimensionality
- We can not do decision tree partition in LDP without public data!

[1] Berrett T B, Györfi L, Walk H. Strongly universally consistent nonparametric regression and classification with privatised data[J]. Electronic Journal of Statistics, 2021, 15: 2430-2453.
[2] Györfi L, Kroll M. On rate optimal private regression under local differential privacy[J]. arXiv preprint arXiv:2206.00114, 2022.
[3] Farokhi F. Deconvoluting kernel density estimation and regression for locally differentially private data. Scientific Reports, 2020, 10(1): 21361.

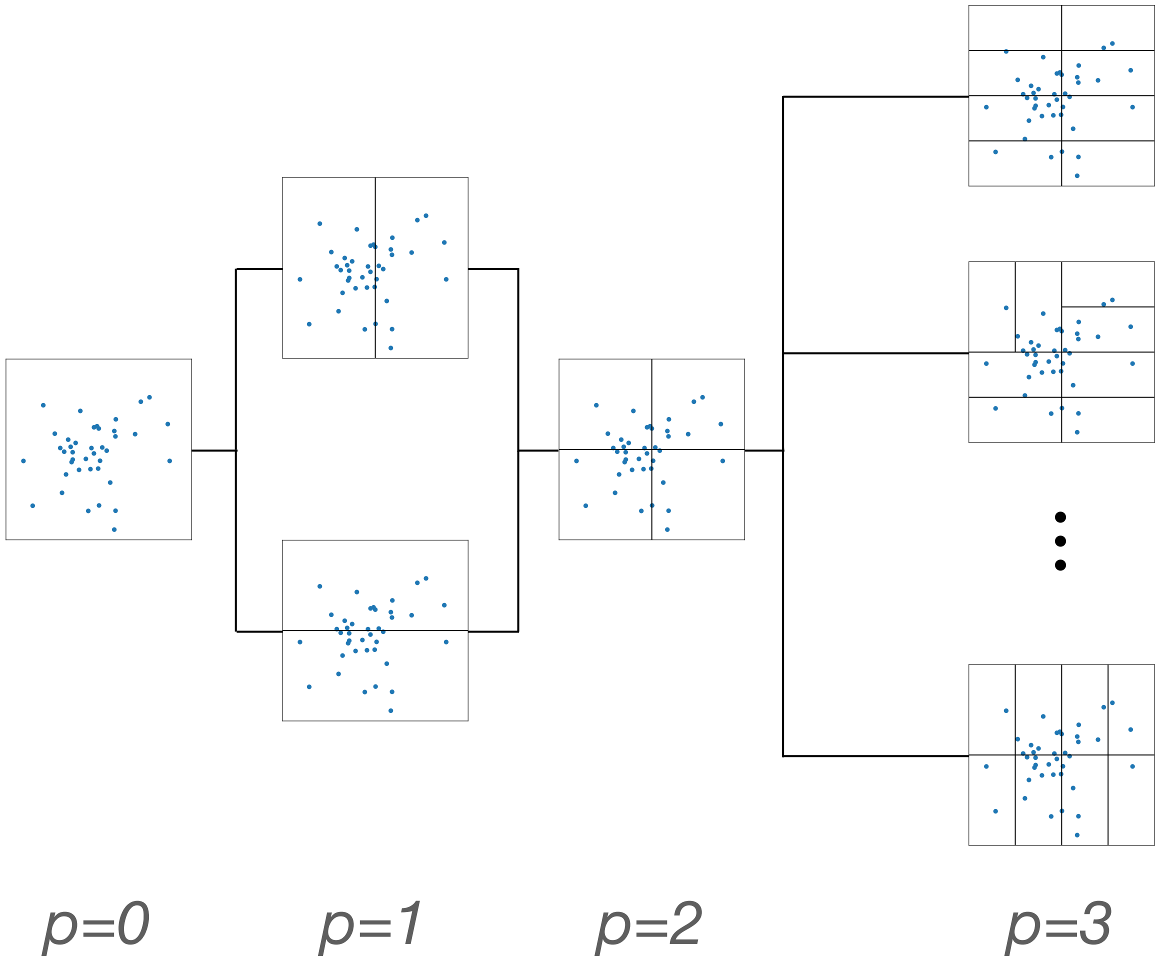
Methodology

Overview

- Given both public and private datasets, we:
    - first create partition on public data
    - then estimate privately on private data
- In doing so, the estimator
    - remains rate optimal in a milder assumption
    - is free of range parameter
    - has significant better empirical performance

Max-edge Partition Rule

- For each grid, the partition rule selects the midpoint of the longest edges that achieves the largest variance reduction
- This procedure continues until there are not enough samples contained in any leaf node, or the depth of the tree reaches its limit
- No private concern



Privacy for Partition Estimation

- Given partition π, let $U_i \in \{0,1\}^{|\mathcal{I}|}$ and $U_i^j = 1\{X_i \in A_j\}$

$$f_\pi(x) = \sum_{j \in \mathcal{I}} 1_{A_j} \frac{\sum_{i=1}^n Y_i \cdot U_i^j}{\sum_{i=1}^n U_i^j}$$

$\approx \int_{A_j} f^*(x) dP(x')$ joint estimation

$\approx \int_{A_j} dP(x')$ marginal estimation

conditional distribution estimation: decision tree

$$\tilde{f}_\pi(x) = \sum_{j \in \mathcal{I}} 1_{A_j} \frac{\sum_{i=1}^n \tilde{Y}_i \cdot \tilde{U}_i^j}{\sum_{i=1}^n \tilde{U}_i^j}$$

private joint estimation

private marginal estimation

private conditional distribution estimation: private decision tree

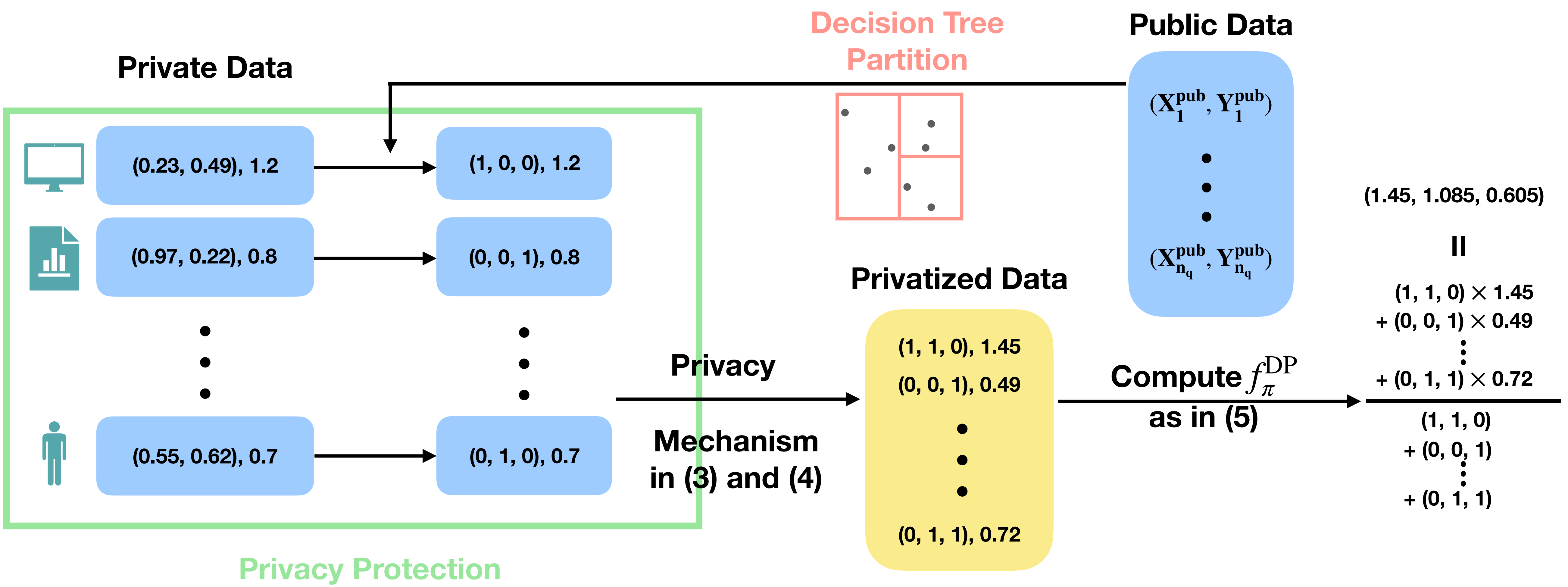
Perturbation Mechanism

- Protect Y by Laplacian noise i.e. $\tilde{Y}_i = Y_i + \frac{4M}{\varepsilon}\xi_i$

- Protect U by random response, i.e.

$$\tilde{U}_i^j = \begin{cases} U_i^j - \frac{1}{1+e^{\varepsilon/4}} & \text{with probability } \frac{e^{\varepsilon/4}}{1+e^{\varepsilon/4}} \\ 1 - U_i^j - \frac{1}{1+e^{\varepsilon/4}} & \text{with probability } \frac{1}{1+e^{\varepsilon/4}} \end{cases}.$$

Locally Private Decision Tree



Theoretical Results

Utility

Assumption 3.2. Let $\alpha \in (0,1]$. Assume the regression function $f : \mathcal{X} \to \mathbb{R}$ is α-Hölder continuous, i.e. there exists a constant $c_L > 0$ such that for all $x_1, x_2 \in \mathcal{X}$, $|f(x_1) - f(x_2)| \leq c_L \|x_1 - x_2\|^\alpha$. Also, assume that the density function of P is upper bounded, i.e. $p(x) \leq \overline{c}$ for some $\overline{c} > 0$.

Assumption 3.3. We assume that there exists some constant $\tau > 1$ such that for all cells $A \in \pi$, there holds $\tau^{-1} \int_A d\mathrm{Q}_X(x) \leq \int_A d\mathrm{P}_X(x) \leq \tau \int_A d\mathrm{Q}_X(x)$.

Theorem 3.4. *Let f_π^{DP} be the LPDT estimator in Algorithm 1. Suppose Assumption 3.2 and 3.3 hold. Then, for $n_q \gtrsim n^{\frac{d}{2\alpha+2d}}$, if we set $p \asymp \log n\varepsilon^2$ and $n_l \asymp n_q/2^p$, there holds*

$$\mathcal{R}_{L,\mathrm{P}}(f_\pi^{\mathrm{DP}}) - \mathcal{R}_{L,\mathrm{P}}^* \lesssim \left(\frac{\log n}{n\varepsilon^2}\right)^{\frac{\alpha}{\alpha+d} \wedge \frac{1}{3}}$$

with probability $1 - 2/n_q^2 - 5/n^2$ with respect to $\mathrm{P}^n \otimes \mathrm{Q}^{n_q} \otimes \mathrm{R}^n$ where R^n is the joint distribution of privacy mechanisms in (3) and (4).

Privacy

Theorem 3.1. *Let* $\pi = \{A_j\}_{j\in\mathcal{I}}$ *be any partition of* $\mathcal{X}$ *with* $\cup_{j\in\mathcal{I}} A_j = \mathcal{X}$ *and* $A_i \cap A_j = \emptyset$, $i \neq j$. *Then the privacy mechanism* $\mathrm{R}(\tilde{U}, \tilde{Y} | X, Y)$ *defined in* (3) *and* (4) *is* ε*-LDP. Consequently, the LPDT estimator* f_π^{DP} *in Algorithm* 1 *is* ε*-LDP.*

Complexity

Table 1: Comparison of complexities of LDP regression methods.

	LPDT	PHIST [9]	DECONV [29]
Training Time Complexity	$\mathcal{O}(n \log n\varepsilon^2 + n_q d \log n\varepsilon^2)$	$\mathcal{O}(nd \log n\varepsilon^2)$	-
Testing Time Complexity	$\mathcal{O}(\log n\varepsilon^2)$	$\mathcal{O}(\log n\varepsilon^2)$	$\mathcal{O}(nd)$
Space Complexity	$\mathcal{O}\left(\left(n\varepsilon^2/\log n\right)^{\frac{d}{2\alpha+2d}}\right)$	$\mathcal{O}\left(\left(n\varepsilon^2/\log n\right)^{\frac{d}{2\alpha+2d}}\right)$	$\mathcal{O}(nd)$

Experiments

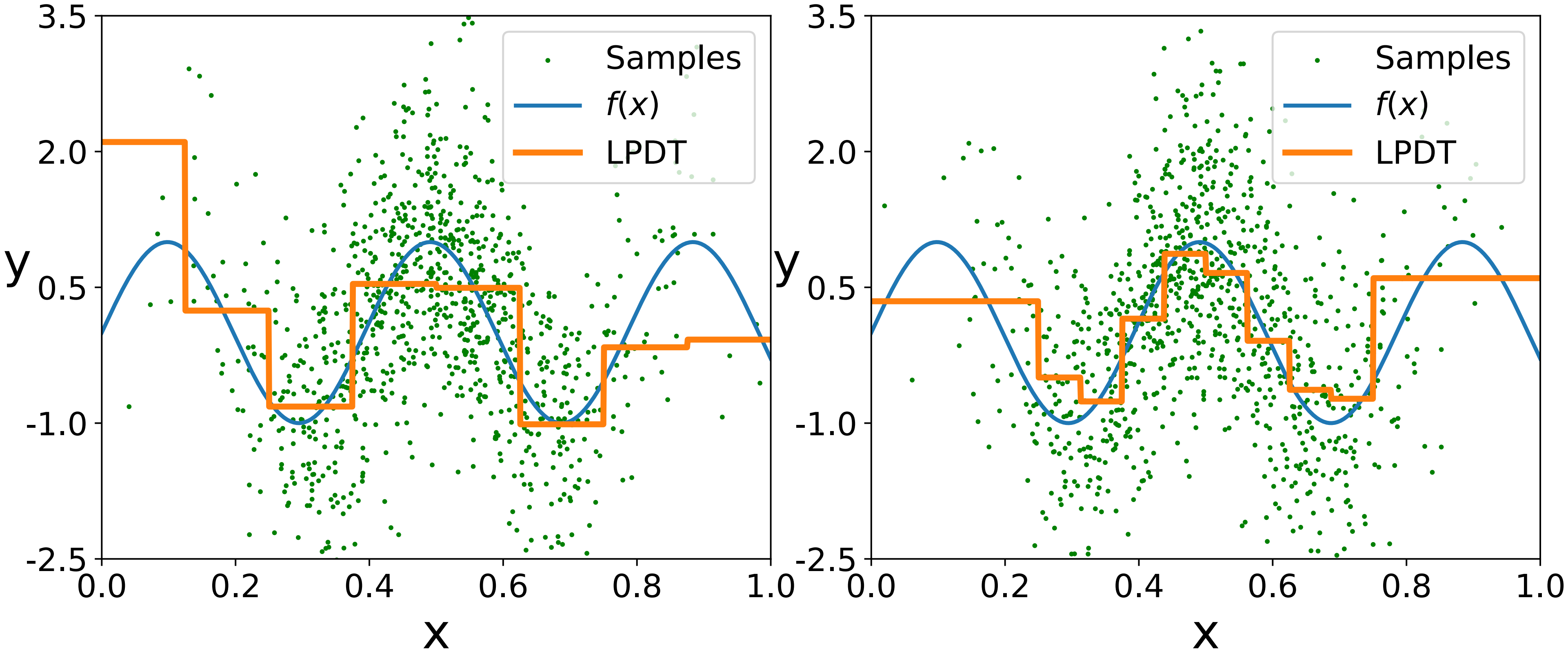
Settings

- Consider $\varepsilon \in [0.5,8]$
- Consider partition rule of CART
- Parameter selection by cross validation in a non-private way, see discussion in [1][2][3].
- Comparison methods: DECONV [4] (deconvolution based), PHIST & APHIST [5][6] (histogram based)

[1] Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. ICLR 2021.
[2] Andrew Lowy, Zeman Li, Tianjian Huang, and Meisam Razaviyayn. Optimal differentially private learning with public data. arXiv preprint arXiv:2306.15056, 2023.
[3] Yuheng Ma, Hanfang Yang. Optimal Locally Private Nonparameteric Classification with Public Data.
[4] Farokhi F. Deconvoluting kernel density estimation and regression for locally differentially private data. Scientific Reports, 2020, 10(1): 21361.
[5] Berrett T B, Györfi L, Walk H. Strongly universally consistent nonparametric regression and classification with privatised data[J]. Electronic Journal of Statistics, 2021, 15: 2430-2453.
[6] Györfi L, Kroll M. On rate optimal private regression under local differential privacy[J]. arXiv preprint arXiv:2206.00114, 2022.

Necessity of Public Data

- n = 6,000, $X \sim N(0.5, 0.16)$, $f^*(x) = \sin(16x) + \varepsilon$, without and with 1,000 public data.

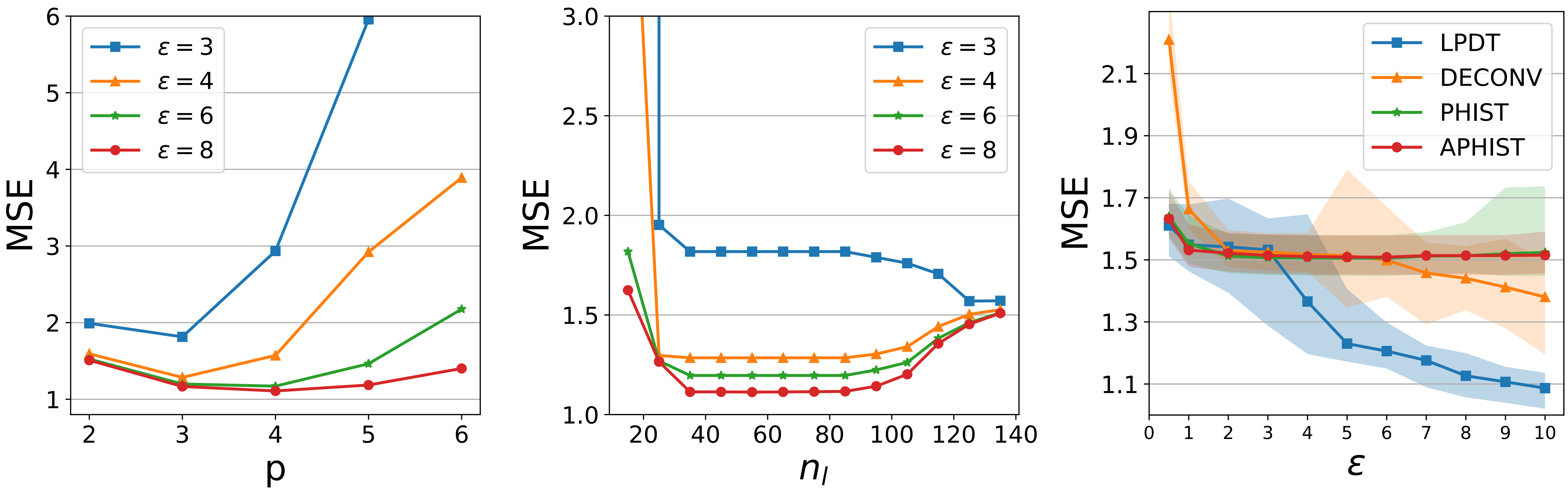


- The low-density regions can be identified and treated with larger cells automatically

Some Analysis

- Under the same distribution and other parameters fixed, examine influence of depth p and minimum leaf samples n_l. When facing higher levels of privacy demand, LPDT cuts down the number of grids to stabilize its estimation.

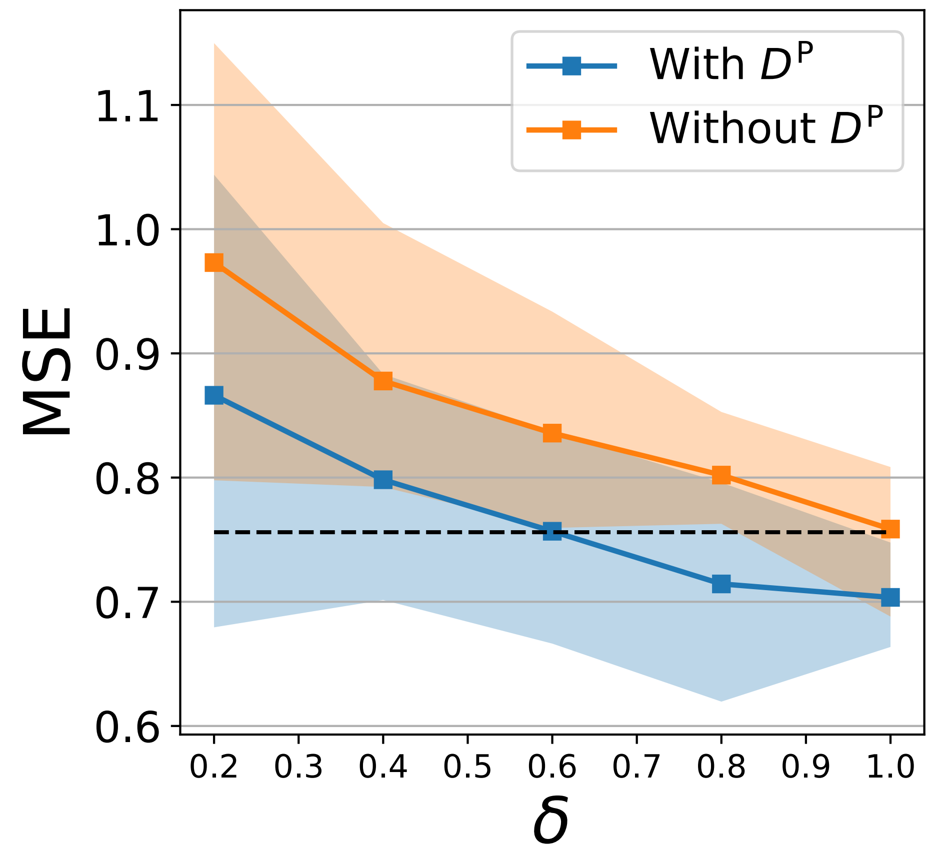
- LPDT achieves best privacy-utility trade off



Identically Distributed Public Data

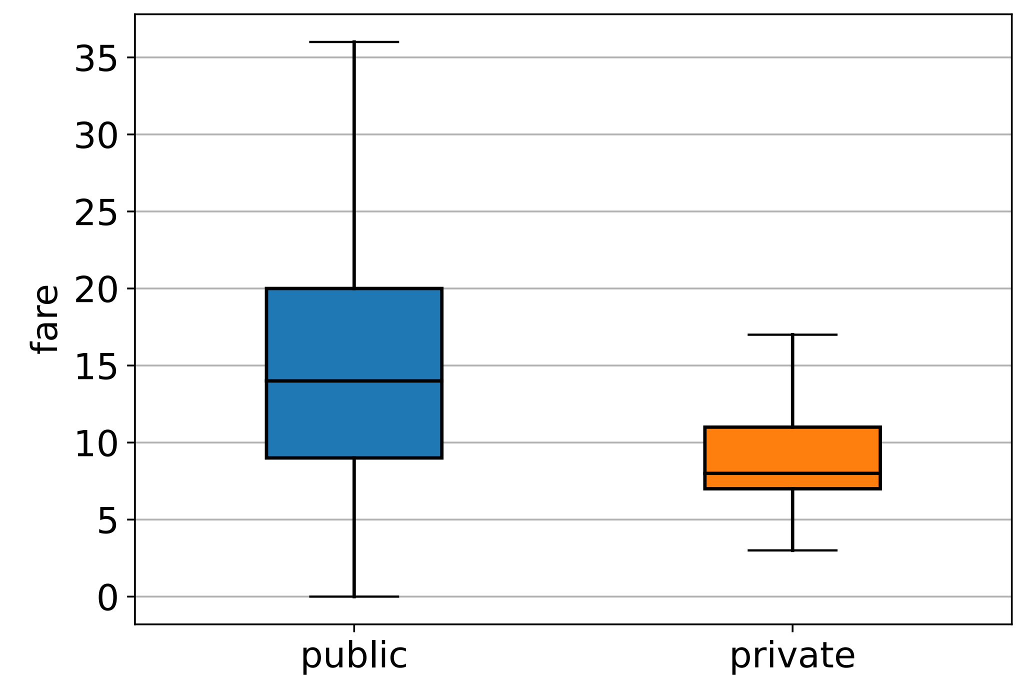
- Over 14 datasets from UCI repository, LPDT outperforms.
- 100 public data and a fraction δ of 1100 private data of wine dataset. The utility increase brought by public data is significance.

	DT	$\varepsilon = 2$					$\varepsilon = 6$				
		LPDT-M	LPDT-V	APHIST	PHIST	DECONV	LPDT-M	LPDT-V	APHIST	PHIST	DECONV
ABA	5.67e+0	**1.01e+1**	1.01e+1	1.89e+1	1.06e+1	1.01e+7	8.38e+0*	**7.34e+0***	2.05e+1	1.05e+1	1.09e+1
AIR	2.26e+1	4.80e+1*	**4.69e+1***	1.31e+3	6.80e+1	3.00e+2	4.49e+1*	**3.60e+1***	1.60e+3	4.98e+1	4.72e+1
ALG	2.12e-2	2.57e-1	**2.43e-1**	2.52e-1	2.52e-1	9.26e+4	**2.44e-1**	2.46e-1	2.63e-1	2.47e-1	3.14e-1
AQU	1.92e+0	2.99e+0*	2.99e+0*	4.01e+0	**2.93e+0***	5.74e+3	2.73e+0*	**2.67e+0***	4.75e+0	2.83e+0	2.96e+0
BUI	1.75e+5	**1.50e+6***	1.64e+6*	-	-	1.20e+9	1.44e+6*	**1.31e+6***	-	-	2.04e+7
CBM	4.08e-27	2.12e+0*	**1.65e+0***	9.53e+0	6.97e+0	2.37e+3	7.62e-1*	**1.23e-1***	4.94e+0	3.21e+0	1.23e+5
CCP	2.19e+1	1.50e+2*	**1.06e+2***	2.07e+4	3.64e+2	3.03e+2	8.42e+1*	**5.18e+1***	2.24e+4	3.28e+2	2.56e+2
CON	9.38e+1	2.94e+2*	**2.89e+2***	3.81e+2	3.00e+2	2.24e+7	2.44e+2*	**2.13e+2***	4.16e+2	2.96e+2	3.13e+2
CPU	2.15e+1	3.41e+2	**9.00e+1***	9.26e+2	3.42e+2	2.15e+5	3.02e+2*	**6.15e+1***	9.98e+2	3.40e+2	3.98e+2
FIS	1.07e+0	2.15e+0*	**2.14e+0***	3.14e+0	2.22e+0	3.47e+3	**1.65e+0***	1.76e+0*	3.60e+0	2.16e+0	2.21e+0
HOU	2.11e+1	**8.10e+1***	8.22e+1*	1.06e+2	8.52e+1	1.92e+4	7.43e+1*	**7.10e+1***	1.23e+2	8.21e+1	2.44e+2
MUS	3.00e+2	3.47e+2*	**3.46e+2***	-	-	9.50e+3	**3.27e+2***	3.27e+2*	-	-	8.09e+3
RED	4.76e-1	7.08e-1*	**7.03e-1***	3.18e+0	7.57e-1	1.23e+8	6.75e-1*	**6.12e-1***	3.80e+0	7.12e-1	8.66e-1
WHI	5.77e-1	8.30e-1	8.42e-1	4.01e+0	**8.15e-1**	1.64e+7	7.03e-1*	**6.61e-1***	4.45e+0	8.03e-1	1.47e+0



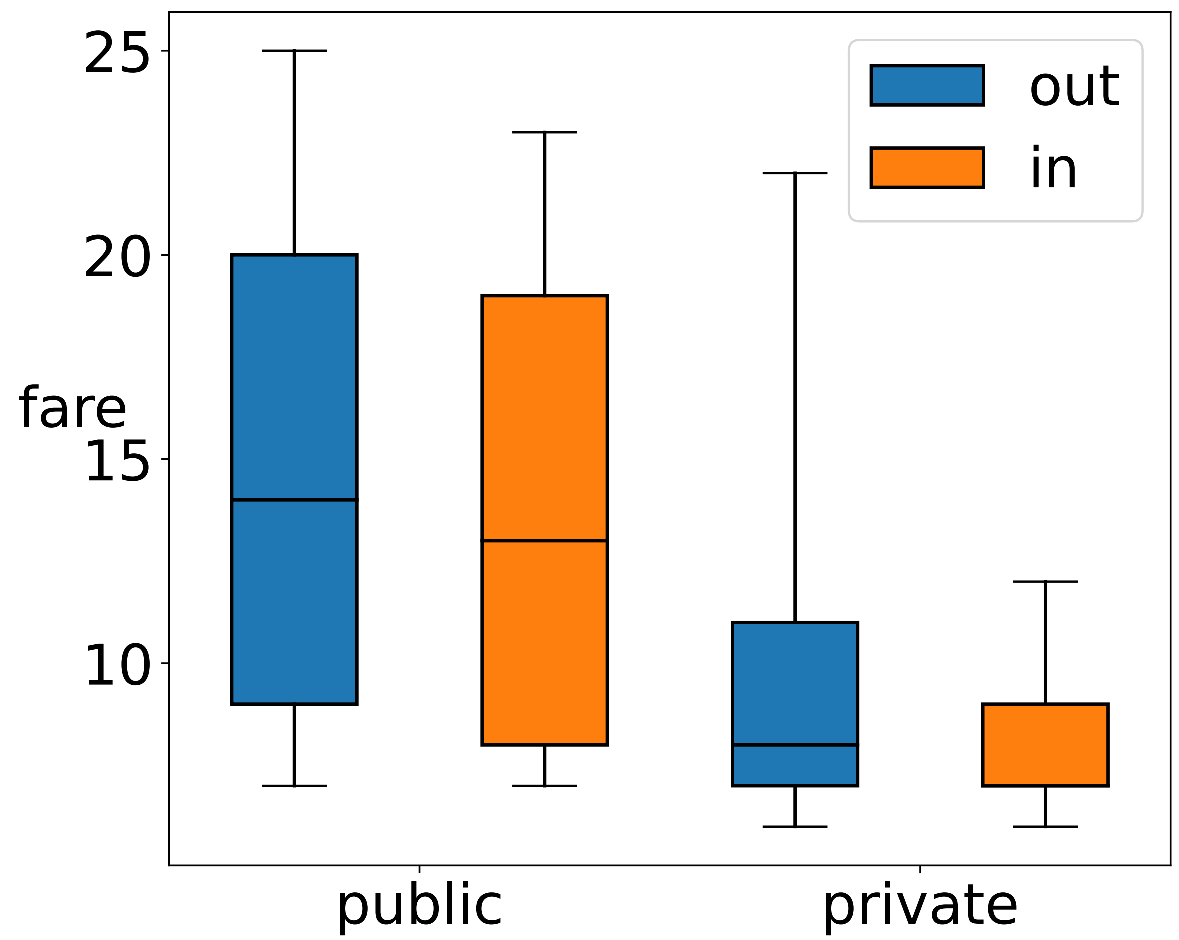
Non-Identically Distributed Public Data

- Taxi trips in Chicago
- Fare ~ time, distance, start/end location, company, paying method. 101 features in total.
- Public: PR card, 24,000 instances
- Private: credit card, 2,100,000 instances
- The distributions are non-identical



How does public data work?

- First split features: whether drop off in district 32?
- Similar pattern, distinct distribution.



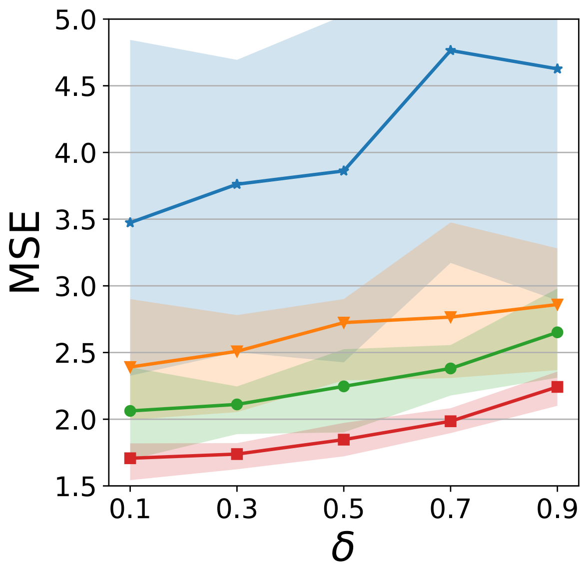
Performance

- With both public and private data, LPDT outperforms with mild privacy constraint.

- Replace a fraction δ of public data by private data. Similar public and private data is better.

Table 3: Average MSE and standard deviation over Chicago taxi data.

DT		LPDT-M						PHIST		APHIST	
Public	Private	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 6$	$\varepsilon = 8$	$\varepsilon = 2$	$\varepsilon = 8$	$\varepsilon = 2$	$\varepsilon = 8$
3.71	0.80	113.45 (14.23)	15.74 (2.20)	4.89 (0.54)	3.35 (0.33)	2.86 (0.10)	2.70 (0.10)	24.72 (0.02)	17.22 (0.00)	38.22 (0.01)	35.5 (0.01)



Code

Q&A