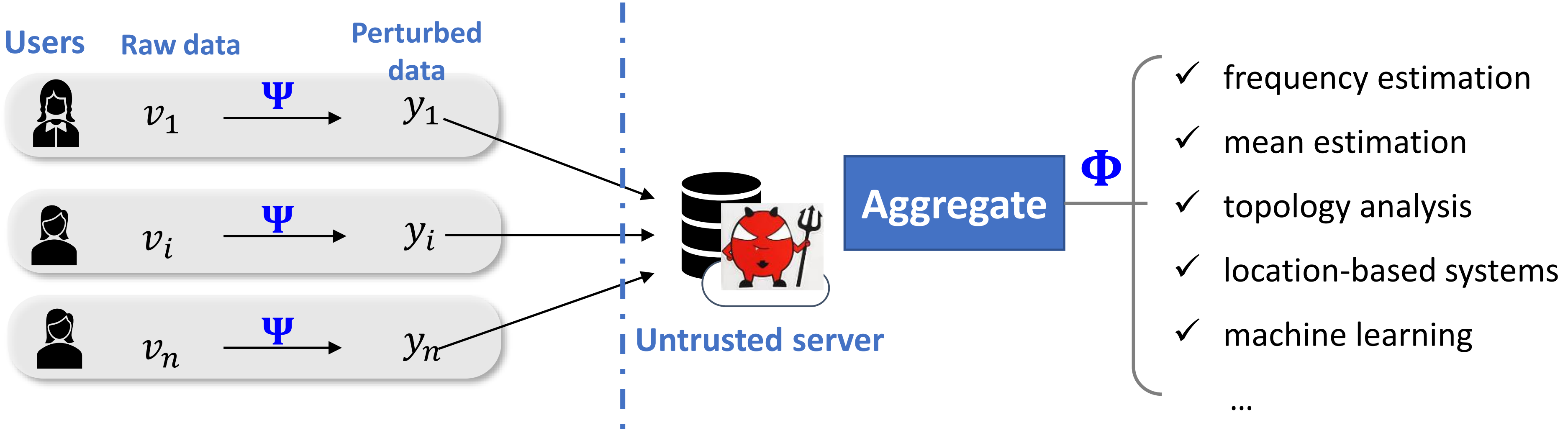
Revisiting EM-based Locally Differentially Private Protocols

Yutong Ye, Tianhao Wang, Min Zhang, Dengguo Feng

Local Differential Privacy [Duchi et al, FOCS'13]

- **Local Differential Privacy (LDP)** is a typical **locally** private data collection model



◆ A mining task under LDP can be formalized as an LDP protocol consisting of a pair of algorithms <**Ψ**,**Φ**>, where **Ψ** is a perturbation algorithm and **Φ** is an aggregation algorithm to extract useful knowledge.

Definition 1: A randomized algorithm **Ψ** satisfies **ε**-local differential privacy, iff for any two inputs $\boldsymbol{v}$ and $\boldsymbol{v}'$ and for any output $\boldsymbol{y}$ of **Ψ**,

$$\Pr[\boldsymbol{\Psi}(\boldsymbol{v}) = \boldsymbol{y}] \leq e^{\varepsilon} \cdot \Pr[\boldsymbol{\Psi}(\boldsymbol{v}') = \boldsymbol{y}]$$

ε ↓ , private ↑ , utility ↓

Local Differential Privacy

- **Fundamental Tasks**
  - Category data: **Frequency estimation, Heavy hitters mining**
    OLH, GRR [Usenix security' 17] RAPPOR [CCS' 14]
  - Numerical data: **Mean estimation, Density estimation**
    SW [Sigmod'20] PM [ICDE'19]

- **Local Differential Privacy is deployed in**
  - Apple iOS/macOS, to collect typing statistics, types of photos at frequently visited locations
  - Google Chrome/Android, to collect browsing statistics
  - Amazon Echo, to collect frequency of voice command statistics
  - Microsoft Windows, to collect telemetry data

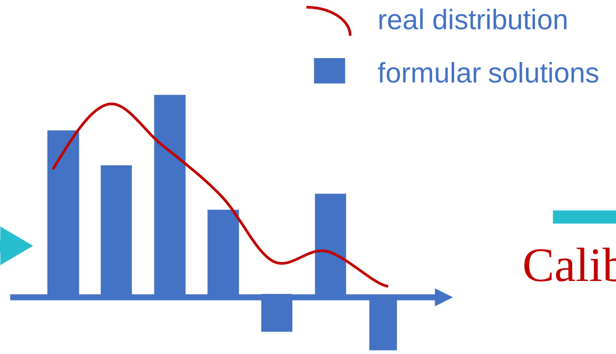
Aggregation methods (Φ)

Unbiased estimation + post-processing.

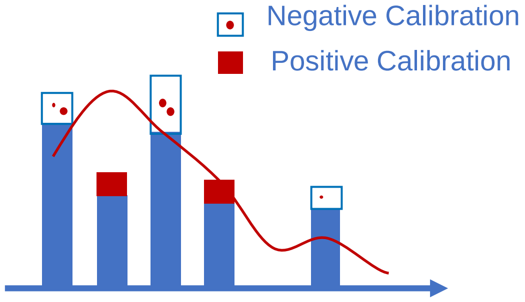
Estimation function is done independently for each value, and then Calibrate

Derive function, i.e.,

$$\hat{f}_\alpha = \frac{\sum_{i=1}^{n} 1_{\tilde{x}_i=\alpha}/n - q}{p-q}.$$



Calibration



Consistency-based calibration

Wang et al. [NDSS' 20]

Prior-knowledge-based calibration

Jia et al. [INFOCOM' 19]

Fang et al. [S&P' 23]

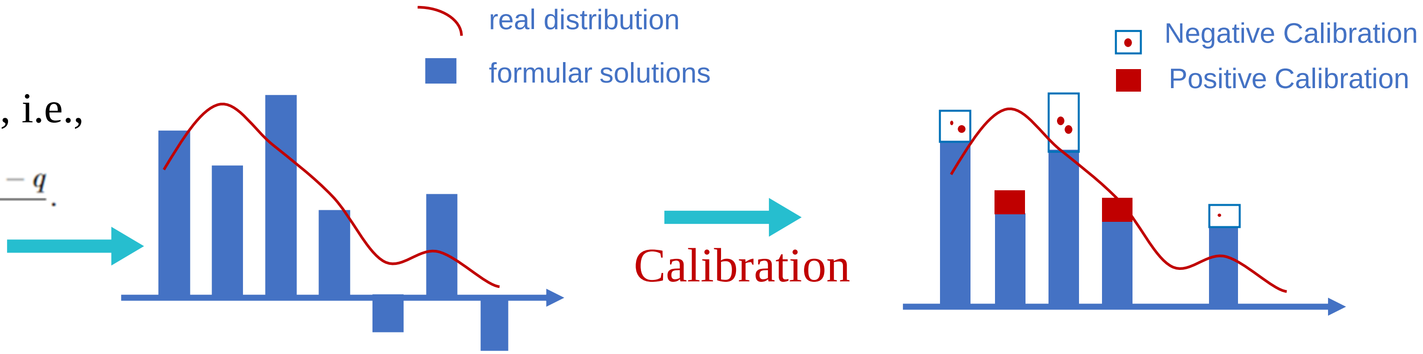
Aggregation methods (Φ)

Unbiased estimation + post-processing.

Estimation function is done independently for each value, and then Calibrate

Derive function, i.e.,

$$\hat{f}_\alpha = \frac{\sum_{i=1}^{n} 1_{\tilde{x}_i=\alpha}/n - q}{p - q}.$$



Consistency-based calibration

Wang et al. [NDSS' 20]

Prior-knowledge-based calibration

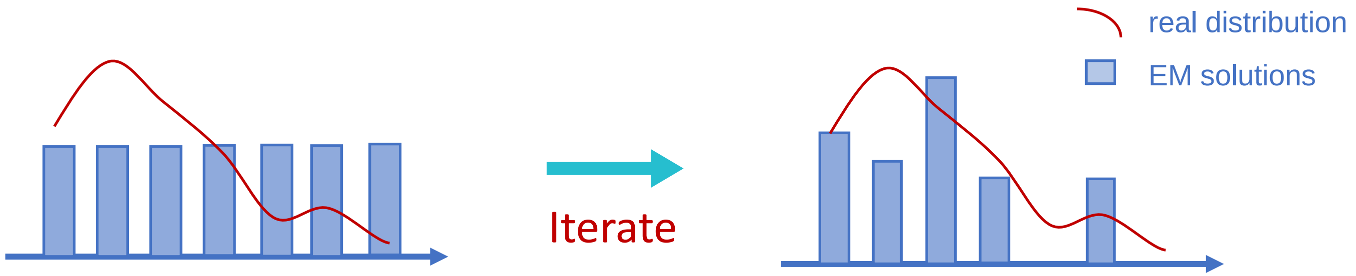
Jia et al. [INFOCOM' 19]

Fang et al. [S&P' 23]

In cases where the function is not easy to derive, or when a reasonable distribution is preferred

Expectation-Maximization (EM) based Maximal Likelihood Estimation(MLE)

Find a distribution that most likely leads to the observed perturbed data



EM-based MLE

Tu et al. [Pets' 19]

Li et al. [SIGMOD' 20]

Problems and Intuitions

Observation 1 (Fig 1) Pursuing a max likelihood value during EM process may lead to worse final error.

Observation 2 (Fig 2) More value need to estimate during EM → larger overall error.

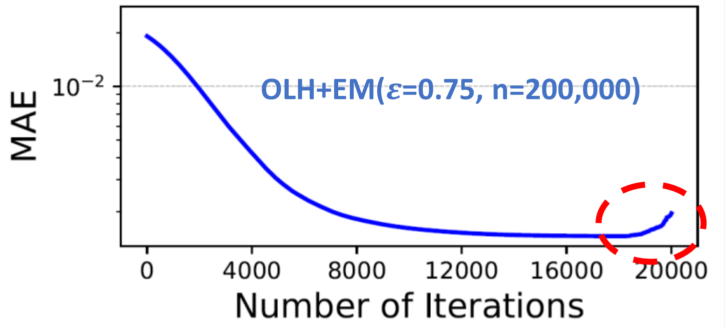


Fig 1. trace the MAE of EM iteration process

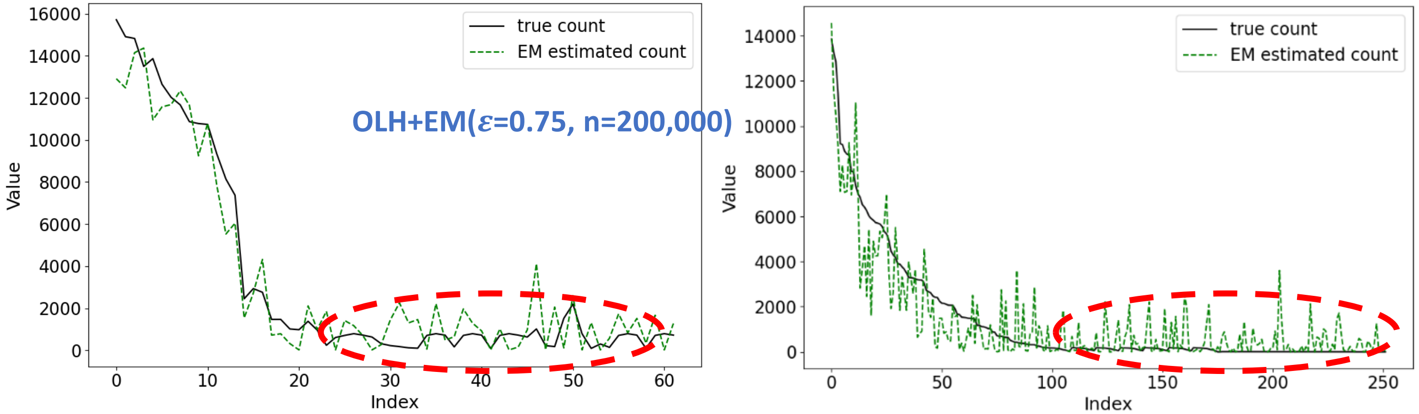


Fig 2. Compare the error between different distribution

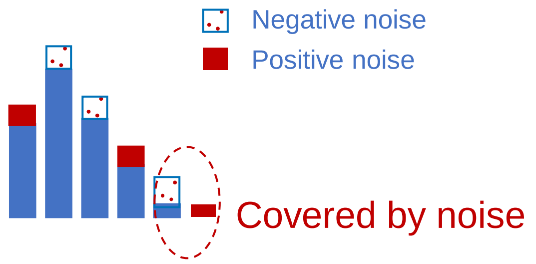
Problems EM-based MLE **is easy to overfit to the noise data**, especially when there is much noise.

Problems and Intuitions

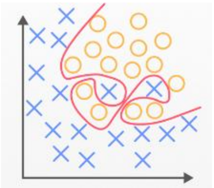
How can we overcome the overfitting issue of EM-based MLE to reduce the overall error?

Intuition

Noise overwhelms the small truth: LDP noise follows zero Gaussians, are likely to cover the small value



Many values → complexity fitting model → easily overfitting: In machine learning, regularization is a well-studied technique for overfitting issue, which penalizes small values in the model

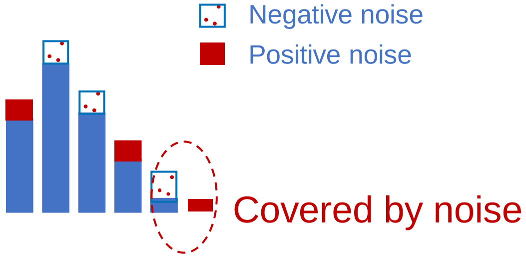


Problems and Intuitions

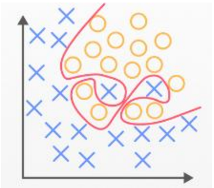
How can we overcome the overfitting issue of EM-based MLE to reduce the overall error?

Intuition

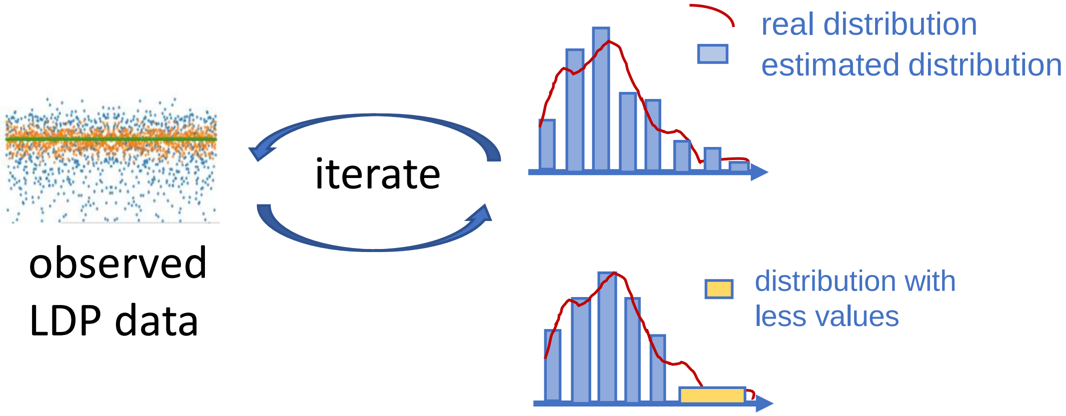
Noise overwhelms the small truth: LDP noise follows zero Gaussians, are likely to cover the small value



Many values → complexity fitting model → easily overfitting: In machine learning, regularization is a well-studied technique for overfitting issue, which penalizes small values in the model



Enhancing the EM-based estimation by reducing the number of values that are likely small across EM iteration process



Our Approach

Review of the EM Algorithm

✓ Iterative optimization technique used for parameter estimation, i.e., Gaussian Mixture model (GMM)

Goal: $$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} \left(\sum_{i}^{n} p(\tilde{x}|\mathbf{w}) \right)$$

E-step:

Calculate the likelihood given $\widehat{\mathbf{W}}^{(t)}$

M-step:

Update $\widehat{\mathbf{W}}^{(t+1)}$ that maximize the likelihood function $\mathcal{L}$

(by taking the derivative of $\mathcal{L}$)

Repeat EM step until converge

Our Approach

Review of the EM Algorithm

✓ Iterative optimization technique used for parameter estimation, i.e., Gaussian Mixture model (GMM)

Goal: $\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} \left(\sum_{i}^{n} p(\tilde{x}|\mathbf{w}) \right)$

E-step:

Calculate the likelihood given $\widehat{\mathbf{W}}^{(t)}$

M-step:

Update $\widehat{\mathbf{W}}^{(t+1)}$ that maximize the likelihood function $\mathcal{L}$

(by taking the derivative of $\mathcal{L}$)

Repeat EM step until converge

Generalize EM under different LDP $\boldsymbol{\Psi}$

Build a LDP mixture model for generalization

$$\phi(\tilde{x}; \mathbf{w}, \boldsymbol{\alpha}) = \sum_{k=1}^{K} \omega_k \mathrm{Pr}\left[\Psi_\varepsilon(\alpha_k) = \tilde{x}\right]$$

- $\mathrm{Pr}\left[\Psi_\varepsilon(\alpha_k) = \tilde{x}\right]$: PMF, also the transfer function of $\boldsymbol{\Psi}$
- ω_k : proportions or weights for each components.
- K : the number of values.

Goal: $\arg\max_{\hat{\mathbf{w}}} \mathcal{L}(\hat{\mathbf{w}}) \quad \text{s.t.} \quad \sum \hat{w}_i = 1,\ \hat{w}_i \geq 0$

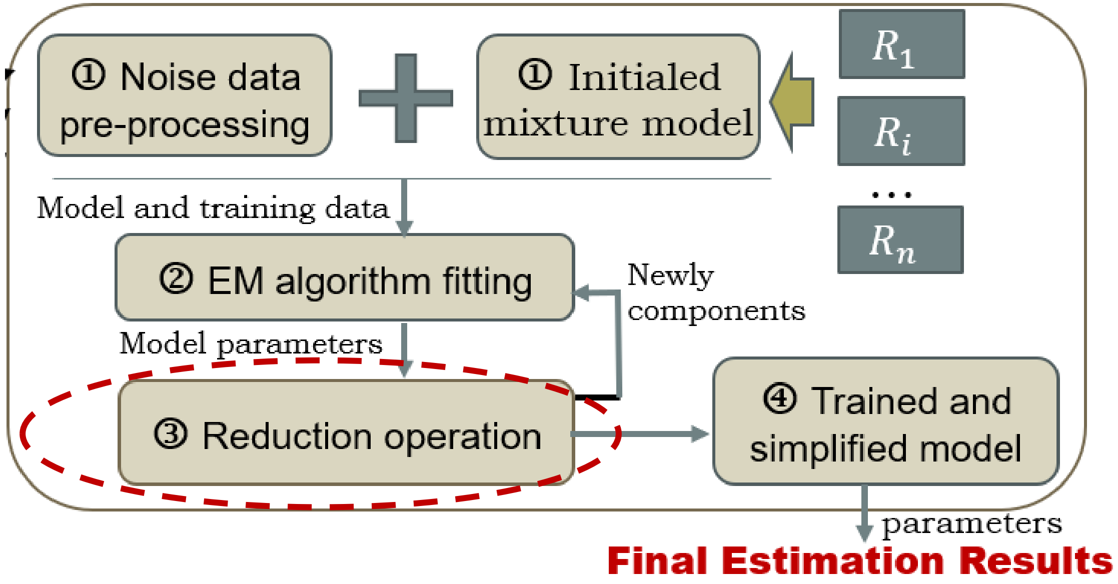
E-step:

$$\gamma_{ik} \leftarrow \frac{\hat{w}_k \mathrm{Pr}\left[\Psi_\varepsilon(\alpha_k) = \tilde{x}_i\right]}{\sum_{j=1}^{K} \hat{w}_j \mathrm{Pr}\left[\Psi_\varepsilon(\alpha_j) = \tilde{x}_i\right]}$$

M-step:

$$\hat{w}_k \leftarrow \frac{1}{n} \sum_{i=1}^{n} \gamma_{ik}$$

Our Approach (Mixture Reduction)



③ Reduction operation

merging step: $(w_{12}, \Psi_\varepsilon(\alpha_{12})) \leftarrow \{(w_1, \Psi_\varepsilon(\alpha_1)), (w_2, \Psi_\varepsilon(\alpha_2))\}$

$$w_{12} = w_1 + w_2$$

$$\Pr[\Psi_\varepsilon(\alpha_{12}) = \tilde{x}] = \sum_{i=1}^{2} \frac{w_i}{w_{12}} \Pr[\Psi_\varepsilon(\alpha_i) = \tilde{x}].$$

④ Judge the model and stop

BIC : Trade-off between model fit and complexity

$$\mathrm{BIC} = -2\log(\mathcal{L}) + K' \log(n)$$

Our Approach (Mixture Reduction)

Generalization: we demonstrate the application of our approach in various LDP tasks

TABLE I
SUMMARY OF METHODS IN EM-BASED MLE

Methods	Description	Pre-process	Probability mass or density function	Time complexity
GRR	FO in small K scenario	-	Equation (2)	$O(K^2 \log(K) I)$
OLH	FO in large K scenario	hash matching	$\frac{e^\varepsilon}{e^\varepsilon + K^* - 1}$ if hash matches	$O(nK \log(K) I)$
PM & SW	numerical FO and mean estimator	binning	Equation (7) and (14)	$O(K^2 \log(K) I)$
Laplace	numerical perturbation	binning	the pdf of Laplace distribution	$O(nK \log(K) I)$
Gaussian	(ε, δ)-LDP for high-dimensional data	binning	the pdf of Gaussian distribution	$O(nK \log(K) I)$
PCKV-PM	key-value data analysis	binning	joint pmf from the combination of PM and FOs	$O(Kd^2 \log(d) I)$

Our Approach (Mixture Reduction)

Generalization: we demonstrate the application of our approach in various LDP tasks

TABLE I
SUMMARY OF METHODS IN EM-BASED MLE

Methods	Description	Pre-process	Probability mass or density function	Time complexity
GRR	FO in small K scenario	-	Equation (2)	$O(K^2 \log(K) I)$
OLH	FO in large K scenario	hash matching	$\frac{e^{\varepsilon}}{e^{\varepsilon}+K^*-1}$ if hash matches	$O(nK \log(K) I)$
PM & SW	numerical FO and mean estimator	binning	Equation (7) and (14)	$O(K^2 \log(K) I)$
Laplace	numerical perturbation	binning	the pdf of Laplace distribution	$O(nK \log(K) I)$
Gaussian	(ε, δ)-LDP for high-dimensional data	binning	the pdf of Gaussian distribution	$O(nK \log(K) I)$
PCKV-PM	key-value data analysis	binning	joint pmf from the combination of PM and FOs	$O(Kd^2 \log(d) I)$

Accuracy analysis (Informal)

The MSE of our approach consists of two components: (1)the estimation error from the EM algorithm applied to the remaining values, and (2)the error introduced by the reduction process:

$$\mathrm{MSE}_{\mathrm{Ours}} = \frac{K'}{K}\mathrm{MSE}_{\mathrm{EM}} + \frac{1}{K}\sum_{i=1}^{t} h_i \sigma_i^2.$$

- K, K' : Initial number of value and remaining number of value
- h_i, σ_i : The number of value and their variance in the i-th merging operation

$\mathrm{MSE}_{\mathrm{Ours}} < \mathrm{MSE}_{\mathrm{EM}}$ **, especially when ε or n is insufficient**

Evaluation

Datasets

S-MN(n=2000 & n=50000), SFC(n=43,386) Income (n=300,000)

Tasks

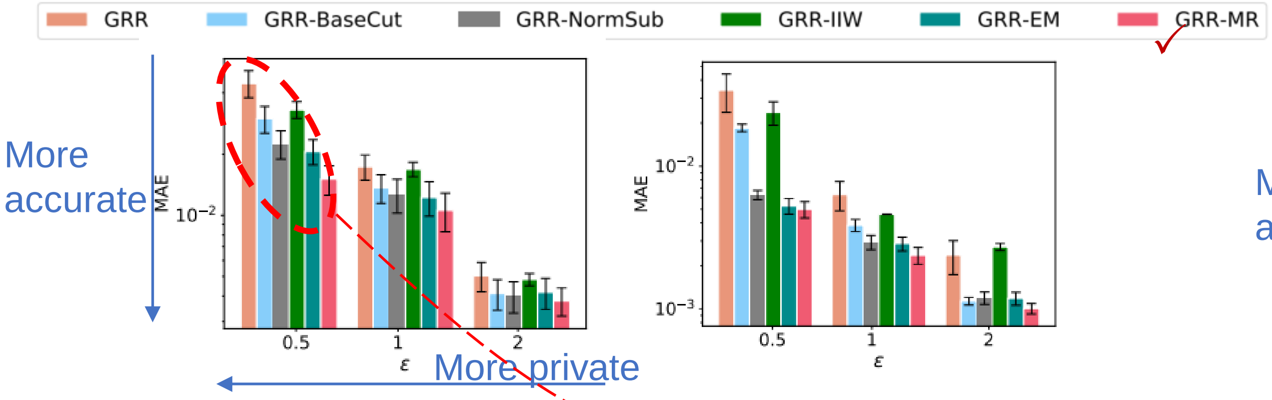
Categorical data: Distribution
Numerical data: Mean & Density
Key-Value data: Conditional mean & density

Evaluation metrics

Mean Absolute Error
Mean Squared Error
Wasserstein Distance
Quantile

Evaluation

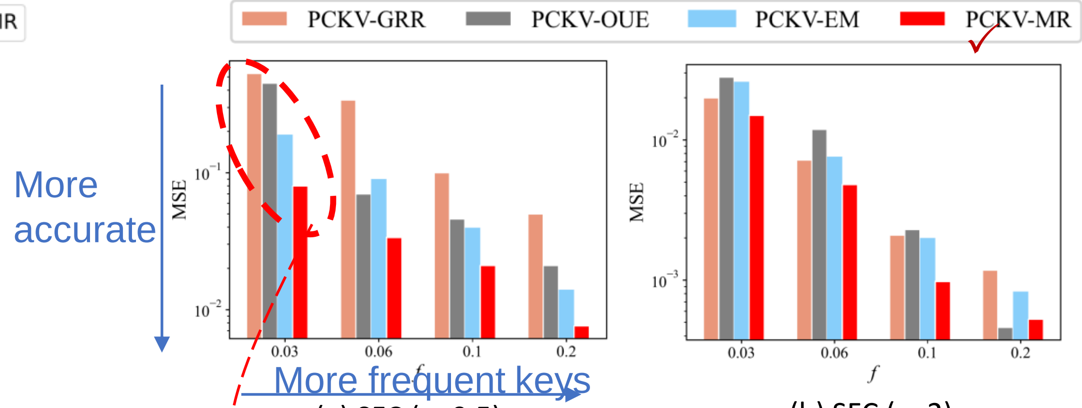
Accuracy comparison on frequency estimation



(a) SFC (full, $K = 60$)

(b) Income (full, $K = 300$)

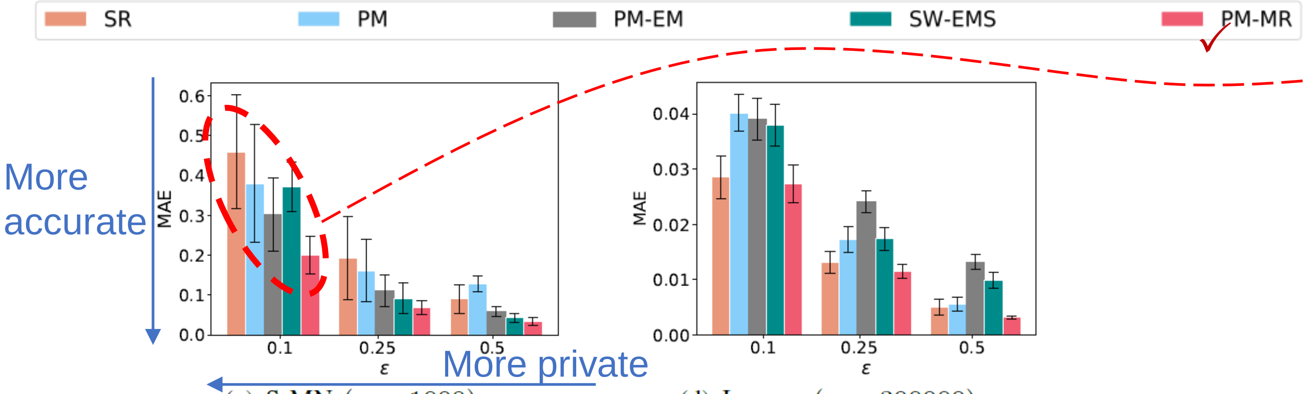
Accuracy comparison on conditional mean



(a) SFC (ε=0.5)

(b) SFC (ε=2)

Accuracy comparison on mean



(a) S-MN ($n = 1000$)

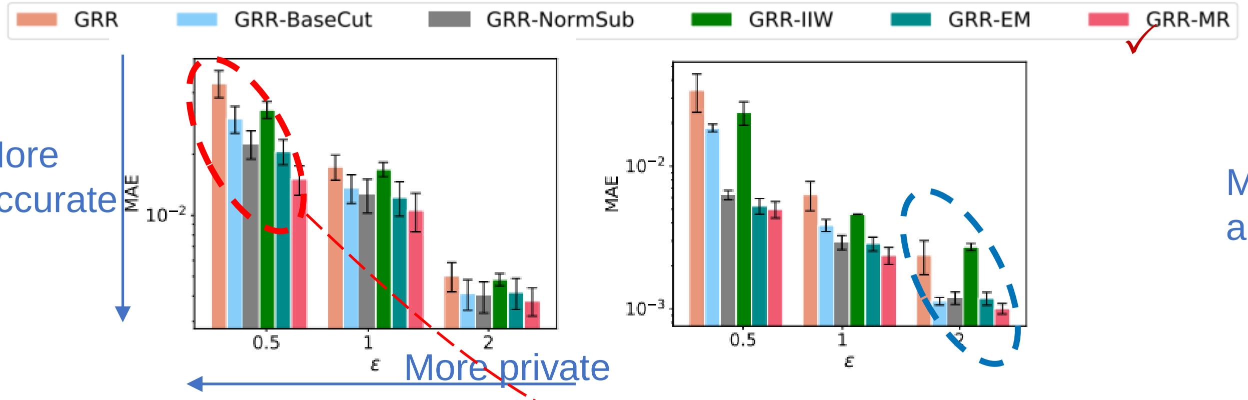
(d) Income ($n = 300000$)

- When there exists substantial noise (low $\boldsymbol{\varepsilon}$ or $\boldsymbol{n}$), our (–MR) reduces the error by almost half.

Evaluation

Accuracy comparison on frequency estimation

More accurate



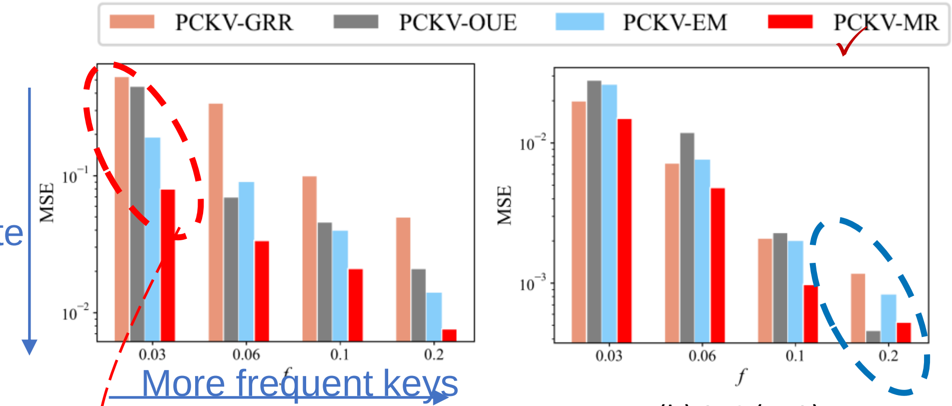
More private

(a) SFC (full, $K = 60$)

(b) Income (full, $K = 300$)

Accuracy comparison on conditional mean

More accurate



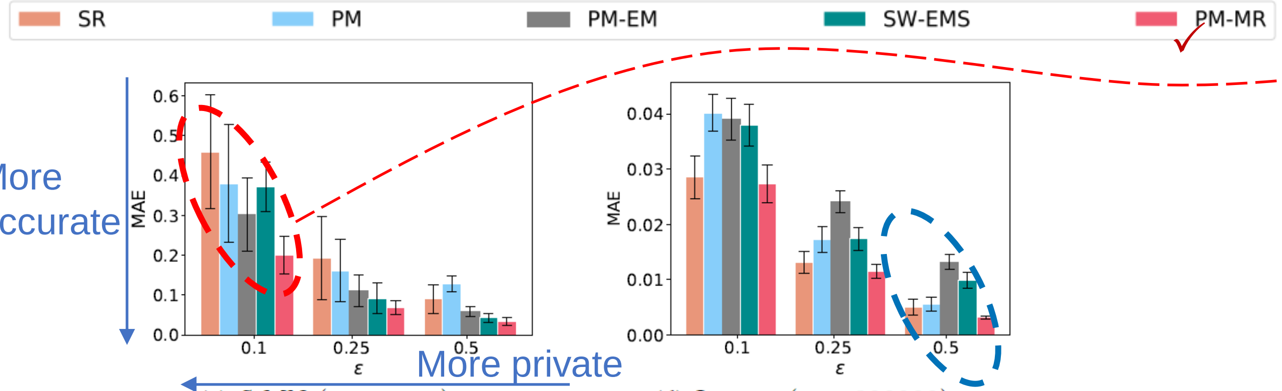
More frequent keys

(a) SFC (ε=0.5)

(b) SFC (ε=2)

Accuracy comparison on mean

More accurate



More private

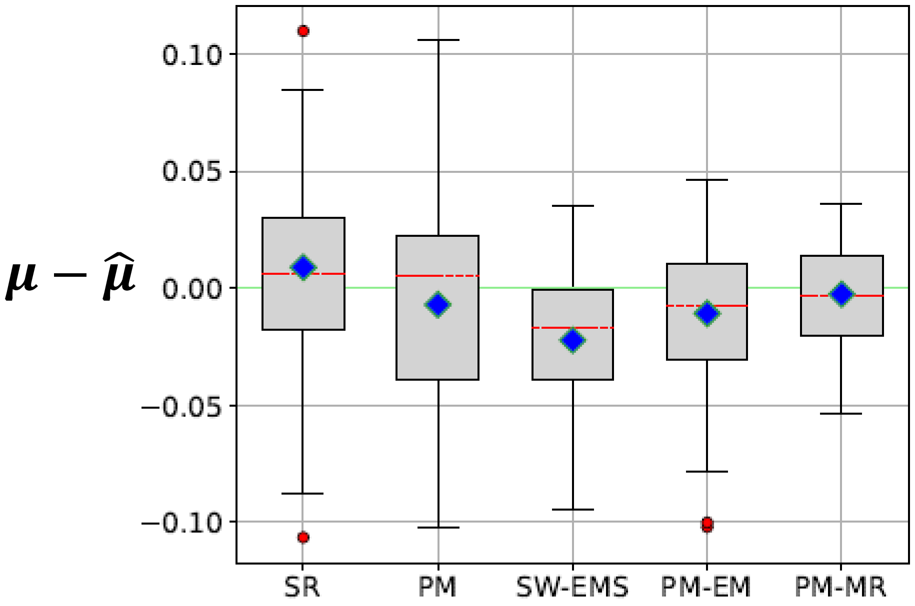
(a) S-MN ($n = 1000$)

(d) Income ($n = 300000$)

- When there exists substantial noise (low $\boldsymbol{\varepsilon}$ or $\boldsymbol{n}$), our (–MR) reduces the error by almost half.
- As $\boldsymbol{\varepsilon}$ or $\boldsymbol{n}$ increases, the advantage of our (-MR) over others gradually diminish ↓.

The reason behind our MR’s high performance

Bias vs Variance

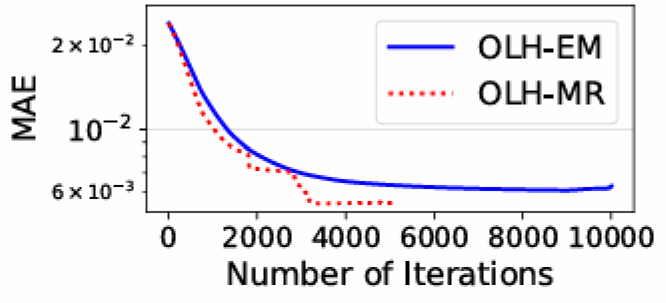


Repeat 100 times, plot the error in boxplot.

EM overfitting → too much bias

Efficiency

Convergence speed on distribution estimation



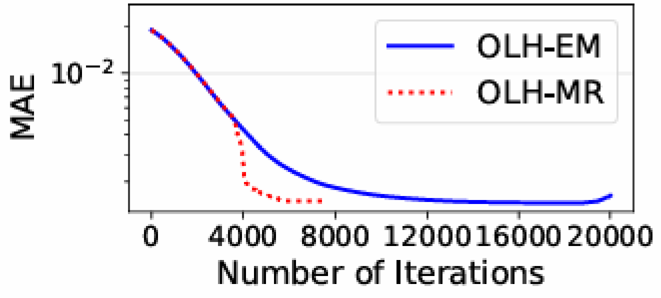
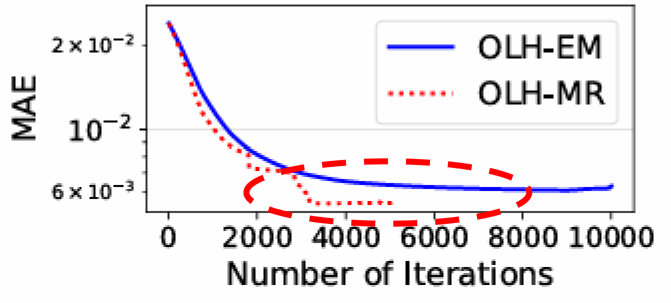


TABLE II
RUNTIME TABLE (SECONDS) OF -EM AND -MR ON DIFFERENT DATASETS, VARYING ε.

	method	ε			
		0.75	1	2	3
SFC	GRR-EM	19	12	9	6
	GRR-MR	10	5	4	4
	OLH-EM	765	502	115	58
	OLH-MR	311	204	83	37
	Laplace-EM	2317	931	416	125
	Laplace-MR	1156	665	306	90
Income	GRR-EM	23	17	12	7
	GRR-MR	11	8	5	4
	OLH-EM	15684	6482	1126	154
	OLH-MR	2837	1697	279	67
	Laplace-EM	12317	8152	2516	823
	Laplace-MR	5457	3003	1026	412

Efficiency

Convergence speed on distribution estimation



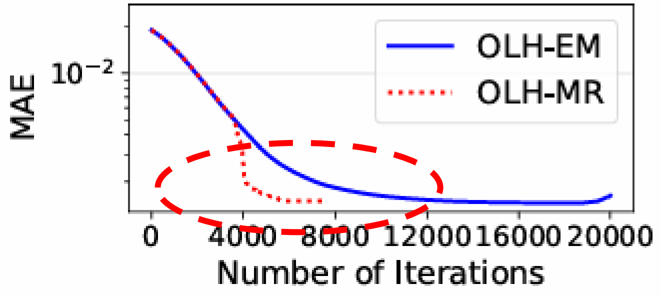


TABLE II
RUNTIME TABLE (SECONDS) OF -EM AND -MR ON DIFFERENT DATASETS, VARYING ε.

	method	ε 0.75	1	2	3
SFC	GRR-EM	19	12	9	6
	GRR-MR	10	5	4	4
	OLH-EM	765	502	115	58
	OLH-MR	311	204	83	37
	Laplace-EM	2317	931	416	125
	Laplace-MR	1156	665	306	90
Income	GRR-EM	23	17	12	7
	GRR-MR	11	8	5	4
	OLH-EM	15684	6482	1126	154
	OLH-MR	2837	1697	279	67
	Laplace-EM	12317	8152	2516	823
	Laplace-MR	5457	3003	1026	412

Ours converged faster

When to use our MR?

MR vs Unbiased estimation & EM

1. MR can replace the traditional EM.
2. When many values need to be estimated.
3. When there exists substantial noise (low $\boldsymbol{\varepsilon}$ or $\boldsymbol{n}$).

Source code is available at https://github.com/yyt20080808/LDP-EM-MR

For additional information contact us:
yutong2017@iscas.ac.cn