Hiding Numerical Vectors in Local Private and Shuffled Messages

Shaowei Wang[1] , **Jin Li**[1*] , **Yuqiu Qian**[2] , **Jiachun Du**[2] , **Wenqing Lin**[2] , **Wei Yang**[3]
[1]Institute of Artificial Intelligence and Blockchain, Guangzhou University
[2]Interactive Entertainment Group, Tencent Inc.
[3]Department of Computer Science and Technology, University of Science and Technology of China
{wangsw,lijin}@gzhu.edu.cn, {yuqiuqian,kevinjcdu,edwlin}@tencent.com, qubit@ustc.edu.cn

Abstract

Numerical vector aggregation has numerous applications in privacy-sensitive scenarios, such as distributed gradient estimation in federated learning, and statistical analysis on key-value data. Within the framework of local differential privacy, this work gives tight minimax error bounds of $O(\frac{ds}{n\epsilon^2})$, where d is the dimension of the numerical vector and s is the number of non-zero entries. An attainable mechanism is then designed to improve from existing approaches suffering error rate of $O(\frac{d^2}{n\epsilon^2})$ or $O(\frac{ds^2}{n\epsilon^2})$. To break the error barrier in the local privacy, this work further consider privacy amplification in the shuffle model with anonymous channels, and shows the mechanism satisfies centralized $(\sqrt{14\ln(2/\delta)\frac{s\cdot e^{\epsilon}+2s-1}{n-1}}, \delta)$-differential privacy, which is domain independent and thus scales to federated learning of large models. We experimentally validate and compare it with existing approaches, and demonstrate its significant error reduction.

1 Introduction

With the enacting of increasingly rigid regulations on data privacy (e.g., the General Data Protection Regulation [Voigt and Von dem Bussche, 2017] in the Europe Union, the California Consumer Privacy Act, and the Civil Code of the People's Republic of China), local differential privacy (LDP) has become the *de facto* notion for data privacy preservation over the Internet. It originates from the classical notion of differential privacy in the database community [Dwork, 2008] without the trust of the data aggregator or other third parties. LDP allows every user/agent to sanitize their personal data locally (e.g., on mobile devices, IoT sensors or edge servers) and provides information-theoretically rigorous privacy protection. Currently, many giant internet service providers (such as Apple [Greenberg, 2016], Google [Erlingsson *et al.*, 2014] and Microsoft [Ding *et al.*, 2017]) are deploying LDP for regulation compliance when collecting and analyzing user data. As a remedy to the unacceptable error barrier due to stringent LDP constraints, researchers recently introduce shuffle model [Erlingsson *et al.*, 2019] where messages from users are permuted (by a shuffler, e.g., anonymous channels) before sent to the aggregator. The linkage between users and their messages are cutted off and messages could hide in the crowd, [Erlingsson *et al.*, 2019] show privacy is amplified with shuffling, thus a lower privacy level can be adopted locally to satisfy a relatively higher privacy level as in the analogised central model.

Plenty of user data are in the form of numerical vectors. Let $\mathbf{x}^i$ denote the numerical vector of user i. For simplicity but without loss of generality, $\mathbf{x}^i$ can be assumed as a d-dimensional s-sparse ternary vector [Wen *et al.*, 2017; Ye *et al.*, 2019; Sun *et al.*, 2019; Gu *et al.*, 2020], that is:

$$\mathcal{X}^s = \{\mathbf{x} \mid \mathbf{x} \in \{-1, 0, 1\}^d \text{ and } \|\mathbf{x}\|_0 = s\}.$$

This work studies the problem of numerical vector aggregation within the local and shuffled differential privacy framework. Many real-world data aggregation tasks could be formulated as this problem, such as gradient estimation in federated learning and sensitive key-value data aggregation for user profile/usage analyses in online services.

1.1 Federated Gradient Estimation

Federated learning [Konečný *et al.*, 2016] studies machine learning systems in the distributed setting so that each party keeps its own data locally for privacy preserving. At each gradient descent iteration for training/updating a machine learning model, locally computed gradients $\mathbf{x}^i$ from participating parties (e.g., from n mobile users) need to be averaged by the federation server (e.g., a parameter server):

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}^i. \tag{1}$$

For communication efficiency, local gradients are usually discretized and sparsified [Wen *et al.*, 2017; Wangni *et al.*, 2018].

The original work [Konečný *et al.*, 2016] deems sharing gradient to be more privacy-resistant than sharing raw data, but a recent work demonstrates that the gradient $\mathbf{x}^i$ is also privacy risky [Zhu *et al.*, 2019] and local raw data might be derived with confidence from several transmitted gradients. This calls for rigid privacy protection on local gradients.

*Corresponding author.

1.2 Key-value Data Aggregation

We refer to key-value data as any paired (key, value) mappings, where the key $j \in [1, d]$ is an index and the value $\mathbf{x}_j$ is numerical. Note that, the value is deemed as 0 when and only when the corresponding key is missing from or not defined in key-value data. For any existing or defined keys, their corresponding values are binary as $\{-1, 1\}$. For example, a user might represent preferences on watched movies as key-value data, in which movies the user likes are assigned with value 1 and movies the user unlikes are assigned with value -1.

Common analysis on key-value data includes estimating both unconditional mean statistics and conditional mean statistics. The unconditional mean estimation about the key j is $\overline{\mathbf{x}}_j = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_j^i$, and the conditional mean estimation about the key j is:

$$\overline{\mathbf{x}}_{\underline{j}} = \frac{\sum_{i=1}^{n}\mathbf{x}_j^i}{\#\{\mathbf{x}_j^i \mid \mathbf{x}_j^i \ for\ i \in [1, n]\ and\ \mathbf{x}_j^i \neq 0\}}. \tag{2}$$

1.3 Existing Results

Within the framework of ϵ-LDP, theoretical minimax lower bounds for many statistical estimation problems have been established, such as multinomial distribution estimation [Duchi *et al.*, 2013], logistic regression/generalized linear model estimation [Duchi *et al.*, 2018] and sparse covariance matrix estimation [Wang and Xu, 2019]. Specifically, the work of [Duchi *et al.*, 2018] derives minimax lower bounds for multi-dimensional mean estimation for numerical vectors with bounded ℓ_1-norm or ℓ_2-norm. However, an s-sparse numerical vector with bounded ℓ_1-norm or ℓ_2-norm is a special case of [Duchi *et al.*, 2018] with identical absolute non-zero entries. *Whether it holds the same bounds as the general case or has tighter bounds* is still an open question. Recently, for a broad family of ϵ-LDP estimation problems that can be cast as a mean estimation one, the work of [Błasiok *et al.*, 2019] studies sample complexity lower bounds under certain error tolerance α, but their sample complexity result for s-sparse numerical vectors has at least a $1/\alpha$ gap to ours result of minimax optimal sample complexity.

Practically, plenty of ϵ-LDP mechanisms have been proposed for statistical data estimation, such as multinomial distribution estimation on categorical data [Duchi *et al.*, 2013; Erlingsson *et al.*, 2014; Kairouz *et al.*, 2016; Wang *et al.*, 2020b], and one-dimensional mean estimation on numerical values [Wang *et al.*, 2019; Sun *et al.*, 2020]. For ϵ-LDP numerical vector or key-value data aggregation, existing approaches deal with both dense numerical vectors (e.g., in [Nguyên *et al.*, 2016; Duchi *et al.*, 2018]) and sparse numerical vectors (e.g., in [Ye *et al.*, 2019; Sun *et al.*, 2019; Gu *et al.*, 2020]). Specifically, the works of [Ye *et al.*, 2019; Sun *et al.*, 2019] uniform-randomly select one dimension from $[1, d]$ and transform the multi-dimensional estimation problem to a one-dimensional numerical/categorical problem. The work of [Gu *et al.*, 2020] follows a similar paradigm, but randomly selects one non-empty dimension from s dimensions. However, as we will show in Section 4, these mechanisms are sub-optimal.

To mitigate the high noise needed for LDP, [Erlingsson *et al.*, 2019] introduces a trusted shuffler to hide private views in the crowd. Recent works [Balle *et al.*, 2020; Ghazi *et al.*, 2020] propose sending multiple unary/binary messages with distributed noise to the shuffler for achieving centralized differential privacy (CDP), while other works [Cheu *et al.*, 2019; Balle *et al.*, 2019] study privacy amplification effects for achieving a lower level of LDP and a higher level of CDP simultaneously via shuffling. Specifically, [Balle *et al.*, 2019; Wang *et al.*, 2020a; Liu *et al.*, 2021] utilize the technique of *privacy blanket* for privacy amplification on binomial/multinomial distribution estimation. This work further show domain-independent privacy amplification is achievable for sparse numerical vector.

1.4 Our Contributions

Minimax Lower Bounds. The MSE lower bound of ϵ-LDP s-sparse numerical vector mean estimation is $O(\frac{ds}{n\epsilon^2})$. Our proof considers s-sparse numerical vectors that are decomposable, hence reduces the bounding procedure to cases of multiple multinomial distribution estimations.

An Optimal Mechanism. Since existing approaches are sub-optimal, we design a mechanism that matches the minimax lower bound. The mechanism has computational complexity of $O(s)$ and communication complexity of $O(\log s)$.

Domain-independent Privacy Amplification. For the shuffle model of s-sparse numerical vector aggregation, the proposed optimal mechanism satisfies centralized $(\sqrt{14\ln(2/\delta)\frac{s\cdot e^{\epsilon}+2s-1}{n-1}}, \delta)$-differential privacy. The privacy loss ϵ_c in CDP is independent of domain size d, thus fits federated learning of large models. In turn, when privacy budget ϵ_c is given, we derive local parameter for optimal utility and show the proposed mechanism is asymptotic near-optimal in terms of user population.

The remainder of the paper is organized as follows. Section 2 provides preliminary knowledge on local differential privacy and minimax risk framework of statistical estimation. The minimax lower bound on the ϵ-LDP numerical vector aggregation problem is then given in Section 3. Next, Section 4 reviews the design of existing mechanisms and shows their sub-optimality, and then propose a new mechanism and prove its optimality. Section 5 shows the proposed mechanism enjoys domain-independent privacy amplification in the shuffle model, and prove its asymptotic optimality. Later, Section 6 demonstrates the superior performance of the proposed mechanism against existing mechanisms. Finally, Section 7 concludes the paper.

2 Preliminaries

2.1 Differential Privacy

For datasets D, D' that are of the same size and differ only in one element, they are called *neighboring datasets*. The centralized differential privacy with budget/level (ϵ, δ) is as follows.

Definition 1 ((ϵ, δ)-CDP [Dwork, 2008]). *Let $\mathcal{D}_K$ denote the output domain, a randomized mechanism K satisfies ϵ-differential privacy iff for any neighboring datasets D, D', and any outputs $\mathbf{z} \subseteq \mathcal{D}_K$,*

$$\mathbb{P}[K(D) \in \mathbf{z}] \leq \exp(\epsilon) \cdot \mathbb{P}[K(D') \in \mathbf{z}] + \delta.$$

Let K denote a randomized mechanism for sanitizing a single user data, the LDP with privacy budget ϵ is as follows.

Definition 2 (ϵ-LDP [Duchi *et al.*, 2013]). *Let $\mathcal{D}_K$ denote the output domain, a randomized mechanism K satisfies local ϵ-differential privacy iff for any data pair $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^s$, and any output $z \in \mathcal{D}_K$,*

$$\mathbb{P}[K(\mathbf{x}) = z] \leq \exp(\epsilon) \cdot \mathbb{P}[K(\mathbf{x}') = z].$$

2.2 Local Private Minimax Risks

Assuming samples $\{x^1, x^2, ..., x^n\}$ that are n i.i.d. drawn from a distribution $P \in \mathcal{P}$. Let $\mathcal{K}_\epsilon$ denote the set of all possible mechanisms $\mathbf{K} = \{K^1, ..., K^n\}$ that satisfy ϵ-LDP for every sample in $\{x^1, x^2, ..., x^n\}$. Taking as input the samples, some mechanism $\mathbf{K} \in \mathcal{K}_\epsilon$ produces a list of sanitized views $\{z^1, z^2, ..., z^n\}$. If the parameter estimator:

$$\hat{\theta} = \hat{\theta}(\{z^1, z^2, ..., z^n\})$$

is derived from these private views while having no access to input samples $\{x^j\}_{j=1}^n$, the minimax MSE risk (under privacy budget ϵ) is then:

$$\begin{aligned}&\mathfrak{M}_n(\theta(\mathcal{P}), ||\cdot||_2^2, \epsilon)\\ &:= \inf_{\mathbf{K}\in\mathcal{K}_\epsilon} \inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_{p,\mathbf{K}}[||\hat{\theta}(z^1, z^2, ..., z^n) - \theta(P)||_2^2].\end{aligned}$$

3 Minimax Lower Bounds

The Assouad's method [Yu, 1997] is a common tool for lower bounding via multiple hypothesis testings. It defines a hypercube $\mathcal{V} = \{-1, 1\}^d$ ($d \in \mathbb{N}$), then defines a family of distributions $\{P_\nu\}_{\nu\in\mathcal{V}}$ indexed by the hypercube, where each P_ν is defined on one common space. It's said that the distribution family induces a 2τ-Hamming separation for the loss $||\cdot||_2^2$, if there exists a vertex mapping (a function $\kappa : \theta(\mathcal{P}) \mapsto \{-1, 1\}^d$) satisfying:

$$||\theta - \theta(P_\nu)||_2^2 \geq 2\tau \sum_{j=1}^{d} \mathbf{1}\{[\kappa(\theta)]_j \neq \nu_j\}.$$

Assume that the nature first uniform-randomly chooses a vector $V \in \{-1, 1\}^d$, and the samples $\{\mathbf{x}^1, ..., \mathbf{x}^n\}$ are drawn from the distribution P_ν with $V = \nu$. These samples are then taken as input into ϵ-LDP mechanisms $\mathbf{K}$. The literature [Duchi *et al.*, 2018] gives an ϵ-LDP version of Assouad's method as follows.

Lemma 1 (Private Assouad bound [Duchi *et al.*, 2018]). *Let $P_{+j} = \frac{1}{2^{d-1}} \sum_{\nu:\nu_j=1} P_\nu$ and $P_{-j} = \frac{1}{2^{d-1}} \sum_{\nu:\nu_j=-1} P_\nu$, we have*

$$\mathfrak{M}_n(\theta(\mathcal{P}), ||\cdot||_2^2) \geq d \cdot \tau[1 - (\frac{n(e^\epsilon - 1)^2}{2d} F_{\mathbb{B}_\infty(\mathcal{X}^s),\mathcal{P}})^{\frac{1}{2}}],$$

where $\mathbb{B}_\infty(\mathcal{X}^s)$ denote the collection of function γ with supremum norm bounded by 1 as:

$$\mathbb{B}_\infty(\mathcal{X}^s) := \{\gamma : \mathcal{X}^s \mapsto \mathbb{R} \mid ||\gamma||_\infty \leq 1\},$$

and maximum possible discrepancy $F_{\mathbb{B}_\infty(\mathcal{X}^s),\mathcal{P}}$ is defined as:

$$\sup_{\gamma\in\mathbb{B}_\infty(\mathcal{X}^s)} \sum_{i=1}^{d} (\int_{\mathcal{X}^s} \gamma(x)(dP_{+j}(x) - dP_{-j}(x)))^2.$$

We consider numerical vectors that can be decomposed the into s buckets, each bucket has $\frac{d}{s}$ indexes with only one non-zero entry. We then define a hypercube of length d and construct a class of $\frac{2\delta^2 s^2}{d^2}$-hamming separated probability distributions. Guided by Lemma 1, we bound the maximum possible marginal distance $F_{\mathbb{B}_\infty(\mathcal{X}^s),\mathcal{P}}$ under the value of $\frac{8\delta^2 s}{d}$. Theorem 1 gives the final lower bounds for the problem of local private numerical vector mean estimation.

Theorem 1. *For the numerical vector aggregation problem, for any ϵ-LDP mechanism, there exists a universal constant $c > 0$ such that for all $\epsilon \in [0, 1]$,*

$$\mathfrak{M}_n(\theta(\mathcal{P}), ||\cdot||_2^2, \epsilon) \geq c \cdot \min\{\frac{s^2}{d}, \frac{ds}{n\epsilon^2}\}.$$

To understanding the minimax rate, we can consider the non-private error rate of decomposable numerical vector aggregation, which is $\mathbb{E}||\hat{\theta} - \theta||_2^2 \leq \sum_{i=1}^{d} \mathbb{E}||\hat{\theta}_i - \theta_i||_2^2 \leq \frac{4s}{n}$. Thus the enforcement of local ϵ-LDP causes the effective sample size decreasing from n to $n\epsilon^2/d$.

4 Optimal Mechanism

Let j_- and j_+ denote events that the j-th element of $\mathbf{x}$ (i.e. $\mathbf{x}_j$) equals to -1 and 1 respectively, a numerical vector $\mathbf{x}$ could be represented in the set form as:

$$\mathbf{Y}_\mathbf{x} = \{j_- \mid j \in [1, d] \text{ and } \mathbf{x}_j = -1\} \bigcup \{j_+ \mid j \in [1, d] \text{ and } \mathbf{x}_j = 1\}.$$

Existing works on ϵ-LDP numerical vector aggregation can be categorized into two types, they do dimension sampling in a data-agnostic manner (e.g., the PrivKV in [Ye *et al.*, 2019; Sun *et al.*, 2019]) or a data-dependent manner (e.g., the PCKV in [Gu *et al.*, 2020]).

The PrivKV Mechanism. The seminal work [Ye *et al.*, 2019] on ϵ-LDP key-value data aggregation propose to firstly randomly sample a dimension (from the domain of keys) $j \in [1, d]$, then applies an ϵ-LDP categorical mechanism on the corresponding (key,value) pair that takes a value from $\{(j, 0), (j, 1), (j, -1)\}$, where $(j, 0)$ means that the key is empty in the key-value data. Essentially, the PrivKV mechanism is equivalent to dividing the population of n into d groups, and each group is employed to estimate $[j_+ \in \mathbf{Y}_\mathbf{x}]$ and $[j_- \in \mathbf{Y}_\mathbf{x}]$ for each $j \in [1, d]$ with privacy budget ϵ. Since the minimax lower error bound for estimating frequencies on population of n' with privacy budget ϵ and domain size d' is $\Theta(\frac{d'}{n'\epsilon^2})$ [Duchi *et al.*, 2018], the estimation error of $[j_+ \in \mathbf{Y}_\mathbf{x}]$ and $[j_- \in \mathbf{Y}_\mathbf{x}]$ is hence $\Theta(\frac{d}{n\epsilon^2})$, as $n' = \frac{n}{d}$ and $d' = 3$. Therefore, its total estimation error of frequencies or mean values of d-dimensional vector is $O(\frac{d^2}{n\epsilon^2})$. It has a gap of $\frac{d}{s}$ from the optimal error rate in Theorem 1. Similar methodology and result also hold for the following-up works in [Sun *et al.*, 2019; Liu *et al.*, 2021].

The PCKV Mechanism. The work of [Gu *et al.*, 2020] proposes to sample one key from existing s keys in a key-value data. Afterwards, an ϵ-LDP categorical mechanism is applied to the corresponding 1-sparse numerical vector, which is equivalent to categorical data with domain size of around $2d$. Recall that the minimax lower error bound for estimating frequencies on population of n' with privacy budget ϵ and domain size d' is $\Theta(\frac{d'}{n'\epsilon^2})$, the total estimation error of scaled $[j_+ \in \mathbf{Y_x}]$ and $[j_- \in \mathbf{Y_x}]$ in the PCKV mechanism is hence $\Theta(\frac{d}{n\epsilon^2})$, as $n' = n$ and $d' = 2d$. Due to the previous sampling procedure, the scale factor is s and the total variation error is amplified by s^2, thus the total estimation error of $[j_+ \in \mathbf{Y_x}]$ and $[j_- \in \mathbf{Y_x}]$ in the PCKV mechanism is $O(\frac{ds^2}{n\epsilon^2})$. It has a gap of s from the optimal error rate in Theorem 1.

4.1 Our Design

The paradigm of *dimension sampling & categorical randomization* fails to achieve optimal statistical rate for ϵ-LDP numerical vector aggregation. Therefore, we consider randomizing the numerical vector as a whole with the exponential mechanism [McSherry and Talwar, 2007], and propose the *Collision mechanism*.

If defining event domain as: $\mathcal{Y} = \{1_-, 1_+, 2_-, 2_+, ..., d_-, d_+\}$, we have $\mathbf{Y_x}$ as a subset of $\mathcal{Y}$ with size s. Define the output domain as $\mathcal{Z} = \{1, 2, ..., t\}$, the Collision mechanism probabilistically outputs one item $z \in \mathcal{Z}$. The outputting probabilities are based on whether each item has collision with hashed events in $\mathbf{Y_x}$. The Collision mechanism is formally given in Definition 3.

Definition 3 ((d, s, ϵ, t)-Collision Mechanism)**.** *Given a random-chosen hash function $H : \mathcal{Y} \mapsto \mathcal{Z}$, take an s-sparse numerical vector $\mathbf{Y_x} \subseteq \mathcal{Y}$ as input, the Collision mechanism randomly outputs an element $z \in \mathcal{Z}$ according to following probability design:*

$$\mathbb{P}[z|\mathbf{x}] = \begin{cases} \frac{e^\epsilon}{\Omega}, & \text{if } \exists y \in \mathbf{Y_x}, z = H(y); \\ \frac{\Omega - e^\epsilon \cdot \#\{H(y) \mid H(y) \; for \; y \in \mathbf{Y_x}\}}{(t - \#\{H(y) \mid H(y) \; for \; y \in \mathbf{Y_x}\}) \cdot \Omega}. & otherwise. \end{cases}$$

The normalization factor is $\Omega = s \cdot e^\epsilon + t - s$. An unbiased estimator of indicator $[j_b \in \mathbf{Y_x}]$ for $b \in \{-1, 1\}$ and $j \in [1, d]$ is (for $s \geq 2$):

$$\widehat{[j_b \in \mathbf{Y_x}]} = \frac{[H(j_b) = z] - 1/t}{e^\epsilon/\Omega - 1/t}.$$

The privacy guarantee of the mechanism is given in Proposition 1, which is obvious as $s \geq \#\{H(y) \mid H(y) \; for \; y \in \mathbf{Y_x}\}$. The utility-optimality guarantee of the mechanism is given in Theorem 2. For $\epsilon = O(1)$, its computational complexity is bounded by $t^* \approx s + 2s - 1 + s \cdot e^\epsilon = O(s)$, and communication complexity is $\log_2(2s - 1 + s \cdot e^\epsilon) = O(\log s)$.

Proposition 1. *The (d, s, ϵ, t)-Collision mechanism in Definition 3 satisfies ϵ-LDP for numerical vector data.*

Theorem 2. *Given privacy budget $\epsilon = O(1)$, with optimal choice of the parameter t^*, the mean estimation error of (d, s, ϵ, t)-Collision mechanism for numerical vector is $O(\frac{ds}{n\epsilon^2})$.*

Proof. Since $[H(j_b) = z]$ are Bernoulli random variables, we have the mean squared error:

$$\begin{aligned} Var[\widehat{\mathbf{x}}] &\leq 2 \sum_{j=1}^{d} \sum_{b \in [-1,1]} Var[\widehat{[j_b \in \mathbf{Y_x}]}] \\ &\leq \frac{2}{n} \cdot \frac{s \cdot e^\epsilon/\Omega(1 - e^\epsilon/\Omega) + (2d - s) \cdot 1/t(1 - 1/t)}{(e^\epsilon/\Omega - 1/t)^2}. \end{aligned}$$

Taking the previous formula as a function of continuous t, actually the function is convex when $d \geq t \geq s$. Choosing approximate optimal t^* at around $2s - 1 + s \cdot e^\epsilon$, we then have (with $e^\epsilon \approx \epsilon + 1$):

$$Var[\widehat{\mathbf{x}}] \leq \frac{2d \cdot \Theta(s^3) + \epsilon \cdot \Theta(s^3)}{n \cdot \epsilon^2 \cdot (-1 + (2 + \epsilon) \cdot s)^2} \leq O(\frac{ds}{n\epsilon^2}).$$

□

5 Privacy Amplification in Shuffle Model

When semi-trusted shufflers lie between users and the aggregator, the aggregator only observes the multi-set of private views $\{z_1, z_2, ...z_n\}$, thus the privacy level of some LDP mechanisms is amplified (w.r.t. CDP). For binary domain of private views $\mathcal{Z} = \{1, 2\}$ generated from randomized response, [Cheu *et al.*, 2019; Balle *et al.*, 2019] show the observed frequencies of $z \in \mathcal{Z}$ satisfies $(\sqrt{14 \ln(2/\delta) \frac{e^\epsilon + 1}{n-1}}, \delta)$-CDP, since about $\frac{2(n-1)}{e^\epsilon + 1}$ users response uniform randomly, they contributed $\mathbf{B}(\frac{2(n-1)}{e^\epsilon + 1}, 0.5)$ Binomial random noises to the frequencies.

Similarly, in the Collision mechanism with private view domain $\mathcal{Z} = \{1, 2, ..., t\}$, about $\frac{n-1}{s \cdot e^\epsilon + t - s}$ users response uniform randomly, they contributed $\mathbf{B}(\frac{n-1}{s \cdot e^\epsilon + t - s}, 1/t)$ Binomial random noises to each frequency of $z \in \mathcal{Z}$. As a result, the Collision mechanism also satisfies $(\sqrt{14 \ln(2/\delta) \frac{s \cdot e^\epsilon + t - s}{n-1}}, \delta)$-CDP. The formal guarantee of privacy amplification of the Collision mechanism is presented in Theorem 3.

Theorem 3. *In the shuffle model, the (d, s, ϵ, t)-Collision mechanism satisfies $(\sqrt{14 \ln(2/\delta) \frac{s \cdot e^\epsilon + t - s}{n-1}}, \delta)$-differential privacy when $n \geq \frac{27(e^\epsilon + t - 1)}{\sqrt{14 \ln(2/\delta) \frac{s \cdot e^\epsilon + t - s}{n-1}}} + 1$.*

Proof. Since private views with hash functions are randomly shuffled and the final estimator is derived from them, to prove the final estimator is $(\sqrt{14 \ln(2/\delta) \frac{s \cdot e^\epsilon + t - s}{n-1}}, \delta)$-CDP, it is enough to show that the distribution of observed private views with hash functions satisfies CDP. Further because the hash functions are randomly chosen from some universe, we only need to show the frequency distribution of observed private views satisfies CDP. The frequency distribution is noised by $N \sim \mathbf{B}(n - 1, \frac{t}{s \cdot e^\epsilon + t - s})$ users, each user contributed with uniform-random t-multinomial distribution in the output domain. Then, according to the tail bounding result on this noise distribution with multiplicative Chernoff bound or Bennett's inequality (the privacy blanket theorem in [Balle *et al.*, 2019]), such a noise distribution satisfies $\sqrt{14 \ln(2/\delta) \frac{s \cdot e^\epsilon + t - s}{n-1}}$-CDP with probability at least $1 - \delta$ when $n \geq \frac{27(e^\epsilon + t - 1)}{\sqrt{14 \ln(2/\delta) \frac{s \cdot e^\epsilon + t - s}{n-1}}} + 1$. □

With optimal choice of $t \approx 2s - 1 + s \cdot e^\epsilon$ in LDP, the privacy amplification bound is thus independent of the domain size d. Alternatively, when the privacy level in the CDP is given as (ϵ_c, δ), we have $\Omega = s \cdot e^\epsilon + t - s = \frac{\epsilon_c^2 (n-1)}{14 \ln 2/\delta}$, then

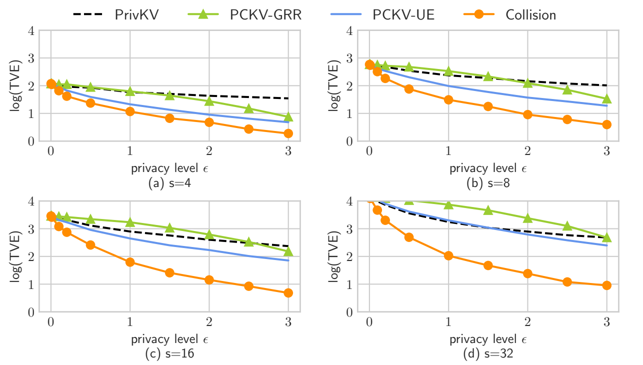

Figure 1: TVE results on $n = 100,000$ users with dimension $d = 256$ when sparsity s ranges from 4 to 32.

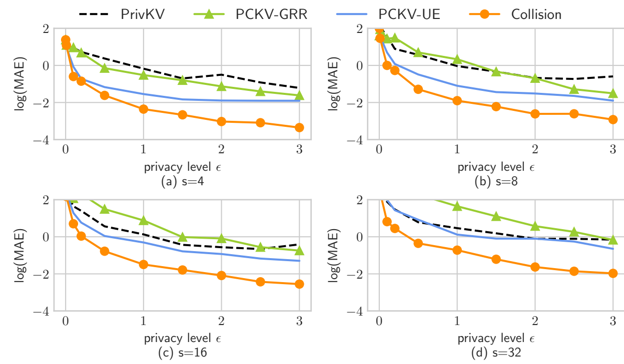

Figure 2: MAE results on $n = 100,000$ users with dimension $d = 256$ when sparsity s ranges from 4 to 32.

the numerical vector estimation error bound is a function of t with constant Ω as:

$$Var[\hat{\bar{\mathbf{x}}}] \leq \frac{\frac{(2d-s)(t-1)}{t^2} + \frac{(\Omega+s-t)(\Omega(s-1)-s+t)}{\Omega^2 s}}{\frac{\Omega+s-t}{\Omega s} - \frac{1}{t}}.$$

Choosing $t \approx \frac{4+\Omega+s+\sqrt{\Omega^2+2\Omega(7s-8)+s^2-16s+16}}{6}$ approximately minimizes the error, which is also independent of the domain size d.

With the optimal choice of t in terms of CDP, when the numerical vector is highly sparse (i.e. $d \gg s$) and the user population is large, we have the asymptotic estimation error in the shuffle setting as:

$$Var[\hat{\bar{\mathbf{x}}}] = O\left(\frac{ds^2 \ln 1/\delta}{n^2\epsilon_c^2} \cdot \max\{1, \frac{14s \ln 2/\delta}{\epsilon_c^2(n-1)}\}\right),$$

which nearly matches the optimal error rate in the centralized differential privacy setting (the sensitivity is $2s$).

6 Experiments

The statistical efficiency of the proposed Collision mechanism for ϵ-LDP numerical vector aggregation is evaluated in this section. Competing mechanisms include the PrivKV mechanism [Ye *et al.*, 2019], the PCKV mechanism with general randomized response as the base randomizer (denoted as PCKV-GRR), and the PCKV mechanism with unary encoding as the base randomizer[Gu *et al.*, 2020] (denoted as PCKV-UE). Since the performances of all these mechanisms are data-independent, it is enough to utilize synthetic datasets for fair evaluation. The parameters of synthetic datasets are listed as follows (default values are in bold form), covering most cases encountered in real-world applications:

i. Number of users n: 10,000, **100,000**.
ii. Dimension d: 256, **1024**.
iii. Sparsity parameter s: 4, 8, **16**, 32.
iv. Privacy budget ϵ: 0.001, 0.01, 0.1, 0.2, 0.4, 0.8, 1.0, 1.5, 2.0, 2.5, 3.0.

During each simulation, the numerical vector of each user is independent-randomly generated, the non-zero entries are uniform-randomly selected from d dimensions, and each dimension has an equal probability of being -1 or 1.

6.1 Evaluation Metric

Previous theoretical results focus on the mean squared error: $\sum_{j\in[1,d]} ||\hat{\bar{\mathbf{x}}}_j - \bar{\mathbf{x}}_j||_2^2$. Here we evaluate mechanisms with metrics on frequency estimators of $[j_b \in \mathbf{Y_X}]$ (including TVE and MAE), which are basic statistics for both the unconditional and conditional mean estimation in Equation (1) and Equation (2). The total variation error (TVE) is defined as:

$$\text{TVE} = \sum_{j\in[1,d],\ b\in\{-1,1\}} |\widehat{[j_b \in \mathbf{Y_X}]} - [j_b \in \mathbf{Y_X}]|_1,$$

and the maximum absolute error (MAE) is defined as:

$$\text{MAE} = \max_{j\in[1,d],\ b\in\{-1,1\}} |\widehat{[j_b \in \mathbf{Y_X}]} - [j_b \in \mathbf{Y_X}]|_1.$$

Since the $\frac{1}{s}$-scaled frequencies lie in the d dimensional probability simplex, the estimated frequencies are projected into the Δ_d-simplex as in [Wang and Carreira-Perpinán, 2013]. All experimental results are the mean *natural logarithm* value of 10 repeated simulations.

6.2 Effects of Sparsity s

Assume that there are $n = 100,000$ users, and the dimension is $d = 256$. When the number of non-zero entries in numerical vectors varies from 4 to 32, the TVE/MAE error results are presented in Figure 1 and Figure 2 respectively. The PCKV-UE mechanism improves upon the PrivKV in the extreme sparse cases, but for other cases (e.g., $s = 32$), the PCKV-UE and the PrivKV mechanism have similar performances. The Collision mechanism outperforms all competing mechanisms in all cases significantly, and averagely reduces more than 60% errors. As the sparsity parameter s gets larger, the performance gaps get larger.

6.3 Effects of Dimension d

Assume that there are $n = 100,000$ users, but the dimension now becomes $d = 1024$. When the number of non-zero entries in numerical vectors still varies from 4 to 32, the results of TVE and MAE are shown in Figure 3 and Figure 4 respectively. Compared to cases of $d = 256$ (i.e. TVE results in Figure 1 and MAE results in Figure 2), it is easy to observe that the TVE/MAE value grows with around $\sqrt{d}$.

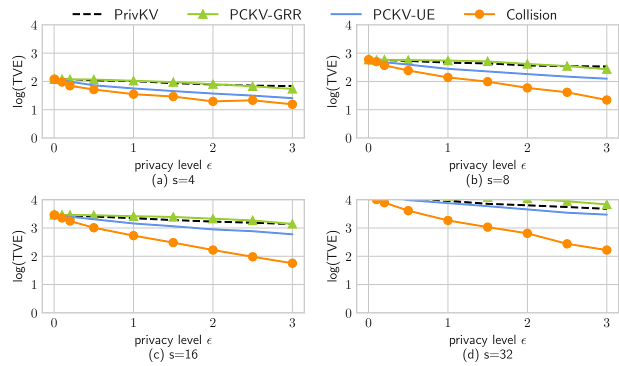

Figure 3: TVE results on $n = 100,000$ users with dimension $d = 1024$ when sparsity s ranges from 4 to 32.

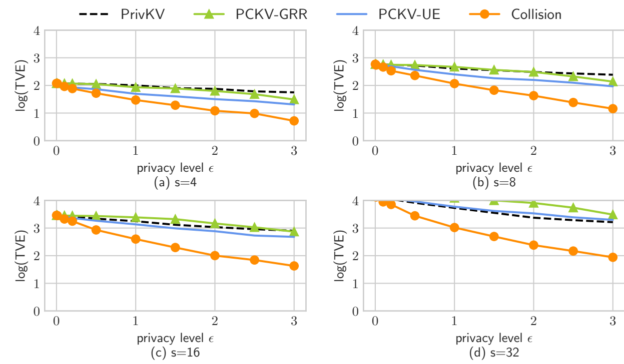

Figure 5: TVE results on $n = 10,000$ users with dimension $d = 256$ when sparsity s ranges from 4 to 32.

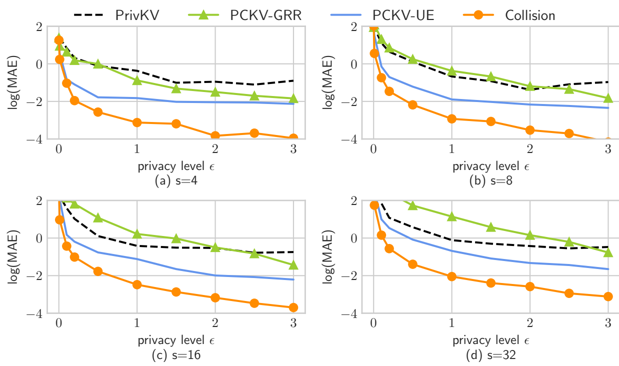

Figure 4: MAE results on $n = 100,000$ users with dimension $d = 1024$ when sparsity s ranges from 4 to 32.

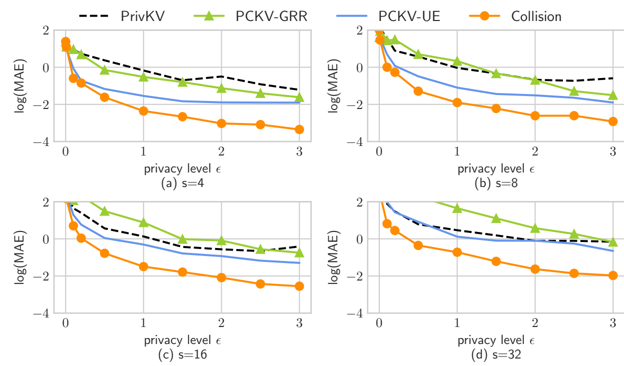

Figure 6: MAE results on $n = 10,000$ users with dimension $d = 256$ when sparsity s ranges from 4 to 32.

6.4 Effects of Number of Users n

Assume that there are only $n = 10,000$ users, and the dimension is $d = 256$. When the number of non-zero entries in numerical vectors varies from 4 to 32, the results of TVE and MAE are listed in Figure 5 and Figure 6 respectively. Compared to the case of $n = 100,000$ (i.e. Figure 1 and Figure 2), the TVE/MAE value is about $\sqrt{100000/10000}$ times larger (i.e. decreases with around $\sqrt{n}$).

6.5 After Shuffling

Considering the privacy budget is amplified in the shuffle model, typically when the number of users is $n = 100000$, privacy budget in CDP is $\epsilon_c = 0.5$ and $\delta = 1/n$, the local privacy budget in the Collision mechanism is scaled to around 2.0 with optimal t. In these region with large local privacy budget, the performance of the Collision is far better than other approaches. Besides, the privacy amplification of PrivKV/PCKV scales poorly with dependance on the domain size, thus when CDP budget (ϵ_c, δ) is given and the domain is relatively large, the performance gap grows.

6.6 Experimental Summary

Through experimental evaluation, we can conclude that the Collision mechanism outperforms existing approaches in all cases. Their performance gaps also support our previous theoretical analysis on error bounds (MAE errors usually have magnitude proportional to the root of mean squared error).

7 Conclusion

Within the framework of distributed differential privacy, this paper studied the problem of numerical vector statistical estimation, which has its applications in federated learning and key-value data aggregation. We provided tight minimax error bounds of $O(\frac{ds}{n\epsilon^2})$ for local differential private mean estimation on numerical vectors. Our proof relies on a novel decomposition technique for data domain with sparse structure and an application of the local private version of Assouad methods. Given that existing approaches are suffering gaps form the optimal error bounds, we further design an optimal mechanism for the problem, and then give an efficient implementation with linear computation/communication complexity. To further break the error bounds, we consider numerical vector estimation in the shuffled differential privacy, and show the proposed mechanism has the advantages of domain-independent privacy amplification and near-optimal utility. Experimental results show averagely 60% error reduction of the optimal mechanism when compared with current approaches.

Acknowledgements

We would like to thank anonymous reviewers for their helpful comments. This work is supported by National Key Project of China (No. 2020YFB1005700), National Natural Science Foundation of China for Joint Fund Project (No. U1936218), Project of Guangzhou University (No. RQ2021006), and the Pazhou lab, Guangzhou, China.

References

[Balle *et al.*, 2019] Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. *CRYPTO*, 2019.

[Balle *et al.*, 2020] Borja Balle, James Bell, Adria Gascon, and Kobbi Nissim. Private summation in the multi-message shuffle model. *CCS*, 2020.

[Błasiok *et al.*, 2019] Jaroslaw Błasiok, Mark Bun, Aleksandar Nikolov, and Thomas Steinke. Towards instance-optimal private query release. *SODA*, 2019.

[Cheu *et al.*, 2019] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. *EUROCRYPT*, 2019.

[Ding *et al.*, 2017] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *NeurIPS*, 2017.

[Duchi *et al.*, 2013] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. *FOCS*, 2013.

[Duchi *et al.*, 2018] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 2018.

[Dwork, 2008] Cynthia Dwork. Differential privacy: A survey of results. *International Conference on Theory and Applications of Models of Computation*, pages 1–19, 2008.

[Erlingsson *et al.*, 2014] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. *CCS*, 2014.

[Erlingsson *et al.*, 2019] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. *SODA*, 2019.

[Ghazi *et al.*, 2020] Badih Ghazi, Pasin Manurangsi, Rasmus Pagh, and Ameya Velingker. Private aggregation from fewer anonymous messages. *EUROCRYPT*, 2020.

[Greenberg, 2016] Andy Greenberg. Apple's 'differential privacy'is about collecting your data–but not your data. *Wired (June 13, 2016)*, 2016.

[Gu *et al.*, 2020] Xiaolan Gu, Ming Li, Yueqiang Cheng, Li Xiong, and Yang Cao. PCKV: Locally differentially private correlated key-value data collection with optimized utility. *USENIX Security*, 2020.

[Kairouz *et al.*, 2016] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. *ICML*, 2016.

[Konečnỳ *et al.*, 2016] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[Liu *et al.*, 2021] Ruixuan Liu, Yang Cao, Hong Chen, Ruoyang Guo, and Masatoshi Yoshikawa. Flame: Differentially private federated learning in the shuffle model. *AAAI*, 2021.

[McSherry and Talwar, 2007] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. *FOCS*, 2007.

[Nguyên *et al.*, 2016] Thông T Nguyên, Xiaokui Xiao, Yin Yang, Siu Cheung Hui, Hyejin Shin, and Junbum Shin. Collecting and analyzing data from smart device users with local differential privacy. *arXiv preprint arXiv:1606.05053*, 2016.

[Sun *et al.*, 2019] Lin Sun, Jun Zhao, Xiaojun Ye, Shuo Feng, Teng Wang, and Tao Bai. Conditional analysis for key-value data with local differential privacy. *arXiv preprint arXiv:1907.05014*, 2019.

[Sun *et al.*, 2020] Lin Sun, Xiaojun Ye, Jun Zhao, Chenhui Lu, and Mengmeng Yang. Bisample: Bidirectional sampling for handling missing data with local differential privacy. *arXiv preprint arXiv:2002.05624*, 2020.

[Voigt and Von dem Bussche, 2017] Paul Voigt and Axel Von dem Bussche. *The EU General Data Protection Regulation (GDPR)*, volume 18. Springer, 2017.

[Wang and Carreira-Perpinán, 2013] Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

[Wang and Xu, 2019] Di Wang and Jinhui Xu. Lower bound of locally differentially private sparse covariance matrix estimation. *IJCAI*, 2019.

[Wang *et al.*, 2019] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and analyzing multidimensional data with local differential privacy. *ICDE*, 2019.

[Wang *et al.*, 2020a] Tianhao Wang, Bolin Ding, Min Xu, Zhicong Huang, Cheng Hong, Jingren Zhou, Ninghui Li, and Somesh Jha. Improving utility and security of the shuffler-based differential privacy. *VLDB*, 2020.

[Wang *et al.*, 2020b] Tianhao Wang, Z Li, N Li, M Lopuhaä-Zwakenberg, and B Skoric. Locally differentially private frequency estimation with consistency. *NDSS*, 2020.

[Wangni *et al.*, 2018] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. *NeurIPS*, 2018.

[Wen *et al.*, 2017] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *NeurIPS*, 2017.

[Ye *et al.*, 2019] Qingqing Ye, Haibo Hu, Xiaofeng Meng, and Huadi Zheng. Privkv: Key-value data collection with local differential privacy. *IEEE S&P*, 2019.

[Yu, 1997] Bin Yu. Assouad, fano, and le cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1997.

[Zhu *et al.*, 2019] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *NeurIPS*, 2019.