Instance-optimal Mean Estimation Under Differential Privacy

Ziyue Huang, Yuting Liang, Ke Yi

{zhuangbq,yliangbs,yike}@cse.ust.hk
Department of Computer Science and Engineering
Hong Kong University of Science and Technology

Abstract

Mean estimation under differential privacy is a fundamental problem, but worst-case optimal mechanisms do not offer meaningful utility guarantees in practice when the global sensitivity is very large. Instead, various heuristics have been proposed to reduce the error on real-world data that do not resemble the worst-case instance. This paper takes a principled approach, yielding a mechanism that is instance-optimal in a strong sense. In addition to its theoretical optimality, the mechanism is also simple and practical, and adapts to a variety of data characteristics without the need of parameter tuning. It easily extends to the local and shuffle model as well.

1 Introduction

Mean estimation is one of the most fundamental problems in statistics, optimization, and machine learning. However, privacy concerns forbid us from using the exact mean in these applications, and the problem of how to achieve the smallest error under a given privacy model has received considerable attention in the literature. *Differential privacy* (DP) is a rigorous mathematical definition for protecting individual privacy and has emerged as the golden standard in privacy-preserving data analysis nowadays, which has been deployed by Apple [18], Google [24], and Microsoft [19].

Given a data set $\mathcal{D} := \{x_i\}_{i\in[n]} \subset \mathcal{U}^d$, where $\mathcal{U} = [u]$, i.e., each coordinate of the input vector is an integer (real-valued coordinates can be handled by quantization; see remark 3.2), our goal is to obtain a differentially private estimation $M(\mathcal{D})$ for the mean $f(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} x_i$ with small ℓ_2 error $\|M(\mathcal{D}) - f(\mathcal{D})\|_2$. Because $f(\cdot)$ has global ℓ_2 sensitivity $\mathrm{GS}_f = \sqrt{d}u/n$, the standard DP mechanism just adds Gaussian noise scaled to GS_f to each coordinate of $f(\mathcal{D})$, which results in an ℓ_2 error proportional to du/n. This simple mechanism is worst-case optimal [31], but it is certainly undesirable in practice, as people often conservatively use a large u (e.g., $u = 2^{32}$) but the actual dataset $\mathcal{D}$ may have much smaller coordinates. Instead, the *clipped-mean estimator* [1] (see Section 3.1 for details) has been widely used as an effective heuristic, but two questions remain unresolved: (1) how to choose the clipping threshold C; and (2) if it can yield any optimality guarantees. We answer these questions in a fairly strong sense in this paper.

1.1 Instance Optimality

As worst-case optimality is theoretically trivial and practically meaningless for the mean estimation problem when the global sensitivity is too large, one may aim at instance optimality. More precisely, let $\mathcal{M}$ be the class of DP mechanisms and let

$$\mathcal{R}_{\mathrm{ins}}(\mathcal{D}) := \inf_{M'\in\mathcal{M}} \inf\{\xi \mid \Pr[\|M'(\mathcal{D}) - f(\mathcal{D})\|_2 \le \xi] \ge 2/3\}$$

be the smallest error any M' can achieve (with constant probability) on $\mathcal{D}$, then the standard definition of instance optimality requires us to design an M such that

$$\Pr[\|M(\mathcal{D}) - f(\mathcal{D})\|_2 \leq c \cdot \mathcal{R}_{\text{ins}}(\mathcal{D})] \geq 2/3 \tag{1}$$

for every $\mathcal{D}$, where c is called the *optimality ratio*. Unfortunately, for any $\mathcal{D}$, one can design a trivial $M'(\cdot) \equiv f(\mathcal{D})$ that has 0 error on $\mathcal{D}$ (but fails miserably on other instances), so $\mathcal{R}_{\text{ins}}(\cdot) \equiv 0$, which rules out instance-optimal DP mechanisms by a standard argument [23].

Since $\mathcal{R}_{\text{ins}}(\cdot)$ is unachievable, relaxed versions can be considered. The above trivial M' exists because it is only required to work well on one instance $\mathcal{D}$. Imposing higher requirements on M' would yield relaxed notions of instance optimality. One natural requirement is that M' should work well not just on $\mathcal{D}$, but also on its neighbors, i.e., we raise the target error from $\mathcal{R}_{\text{ins}}(\mathcal{D})$ to

$$\mathcal{R}_{\text{nbr}}(\mathcal{D}) := \inf_{M' \in \mathcal{M}} \sup_{\mathcal{D}' : d_{\text{ham}}(\mathcal{D}, \mathcal{D}') \leq 1} \inf\{\xi \mid \Pr[\|M'(\mathcal{D}') - f(\mathcal{D}')\|_2 \leq \xi] \geq 2/3\}.$$

Vahdan [40] observes that $\mathcal{R}_{\text{nbr}}(\mathcal{D})$ is exactly $\text{LS}_f(\mathcal{D})$, the *local sensitivity* of f at $\mathcal{D}$, up to constant factors. However, $\text{LS}_f(\cdot)$ may not be an appropriate target to shoot at, depending on what f is. For the MEDIAN problem, $\text{LS}_f(\mathcal{D}) = 0$ for certain $\mathcal{D}$'s and no DP mechanisms can achieve this error [37], while for mean estimation, $\text{LS}_f(\mathcal{D}) = \Theta(\text{GS}_f) = \Theta(\sqrt{d}u/n)$ for all $\mathcal{D}$, so this relaxation turns instance optimality into worst-case optimality.

The reason why the above relaxation is "too much" for the mean estimation problem is that $\mathcal{D}'$ may change one vector of $\mathcal{D}$ *arbitrarily*, e.g., from $(0, \dots, 0)$ to $(u, \dots, u)$. We restrict this. More precisely, letting $\text{supp}(\mathcal{D})$ denote the set of distinct vectors in $\mathcal{D}$, we consider the target error

$$\mathcal{R}_{\text{in-nbr}}(\mathcal{D}) := \inf_{M' \in \mathcal{M}} \sup_{\mathcal{D}' : d_{\text{ham}}(\mathcal{D}, \mathcal{D}') \leq 1, \text{supp}(\mathcal{D}') \subseteq \text{supp}(\mathcal{D})} \inf\{\xi \mid \Pr[\|M'(\mathcal{D}') - f(\mathcal{D}')\|_2 \leq \xi] \geq 2/3\},$$

namely, we require M' to work well only on $\mathcal{D}$ and its *in-neighbors*, in which a vector can only be changed to another one already existing in $\mathcal{D}$. Correspondingly, an instance-optimal M (w.r.t. the in-neighborhood) is one such that (1) holds where $\mathcal{R}_{\text{ins}}$ is replaced by $\mathcal{R}_{\text{in-nbr}}$.

We make a few notes on this notion of instance optimality: (1) This optimality is only about the utility of the mechanism, not its privacy. We still require the mechanism to satisfy the DP requirement between *any* $\mathcal{D}, \mathcal{D}'$ such that $d_{\text{ham}}(\mathcal{D}, \mathcal{D}') = 1$, not necessarily one and its in-neighbors. (2) In general, a smaller neighborhood leads to a stronger notion of instance optimality. Thus, the optimality using in-neighbors is stronger than that using all neighbors, which is in turn stronger than worst-case optimality (i.e., $\mathcal{D}'$ can be any instance), while the latter two are actually the same for the mean estimation problem. (3) For an instance-optimal M (by our notion), there still exist $\mathcal{D}, M'$ such that M' does better on $\mathcal{D}$ than M, but it is not possible for M' to achieve a smaller error than the error of M on $\mathcal{D}$ over all in-neighbor of $\mathcal{D}$. This is more meaningful than ranging over all neighbors of $\mathcal{D}$, some of which (e.g., one with $(u, \dots, u)$ as a datum) are unlikely to be the actual instances encountered in practice.

1.2 Our Results

To design an $M(\mathcal{D})$ for the mean function $f(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} x_i$ that achieves an error w.r.t. $\mathcal{R}_{\text{in-nbr}}(\mathcal{D})$ for all $\mathcal{D}$, we need an upper bound and a lower bound. For the lower bound, we show that $\mathcal{R}_{\text{in-nbr}}(\mathcal{D}) = \Omega(w(\mathcal{D})/n)$, where $w(\mathcal{D}) := \max_{1 \leq i < j \leq n} \|x_i - x_j\|_2$ is the *diameter* of $\mathcal{D}$. Thus, from the upper bound side, it suffices to show that the mechanism's error is bounded by $c \cdot w(\mathcal{D})/n$. This is achieved in two steps. First, we use the clipped-mean estimator, but find the clipping threshold C that optimizes its bias-variance trade-off, which is a certain quantile of the norms of the vectors in $\mathcal{D}$. However, we cannot use the optimal C directly, as it would violate DP. Thus, we use a simple binary search based algorithm that can find any specific quantile privately with an optimal rank error. This results in a DP mechanism with error $\tilde{O}(\sqrt{d/\rho}) \cdot r(\mathcal{D})/n$, where $r(\mathcal{D}) := \max_i \|x_i\|_2$ and ρ is the privacy parameter (formal definition given in Section 2). To reduce the error from $r(\mathcal{D})$ to $w(\mathcal{D})$, in the second step, we rotate and shift $\mathcal{D}$ into a $\tilde{\mathcal{D}}$ such that $r(\tilde{\mathcal{D}}) = O(w(\mathcal{D}))$ w.h.p., and apply the clipped-mean estimator (with our privatized optimal clipping threshold) on $\tilde{\mathcal{D}}$, leading to an error of $\tilde{O}(\sqrt{d/\rho}) \cdot w(\mathcal{D})/n$ for $n = \tilde{\Omega}(\sqrt{d/\rho})$. We also show that the optimality ratio $c = \tilde{O}(\sqrt{d/\rho})$ is optimal, i.e., any mechanism $M(\mathcal{D})$ having error $c \cdot w(\mathcal{D})/n$ for all $\mathcal{D}$ must have $c = \tilde{\Omega}(\sqrt{d/\rho})$ for $\rho < \tilde{O}(\sqrt{d/n})$.

Our mechanism can be applied directly to *statistical mean estimation*, where the vectors in $\mathcal{D}$ are i.i.d. samples from a certain distribution and one would like to estimate the mean of the distribution (in contrast, the version defined above is referred to as *empirical mean estimation*). For concreteness, we show how this is done for the multivariate Gaussian distributions $\mathcal{N}(\mu, \Sigma)$. For the case $\Sigma = \mathbf{I}$, our algorithm achieves an ℓ_2 error of α using $n = \tilde{O}(\frac{d}{\alpha^2} + \frac{d}{\alpha\sqrt{\rho}})$ samples for $\alpha \leq O(1)$, matching the optimal bound in the statistical setting [9]. For a non-identity, unknown Σ, the ℓ_2 error is proportional to $\|\Sigma^{1/2}\|_F$, the same as in [9]. However, our mechanism requires only crude *a priori* bounds on μ and Σ (i.e., the error depends on these bounds logarithmically), while [9] needs a constant-factor approximation of Σ, which can be obtained using $n = \tilde{\Omega}(d^{3/2}/\sqrt{\rho})$ samples [29]. Note that this can be a $\sqrt{d}$-factor higher than the sample complexity of mean estimation. Fundamentally, estimating Σ is harder than estimating μ, and we bypass the former so as to retain the linear dependency on d. In practice, estimating Σ first would consume the privacy budget from the mean estimation problem itself. On the other hand, the benefit of estimating Σ first is that one can obtain an error guarantee under the Mahalanobis distance [29], which cannot be achieved by our method. Interestingly, [10] very recently shows how to achieve the same Mahalanobis error guarantee without estimating Σ, thereby retaining the linear dependency on d, but unfortunately the running time is exponential. It is an interesting open problem whether one can achieve a constant Mahalanobis error with $\tilde{O}(d)$ samples in polynomial time.

Furthermore, by simply changing the primitive operations, our mechanism easily extends to the local and shuffle model of differential privacy. In doing so, we also extend the one-dimensional summation/mean estimation protocol in the shuffle model [6, 27] to high dimensions.

In addition to the theoretical optimality, our mechanism is also simple and practical. Most importantly, there is no (internal) parameter to tune. Yet, our experimental results demonstrate that our mechanism outperforms the state-of-the-art algorithm [9] with the best parameters tuned for each specific setting.

1.3 Related Work

Asi and Duchi [5] recently initialized the study on instance optimality under DP. They propose two ways to relax (equivalently, strengthen the requirement on M') the strict instance optimality, which is unachievable. The first is to require M' to be unbiased. This is not appropriate for mean estimation, since many estimators, including clipped-mean, is not unbiased. The second is to require M' to work well over all the r-distance neighbors of $\mathcal{D}$ for $r \geq 1$. Thus, their optimality is weaker than using $\mathcal{R}_{\text{nbr}}(\cdot)$, hence not appropriate for the mean estimation problem (i.e., their optimality is the same as worst-case optimality). Instance optimality has not been studied in the local or shuffle model; existing protocols in these two models [8, 21, 6] all have errors proportional to the global sensitivity.

How to choose the clipping threshold C for the clipped mean estimator has been extensively studied [3, 4, 38, 36], but existing methods do not offer any optimality guarantees. In particular, Andrew et al. [4] also use a quantile (actually, median) as C, but as we shall see, median is actually not the optimal choice. Furthermore, they use online gradient descent to find a privatized quantile, which does not have any theoretical error guarantees. Amin et al. [3] attempt to select an optimal quantile as the clipping threshold to truncate the number of contributions from each user, instead of clipping the actual samples in high dimensions as in our paper.

In the statistical setting, where the data are i.i.d. samples from some specific distribution, there are numerous methods [29, 9, 30, 32] that can avoid an error proportional to the global sensitivity, by exploiting the concentration property of the distribution. In particular, Biswas et al. [9] provide a simple and practical mechanism for multivariate Gaussian data. Levy et al. [34] propose a private mean estimator with error scaling with the concentration radius τ of the distribution rather than the entire range, but their algorithm requires τ to be publicly known in advance. In the local model, the algorithm in [25] uses a quantile estimation procedure based on binary search as a subroutine for one-dimensional Gaussian data.

Very recently, the relationship between the error of mean estimation and the diameter of the dataset has been exploited in [17] for low-communication protocols, but they do not consider privacy. Our DP protocols in the local and shuffle models have communication cost $\tilde{O}(d)$ per user (we do not state the communication costs in the theorems as they are not our major concern); it would be interesting to see if ideas from [17] can be used to reduce it further.

2 Preliminaries

2.1 Differential Privacy in the Central Model

Definition 2.1 (Differential Privacy (DP) [23]). For $\varepsilon > 0$ and $\delta \geq 0$, a randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-differentially private if for any neighboring datasets $\mathcal{D} \sim \mathcal{D}'$ (i.e., $d_{\text{ham}}(\mathcal{D}, \mathcal{D}') = 1$) and any $E \subseteq \mathcal{Y}$,

$$\mathsf{Pr}[M(\mathcal{D}) \in E] \leq e^{\varepsilon} \cdot \mathsf{Pr}[M(\mathcal{D}') \in E] + \delta.$$

Definition 2.2 (Concentrated Differential Privacy (zCDP) [12]). For $\rho > 0$, a randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is ρ-zCDP if for any $\mathcal{D} \sim \mathcal{D}'$,

$$D_{\alpha}(M(\mathcal{D}) || M(\mathcal{D}')) \leq \rho\alpha$$

for all $\alpha > 1$, where $D_{\alpha}(M(\mathcal{D}) || M(\mathcal{D}'))$ is the α-Rényi divergence between $M(\mathcal{D})$ and $M(\mathcal{D}')$.

Note that $(\varepsilon, 0)$-DP implies $\frac{\varepsilon^2}{2}$-zCDP, which implies $(\frac{\varepsilon^2}{2} + \varepsilon\sqrt{2 \log \frac{1}{\delta}}, \delta)$-DP for any $\delta > 0$. To release a numeric function $f(\mathcal{D})$ taking values in $\mathbb{R}^d$, the most common technique for achieving zCDP is by masking the result with Gaussian noise calibrated to the ℓ_2-sensitivity of f.

Lemma 2.1 (Gaussian Mechanism [12]). *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be a function with global ℓ_2-sensitivity $\mathrm{GS}_f := \max_{\mathcal{D} \sim \mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_2$. For a given data set $\mathcal{D} \in \mathcal{X}^n$, the mechanism that releases $f(\mathcal{D}) + \mathcal{N}\left(0, \frac{\mathrm{GS}_f^2}{2\rho} \cdot I_{d \times d}\right)$ satisfies ρ-zCDP.*

Lemma 2.2 (Composition Theorem [12, 23]). *If M is an adaptive composition of differentially private algorithms $M_1, M_2, \dots, M_k$, then*

1. *If each M_i satisfies $(\varepsilon_i, \delta_i)$-DP, then M satisfies $(\sum_i \varepsilon_i, \sum_i \delta_i)$-DP.*

2. *For all $\varepsilon, \delta, \delta' \geq 0$, if each M_i satisfies (ε, δ)-DP, then M satisfies $(\varepsilon', k\delta + \delta')$-DP, where $\varepsilon' = \sqrt{2k \log \frac{1}{\delta'}}\varepsilon + k\varepsilon(e^{\varepsilon} - 1)$.*

3. *If each M_i satisfies ρ_i-zCDP, then M satisfies $(\sum_i \rho_i)$-zCDP.*

2.2 Differential Privacy in the Local Model and Shuffle Model

The above definitions of DP and zCDP assume that $\mathcal{D}$ is handled by a trusted curator and only the output of the mechanism will be released to the public. Therefore, if the curator is corrupted, the privacy of all users will be breached. For weaker trust assumptions, the most popular models are the local model and the shuffle model, where each user holds their datum and locally privatizes (by some randomized mechanism) the message before sending it out for analysis. Hence, there is no third-party who has direct access to $\mathcal{D}$. Formally, each user holds one datum $x_i \in \mathcal{D}$, and the protocol interacts with the dataset using some local randomizer $R : \mathcal{X} \to \mathcal{Y}$, and the privacy guarantee is defined over the transcript (all messages sent during the protocol). For simplicity, we only present the definition for one-round protocols; the privacy guarantee of multi-round protocols can be composed across all rounds by the composition theorem. The definition below uses zCDP; other DP notions can be defined similarly.

Definition 2.3 (Local Model (LDP))**.** A protocol using $R(\cdot)$ as the local randomizer satisfies ρ-zCDP in the local model if for any $x, x' \in \mathcal{X}$, any $\alpha > 1$, $D_\alpha(R(x)||R(x')) \le \rho\alpha$.

Due to the much stronger privacy requirement, the best accuracy guarantee of LDP protocols for several fundamental problems [14, 7, 20, 35] is a $\sqrt{n}$-factor worse than that in the central model. The shuffle model is established on an intermediary level of trust assumption between the local model and the central model and aims for obtaining errors closer to the central model. The key feature of the shuffle model is a trusted shuffler $\mathcal{S}$, which can permute all messages randomly before sending them to the analyzer, so that an adversary cannot identify the source of any message. Specifically, we consider the multi-message shuffle model, where each local randomizer $R : \mathcal{X} \to \mathcal{Y}^m$ outputs m messages, and the transcript of the protocol $\Pi_P(\mathcal{D})$ is a random permutation of all mn messages. The following definition uses (ε, δ)-DP; the other two DP notions can also be defined similarly, but they do not offer the improvements that we want over LDP protocols.

Definition 2.4 (Shuffle Model)**.** A protocol P satisfies (ε, δ)-DP in the shuffle model if for any $\mathcal{D} \sim \mathcal{D}'$, and any set $E \subseteq \mathcal{Y}^{mn}$, $\mathsf{Pr}[\Pi_P(\mathcal{D}) \in E] \le e^\varepsilon \cdot \mathsf{Pr}[\Pi_P(\mathcal{D}') \in E] + \delta$.

3 Our Method

3.1 Clipped-Mean Estimator

In the rest of the paper, we focus on the mean function $f(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^n x_i$. Since GS_f is large, a very natural idea is to clip each vector in its ℓ_2 norm by some threshold C. This reduces GS_f to $2C/n$, leading to the clipped-mean estimator [1]:

$$M_C(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^n \min\left\{\frac{C}{\|x_i\|_2}, 1\right\} \cdot x_i + \mathcal{N}\left(\mathbf{0}, \frac{2C^2}{\rho n^2}\mathbf{I}\right). \tag{2}$$

Lemma 3.1. *For any given C, $M_C(\mathcal{D})$ satisfies ρ-zCDP, and has an expected ℓ_2 error at most*

$$\mathsf{E}\left[\|M_C(\mathcal{D}) - f(\mathcal{D})\|_2\right] \le \mathcal{E}(C; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^n \max\{\|x_i\|_2 - C, 0\} + \frac{C}{n} \cdot \sqrt{\frac{2d}{\rho}}.$$

Proof. The privacy guarantee easily follows from Lemma 2.1. The error of $M_C(\mathcal{D})$ is composed of two parts: the bias from clipping and the (square root of the) variance from the Gaussian noise $\mathcal{N}(0, 2C^2/(\rho n) \cdot \mathbf{I})$. Because the ℓ_2 clipping does not change the direction of the input vector, the bias introduced by clipping is at most $\frac{1}{n}\sum_i \max\{\|x_i\|_2 - C, 0\}$. The variance introduced by the Gaussian noise is at most $C^2/n^2 \cdot 2d/\rho$ by Jensen inequality. □

An important remaining question is how to set the clipping threshold C. Setting it too low will result in a large bias, while setting it too high will introduce a large amount of noise. We show how to choose the optimal C to balance this bias-variance trade-off. It is easy to see that the error $\mathcal{E}(C; \mathcal{D})$ is a convex function w.r.t. C, thus the optimal C can be found by setting the derivative of $\mathcal{E}(C; \mathcal{D})$ to zero, i.e.,

$$\frac{\partial \mathcal{E}(C; \mathcal{D})}{\partial C} = \frac{1}{n} |\{i \in [n] \mid \|x_i\|_2 > C\}| - \frac{1}{n} \cdot \sqrt{2d/\rho} = 0.$$

Therefore, the optimal choice of C is the $(n - \sqrt{2d/\rho})$-th quantile of $\{\|x_i\|_2\}_{i \in [n]}$.

3.2 Private Quantile Selection

However, we cannot use the optimal C directly, as it would violate DP. Instead, we find a privatized quantile with small rank error. Specifically, for this problem, $\mathcal{D}$ consists of a sequence of ordered integers $0 \le x_{(1)} \le \dots \le x_{(n)} \le u$. We would like to design a DP mechanism that, for a given m, returns an x (which is not necessarily an element in $\mathcal{D}$) such that $x_{(m-\tau)} \le x \le x_{(m+\tau)}$[1] w.h.p. Here τ is referred to as the *rank error*. Existing methods on private range counting queries [13, 22] can be used for this purpose, but they actually find all quantiles, which is an overkill. Instead, we use a simple binary search algorithm [28, 16], which not only simplifies the algorithm, but also reduces the rank error (by polylog(u) factors) to nearly optimal. Our algorithm `PrivQuant` makes use of a function $\texttt{NoisyRC}([a, b], \mathcal{D})$ that returns a noisy count of $|\mathcal{D} \cap [a, b]|$.

Algorithm 1 DP Quantile Selection by Binary Search; `PrivQuant`

```
Input: the data set $\mathcal{D} : 0 \le x_{(1)} \le \dots \le x_{(n)} \le u$; $m \in [n]$.
Output: a DP approximation to $x_{(m)}$.
left ← 0, right ← $u$
while left < right do
    mid ← $\lfloor(\text{left} + \text{right})/2\rfloor$
    $\tilde{c}$ ← NoisyRC([0, mid], $\mathcal{D}$)
    if $\tilde{c} \le m$ then
        left ← mid + 1
    else
        right ← mid
return $\lfloor(\text{left} + \text{right})/2\rfloor$
```

The following lemma is straightforward:

Lemma 3.2. *If* $|\texttt{NoisyRC}([0, mid], \mathcal{D}) - |\mathcal{D} \cap [0, mid]|| \le \tau$ *for every call to* $\texttt{NoisyRC}([0, mid], \mathcal{D})$*, then Algorithm 1 returns a quantile with rank error* τ.

In the central DP model, we simply use $\texttt{NoisyRC}([0, \text{mid}], \mathcal{D}) = |\mathcal{D} \cap [0, \text{mid}]| + \mathcal{N}(0, \log u/(2\rho))$.

Theorem 3.3. *The algorithm* `PrivQuant` *preserves* ρ*-CDP, and it returns a quantile with rank error* τ *with probability at least* $1 - \beta$ *for* $\tau = \sqrt{\log u \log \frac{\log u}{\beta} / (2\rho)}$.

Proof. It is clear that the range query $|[0, \text{mid}] \cap \mathcal{D}|$ has sensitivity 1, thus adding noise drawn from $\mathcal{N}(0, \log u/(2\rho))$ preserves $\frac{\rho}{\log u}$-CDP for each invocation. Because there are $\log u$ iterations in the while-loop, the privacy guarantee follows from the composition theorem of CDP.

[1] Define $x_{(j)} = 0$ for $j < 1$ and $x_{(j)} = u$ for $j > n$.

In the algorithm, we draw at most $\log u$ Gaussian noises whose absolute values are simultaneously bounded by τ with probability $1-\beta$ by a union bound. Conditioned upon this event, the theorem follows from Lemma 3.2. □

In Section 4, we prove an $\Omega(\sqrt{\log u/\rho})$ lower bound (Corollary 4.7) on the rank error under zCDP for constant β. Thus the algorithm is optimal up to just an $O(\sqrt{\log\log u})$-factor.

We can now use `PrivQuant` to find an approximately optimal clipping threshold. Specifically, we invoke `PrivQuant` with $\rho' = \rho/4$ to find the $\max\{n - \max\{\sqrt{2d/\rho}, \tau\}, 1\}$-th quantile of $\{\|x_i\|_2^2\}_{i\in[n]}$. They are integers no more than du^2, so replacing u by du^2 in Theorem 3.3 yields a rank error of $\tau = 2\sqrt{\log(du)\log\frac{\log(du)}{\beta}/\rho}$. Then we set $\tilde{C}$ as the square root of the returned quantile. Finally, we return the clipped mean estimator $M_{\tilde{C}}(\mathcal{D})$ with $\rho' = 3\rho/4$. The following theorem analyzes its error.

Theorem 3.4. *Our mean estimation mechanism is ρ-zCDP and has ℓ_2 error $O(\sqrt{d/\rho}+\tau)\cdot r(\mathcal{D})/n$ with probability $1-\beta$, where $\tau = 2\sqrt{\log(du)\log\frac{\log(du)}{\beta}/\rho}$.*

Proof. The privacy guarantee easily follows from the composition theorem of zCDP. Next, we analyze the accuracy. By the rank error guarantee, at most $\sqrt{2d/\rho}+\tau$ vectors are clipped by the threshold $\tilde{C}$. Each clipped vector has norm at most $r(\mathcal{D})$, so the bias is at most $(\sqrt{2d/\rho}+\tau)\cdot r(\mathcal{D})/n$. For the error due to the noise, we use the following tail bound of the multivariate Gaussian distribution:

Lemma 3.5 ([33])**.** *If $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then* $\Pr\left[\|X\|_2 \geq \sqrt{d + 2\sqrt{d\log(1/\beta)} + 2\log(1/\beta)}\right] \leq \beta$.

Thus, with probability $1-\beta$, the norm of the noise is bounded by $O\left(\sqrt{d+\log\frac{1}{\beta}}\cdot\frac{\tilde{C}}{n\sqrt{\rho}}\right) \leq O\left(\sqrt{d/\rho}+\tau\right)\cdot r(\mathcal{D})/n$. This inequality requires $\tilde{C} \leq r(\mathcal{D})$, which holds as long as $n > \max\{\sqrt{2d/\rho}, \tau\}$. If this is not the case (note that checking this condition is DP as it does not involve $\mathcal{D}$), we can just return $\mathbf{0}$, which trivially achieves error $r(\mathcal{D}) \leq O(\sqrt{d/\rho}+\tau)\cdot r(\mathcal{D})/n$. □

3.3 Shifted-Clipped-Mean Estimator

To reduce the error from being proportional to $r(\mathcal{D})$ to being proportional to $w(\mathcal{D})$, we perform a random rotation on $\mathcal{D}$ followed by a translation. The rotation is done by $\hat{x}_i := HDx_i$, where H is the Hadamard matrix, D is a diagonal matrix whose diagonal entry is independently and uniformly drawn from $\{-1,+1\}$. Note that for now we omit the normalization coefficient $\frac{1}{\sqrt{d}}$ so that each coordinate of $\hat{x}_i$ is still an integer; we will apply the normalization to the final estimator instead. Then, for each $j \in [d]$, we invoke `PrivQuant` with $\rho' = \rho/(4d)$ to find an approximate median of $\{\hat{x}_i\}_{i\in[n]}$ along dimension j, denoted as $\tilde{c}_j$. Next, we shift the dataset to be centered around $\tilde{c} = (\tilde{c}_1, \dots, \tilde{c}_d)$, obtaining $\tilde{\mathcal{D}} = \{\tilde{x}_i := \hat{x}_i - \tilde{c}\}_{i\in[n]}$. Note that $\tilde{c}$ has integer coordinates, so does $\tilde{x}_i$. Finally, we apply the clipped-mean estimator in Theorem 3.4 with $\rho' = \frac{3}{4}\rho$ on $\tilde{\mathcal{D}}$, obtaining an estimation $\tilde{y}$, and return $y := (\frac{1}{\sqrt{d}}HD)^{-1}\frac{1}{\sqrt{d}}(\tilde{y}+\tilde{c})$ as the mean estimator over $\mathcal{D}$.

Theorem 3.6. *Set $\tau = \sqrt{\log(du)\log\frac{d\log(du)}{\beta}/\rho}$ and assume $n = \Omega(\tau\sqrt{d})$. Our mean estimation mechanism is ρ-zCDP, and has ℓ_2 error $O\left((\sqrt{d/\rho}+\tau)\sqrt{\log\frac{nd}{\beta}}\right)\cdot w(\mathcal{D})/n$ with probability $1-\beta$.*

Proof. The privacy guarantee follows from the composition theorem of ρ-zCDP, as $\sum_{j=1}^{d}\rho/(4d) + 3\rho/4 = \rho$. Next, we analyze the error. We need a lemma from [2], which intuitively says that the random rotation "evenly spreads out" the norm to all the dimensions:

Lemma 3.7 ([2]). *Let H and D be defined as above. Then, for any $x \in \mathbb{R}^d$ and any $\beta > 0$,*

$$\Pr\left[\left\|\frac{1}{\sqrt{d}}HDx\right\|_\infty \geq \frac{\|x\|_2}{\sqrt{d}} \cdot \sqrt{2\log\frac{4d}{\beta}}\right] \leq \beta.$$

Moreover, note that the transformation by $\frac{1}{\sqrt{d}}HD$ or $(\frac{1}{\sqrt{d}}HD)^{-1}$ is orthogonal, so the ℓ_2 norm of any vector will be preserved.

Applying Lemma 3.7 on $x_i - x_j$ for all $i, j \in [n]$ and a union bound, we have $\max_{i,j} \|\hat{x}_i - \hat{x}_j\|_\infty = O(\sqrt{\log\frac{nd}{\beta}}) \cdot w(\mathcal{D})$ with probability $1 - \beta/3$. Over the rotated dataset $\{\hat{x}_i\}_{i\in[n]}$, we use `PrivQuant` to find an approximate median $\tilde{c}_j$ along each dimension $j \in [d]$ with privacy parameter $\rho' = \rho/(4d)$. By the rank error guarantee of `PrivQuant` (Theorem 3.3) and a union bound, if $n = \Omega(\tau\sqrt{d})$, we have $\min_i \hat{x}_{i,j} \leq \tilde{c}_j \leq \max_i \hat{x}_{i,j}$ for all $j \in [d]$ with probability $1 - \beta/3$. Note that the length of this interval is $|\min_i \hat{x}_{i,j} - \max_i \hat{x}_{i,j}| = O(\sqrt{\log\frac{nd}{\beta}}) \cdot w(\mathcal{D})$. Thus the region $(\tilde{c}_1 \pm O(\sqrt{\log\frac{nd}{\beta}}) \cdot w(\mathcal{D}), \dots, \tilde{c}_d \pm O(\sqrt{\log\frac{nd}{\beta}}) \cdot w(\mathcal{D}))$ contains every data point $\hat{x}_i$, hence $\max_i \|\hat{x}_i - \tilde{c}\|_2 = O(\sqrt{d\log\frac{nd}{\beta}}) \cdot w(\mathcal{D})$. This means that the shifted data set $\tilde{\mathcal{D}} = \{\hat{x}_i - \tilde{c}\}_{i\in[n]}$ has $r(\tilde{\mathcal{D}}) = O(\sqrt{d\log\frac{nd}{\beta}}) \cdot w(\mathcal{D})$. Thus, when we apply the clipped-mean estimator in Theorem 3.4 over $\tilde{\mathcal{D}}$ to obtain its mean estimation $\tilde{y}$, we have $\|(\tilde{y}+\tilde{c}) - \frac{1}{n}\sum_i \hat{x}_i\|_2 = O((\sqrt{d/\rho} + \tau)\sqrt{d\log\frac{nd}{\beta}}) \cdot w(\mathcal{D})/n$. Finally, we use $y = (\frac{1}{\sqrt{d}}HD)^{-1}\frac{1}{\sqrt{d}}(\tilde{y} + \tilde{c})$ as the mean estimation for $\mathcal{D}$, and conclude that

$$\left\|y - \frac{1}{n}\sum_i x_i\right\|_2 = \left\|\left(\frac{1}{\sqrt{d}}HD\right)^{-1}\frac{1}{\sqrt{d}} \cdot \left((\tilde{y} + \tilde{c}) - \frac{1}{n}\sum_i \hat{x}_i\right)\right\|_2 = O\left(\left(\sqrt{d/\rho} + \tau\right)\sqrt{\log\frac{nd}{\beta}}\right) \cdot \frac{w(\mathcal{D})}{n}.$$

□

Remark 3.1. The Hadamard transform requires d to be some power of 2. If this is not the case, we can pad each x_i with extra 0's to dimension $\bar{d} = 2^{\lceil\log d\rceil}$, denoted as $\bar{x}_i$. If there is an estimation $\bar{y}$ for $\frac{1}{n}\sum_i \bar{x}_i$, we discard the last $\bar{d} - d$ coordinates of $\bar{y}$ to obtain y as the estimation for $\frac{1}{n}\sum_i x_i$. Then, we have $\|y - \sum_i x_i/n\|_2 \leq \|\bar{y} - \sum_i \bar{x}_i/n\|_2$, since the last $\bar{d} - d$ coordinates of each $\bar{x}_i$ are 0. The padding does not change $w(\mathcal{D})$, so Theorem 3.6 still holds.

Remark 3.2. For a dataset with real coordinates bounded by R (in absoluate value), one can quantize each coordinate to an integer using bucket size $\alpha/\sqrt{d}$, for any $0 < \alpha < R$, and then apply our algorithm over an integer universe of size $u = 2R\sqrt{d}/\alpha$. This just brings an additive α error to the error bound of Theorem 3.6.

3.4 Statistical Mean Estimation

Suppose $\mathcal{D}$ consists of i.i.d. samples drawn from the multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, and we wish to estimate μ, assuming *a priori* bounds $\|\mu\|_2 \leq R$ and $\sigma_{\min}^2\mathbf{I} \preceq \Sigma \preceq \sigma_{\max}^2\mathbf{I}$. Note that in the statistical setting, the privacy requirement should be satisfied between any two neighboring instances (not i.i.d.), but utility is analyzed under the i.i.d. assumption.

We first clip each sample $x_i \leftarrow x_i \cdot \min\{R'/\|x_i\|_2, 1\}$ where $R' := R + 2\sigma_{\max}\sqrt{d + \log\frac{4n}{\beta}}$. Then all coordinates are bounded by R' and we apply our mechanism with bucket size $\alpha/\sqrt{d}$ where $\alpha = \sigma_{\min}\sqrt{d/n}$. Privacy is straightforward, since two instances are neighbors after the R'-clipping only if they are neighbors before the clipping. The error of the estimator will depend on $w(\mathcal{D})$, which can be easily bounded by $\tilde{O}(\|\Sigma^{1/2}\|_2 \cdot \sqrt{d})$ via standard matrix analysis. Below we give a tighter bound using the Hanson-Wright inequality:

Lemma 3.8 (Hanson-Wright inequality [39]). *Let A be any $d \times d$ matrix. Consider a random vector $X = (X_1, X_2, \ldots, X_d)$ where the X_i's are independent random variables drawn from $\mathcal{N}(0,1)$. Then for any $t > 0$, we have*

$$\Pr\left[|\|AX\|_2 - \|A\|_F| > t\right] \le 2\exp\left(-\frac{ct^2}{\|A\|_2^2}\right),$$

where c is an absolute constant.

Corollary 3.9. *Set $\tau = \sqrt{\log(du)\log\frac{4d\log(du)}{\beta}/\rho}$ where $u = 2R'\sqrt{n}/\sigma_{\min}$, and assume $n = \Omega(\tau\sqrt{d})$. Then our algorithm returns a $\hat{\mu}$ such that with probability $1-\beta$,*

$$\|\hat{\mu}-\mu\|_2 = O\left(\left(\|\Sigma^{1/2}\|_F + \|\Sigma^{1/2}\|_2 \cdot \sqrt{\log\frac{n}{\beta}}\right)\cdot\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log\frac{nd}{\beta}}}{\sqrt{\rho}n} + \frac{\tau\sqrt{\log\frac{nd}{\beta}}}{n}\right)\right).$$

Proof. We may assume that no sample gets clipped by R', because by Lemma 3.5, with probability $1-\beta/4$, no sample has norm greater than R'. The error consists of two parts, the sampling error $\|f(\mathcal{D})-\mu\|_2$ and the empirical error $\|\hat{\mu}-f(\mathcal{D})\|_2$. The former is bounded by $O\left(\left(\|\Sigma^{1/2}\|_F + \|\Sigma^{1/2}\|_2 \cdot \sqrt{\log\frac{1}{\beta}}\right)\cdot\frac{1}{\sqrt{n}}\right)$ with probability $1-\beta/4$ by Lemma 3.8. To bound the latter using Theorem 3.6, we note that $w(\mathcal{D}) \le O\left(\|\Sigma^{1/2}\|_F + \|\Sigma^{1/2}\|_2 \cdot \sqrt{\log\frac{n}{\beta}}\right)$ with probability $1-\beta/4$ by Lemma 3.8 and a union bound over all $x_i - x_j$. Plugging this into Theorem 3.6 yields the error bound in the corollary. Note that the additive α error due to quantization is dominated by $O(\|\Sigma^{1/2}\|_F/\sqrt{n})$. □

Remark 3.3. Considering $\|\Sigma^{1/2}\|_F \ge \|\Sigma^{1/2}\|_2$ and ignoring the logarithmic factors, the error in Corollary 3.9 simplifies to $\tilde{O}\left(\|\Sigma^{1/2}\|_F\left(\frac{1}{\sqrt{n}} + \frac{1}{n}\sqrt{\frac{d}{\rho}}\right)\right)$. When $\Sigma = \mathbf{I}$, we have $\|\Sigma^{1/2}\|_F = O(\sqrt{d})$, so the bound further simplifies to $\tilde{O}\left(\sqrt{\frac{d}{n}} + \frac{d}{\sqrt{\rho}n}\right)$, matching the known optimal bound for Gaussian mean estimation [9, 29].

4 Lower Bounds

In this section we establish the instance optimality of Theorem 3.6 via three lower bounds: (1) $\mathcal{R}_{\text{in-nbr}}(\mathcal{D}) = \Omega(w(\mathcal{D})/n)$ for all $\mathcal{D}$; (2) an $\tilde{\Omega}(\sqrt{d/\rho})$ lower bound on the optimality ratio, and (3) that the condition $n = \tilde{\Omega}(\sqrt{d/\rho})$ is necessary.

The first lower bound follows from an observation by Vadhan [40]:

Lemma 4.1 ([40]). *For any f, any (ε,δ)-DP mechanism M', and any neighboring datasets $\mathcal{D}_0 \sim \mathcal{D}_1$, there is a $b \in \{0,1\}$ such that*

$$\Pr[\|M'(\mathcal{D}_b) - f(\mathcal{D}_b)\|_2 < \|f(\mathcal{D}_0) - f(\mathcal{D}_1)\|_2/2] < \frac{1+\delta}{1+e^{-\varepsilon}}.$$

Theorem 4.2. *For $\varepsilon < 0.1, \delta < 0.1$, $\mathcal{R}_{\text{in-nbr}}(\mathcal{D}) = \Omega(w(\mathcal{D})/n)$.*

Proof. By the definition of $\mathcal{R}_{\text{in-nbr}}(\mathcal{D})$, it suffices to show that there exists an in-neighbor $\mathcal{D}'$ of $\mathcal{D}$ such that any M' must incur error $\Omega(w(\mathcal{D})/n)$ with probability at least $1/3$ on either $\mathcal{D}$ or $\mathcal{D}'$. Let x_i, x_j be the two vectors in $\mathcal{D}$ that attain the diameter, i.e., $\|x_i - x_j\|_2 = w(\mathcal{D})$. We let $\mathcal{D}'$ be the dataset obtained by changing x_i to x_j in $\mathcal{D}$. It can be verified that $\|f(\mathcal{D}) - f(\mathcal{D}')\|_2 = w(\mathcal{D})/n$ for the mean function f. Then plugging $\mathcal{D}_0 = \mathcal{D}, \mathcal{D}_1 = \mathcal{D}'$ into Lemma 4.1 proves the theorem. □

The lower bound on the optimality ratio is by the reduction from statistical mean estimation, for which there are known lower bounds:

Lemma 4.3 ([29]). *For a Gaussian distribution with unknown mean $\mu \in [-R, R]^d$ and known covariance $\sigma^2 \mathbf{I}$, any (ε, δ)-DP mechanism (for $\delta = \tilde{O}(\sqrt{d}/(nR))$) for estimating μ must incur ℓ_2 error $\tilde{\Omega}(\sigma d/(\varepsilon n))$ with constant probability.*

Theorem 4.4. *Let M be any ρ-zCDP mechanism for mean estimation that has ℓ_2 error $c \cdot w(\mathcal{D})/n$ with constant probability for any $\mathcal{D} = \{x_i\}_{i\in[n]}$ drawn from $[u]^d$. If $\rho < \tilde{O}(\sqrt{d/n})$, then $c = \tilde{\Omega}(\sqrt{d/\rho})$.*

Proof. Lemma 4.3 implies a lower bound of $\tilde{\Omega}(\sigma d/(n\sqrt{\rho}))$ for ρ-zCDP mechanisms, since a ρ-zCDP mechanism is $(\tilde{O}(\sqrt{\rho}), \delta)$-DP. Since $w(\mathcal{D}) = \tilde{O}(\sigma\sqrt{d})$ for Gaussian data $\mathcal{N}(\mu, \sigma^2\mathbf{I})$ w.h.p., the reduction in Corollary 3.9 converts a $c \cdot w(\mathcal{D}/n)$ error for empirical mean to an error of $O(\sigma\sqrt{d/n}) + c \cdot w(\mathcal{D})/n = O(\sigma\sqrt{d/n}(1 + \frac{c}{\sqrt{n}}))$ for Gaussian mean estimation. Comparing with the above lower bound, we obtain

$$1 + \frac{c}{\sqrt{n}} = \tilde{\Omega}\left(\sqrt{\frac{d}{\rho n}}\right).$$

If $\rho = \tilde{O}(\sqrt{d/n})$, the RHS is $\tilde{\Omega}(1)$, then $c = \tilde{\Omega}(\sqrt{d/\rho})$. □

For the lower bound on n, we consider a weaker problem (so the lower bound is stronger), which is the d-dimensional version of the *interior point problem* [11]: Given a dataset $\mathcal{D} = \{x_i\}_{i\in[n]}$ drawn from $[u]^d$, the mechanism is only required to return a $y \in [u]^d$ such that $\min_i x_{ij} \le y_j \le \max_i x_{ij}$ for all j with constant probability.

Theorem 4.5. *If there exists a ρ-zCDP mechanism that solves the interior point problem with success probability $2/3$, then $n = \Omega(\sqrt{d \log u/\rho})$.*

Proof. We need a lemma from [12] to bound the mutual information of a ρ-zCDP mechanism.

Lemma 4.6 ([12]). *Let $M' : \mathcal{X}^n \to \mathcal{Y}$ satisfy ρ-zCDP. Let X be a random variable in $\mathcal{X}^n$. Then, $I(X; M'(X)) \le \rho n^2$, where $I(\cdot;\cdot)$ denotes mutual information.*

Take $\mathcal{X} = [u]^d$ and let M' be a ρ-zCDP mechanism for the interior point problem. Let X be n copies of Z, which is uniformly drawn from $\mathcal{X}$, i.e., $X = \{x_i = Z\}_{i\in[n]}$. Note that X is a random variable drawn from a universe of size $|\mathcal{X}|$, and by the accuracy guarantee of M', Z can be recovered from $M'(X)$ with probability $2/3$. Then, by the Fano's inequality, we have

$$\begin{aligned} I(X; M'(X)) &= H(X) - H(X \mid M'(X)) \\ &= \log|\mathcal{X}| - H(X \mid M'(X)) \\ &\ge \log|\mathcal{X}| - \log 2 - \frac{2}{3}\log|\mathcal{X}| = \Omega(\log|\mathcal{X}|). \end{aligned}$$

Then the theorem follows from Lemma 4.6 and the fact that $\log|\mathcal{X}| = d\log u$. □

Given a quantile selection mechanism with rank error τ, by finding the median, the 1-dimensional interior point problem can be solved when $n = O(\tau)$. The following corollary then follows from Theorem 4.5.

Corollary 4.7. *Any ρ-zCDP mechanism for the quantile selection problem must have rank error $\Omega(\sqrt{\log u/\rho})$.*

5 Extension to the Local Model and Shuffle Model

Our mean estimation framework can be summarized as follows: (1) Given $\mathcal{D} = \{x_1, \dots, x_n\}$, perform a random rotation, obtaining $\hat{\mathcal{D}} = \{\hat{x}_i := HDx_i\}_{i\in[n]}$; (2) For each $j \in [d]$, find an approximate median $\tilde{c}_j$ of $\hat{\mathcal{D}}$ along dimension j. Shift $\hat{\mathcal{D}}$ to be centered around $\tilde{c}$, obtaining $\tilde{\mathcal{D}} = \{\tilde{x}_i := \hat{x}_i - \tilde{c}\}$; (3) Find a clipping threshold C, which is the m-th quantile over the ℓ_2 norms of the vectors in $\tilde{\mathcal{D}}$. In the central model, the optimal choice is $m = n - \sqrt{2d/\rho}$; (4) Perform ℓ_2 clipping over $\tilde{\mathcal{D}}$ using C, and obtain a mean estimator $\tilde{y}$ of the clipped vectors. Finally, return $y = (\frac{1}{\sqrt{d}}HD)^{-1}\frac{1}{\sqrt{d}}(\tilde{y} + \tilde{c})$.

We note that each step has their counterparts in the local and the shuffle model: Step (1) is easy, where the randomized diagonal matrix D can be generated using public randomness, or sent from the aggregator to each user if public randomness is not available. Step (2) and (3) both rely on quantile selection, which have alternatives in the local model and shuffle model. Step (4) is also easy since all vectors have norms bounded by C. Below we elaborate on the details.

5.1 The Local Model

For step (4), the standard LDP mechanism is that each user applies the Gaussian mechanism (Lemma 2.1) with GS $= 2C$ to inject noise to their clipped vector, sends it out, and the aggregator adds them up and divides the sum by n. Thus, the bias of the clipped-mean estimator is the same as that in Lemma 3.1, while the noise increases by a $\sqrt{n}$-factor, to $C\sqrt{\frac{2d}{\rho n}}$. Correspondingly, the optimal choice of C becomes the $m = (n - \sqrt{2dn/\rho})$-th quantile.

For quantile selection in step (2) and (3), we use the LDP range counting protocol in [15]. This one-round protocol returns a data structure from which any range counting query can be answered.

Lemma 5.1 ([15])**.** *There is a one-round ρ-zCDP[2] protocol in the local model that answers all range counting queries within error $O(\sqrt{n/\rho}\log^2 u \log\frac{u}{\beta})$ with probability at least $1 - \beta$.*

Putting things together, we obtain the following result:

Theorem 5.2. *Set $\tau = \sqrt{n/\rho}\log^2(du)\log\frac{du}{\beta}$ and assume $n = \Omega(\tau\sqrt{d})$. There is a 3-round ρ-zCDP mean estimation mechanism in the local model, achieving an ℓ_2-error of*

$$O\left((\sqrt{dn/\rho} + \tau)\sqrt{\log\frac{nd}{\beta}}\right) \cdot w(\mathcal{D})/n$$

with probability $1 - \beta$.

Proof. The proof is the basically the same as that of Theorem 3.4 and 3.6, except that m and τ now take different values. Step (2), (3), and (4) each require a round. □

5.2 The Shuffle Model

We see that the error in the local model is worse than that in the central model by a $\sqrt{n}$-factor. It turns out that in the shuffle model, we can match the result in the central model up to logarithmic factors, albeit with (ε, δ)-DP. This is mostly due to highly accurate summation and range counting protocols discovered recently for the shuffle model. We start with the summation protocol, restated for 1D mean estimation:

[2] The protocol in [15] actually achieves $(\varepsilon, 0)$-DP, but we only need its zCDP version.

Lemma 5.3 ([6, 27])**.** *Given n real values $\mathcal{D} = \{x_i\}_{i\in[n]}$ where $|x_i| \le C$, there is an (ε, δ)-DP mean estimation protocol in the shuffle model that returns a y such that $\mathsf{E}[(y - f(\mathcal{D}))^2] = O\left(\left(\frac{C}{\varepsilon n}\right)^2\right)$.*

Directly applying this protocol in step (4) over C-clipped vectors along each dimension would result in an ℓ_2 error of $\tilde{O}(Cd/(\varepsilon n))$. Below we show how to reduce it to $\tilde{O}(C\sqrt{d}/(\varepsilon n))$, hence matching the error of the clipped-mean estimator in the central model. Given d-dimensional vectors $\mathcal{D} = \{x_i\}_{i\in[n]}$ with ℓ_2 norm bounded by C (in the full algorithm, this would be the dataset after rotation, shifting, and clipping, but we abuse the notation and still use $\mathcal{D}$), we apply another random rotation $\hat{x}_i = HDx_i$. By Lemma 3.7, the coordinates of all $\hat{x}_i$ are bounded (in absolute value) by $C' = O(C\sqrt{\log(nd)})$ w.h.p. Next, we clip each coordinate to C' and invoke Lemma 5.3 with privacy parameter $\varepsilon' = \varepsilon / \left(2\sqrt{d\log\frac{d}{\delta}}\right)$, $\delta' = \delta/d$ along each dimension. This yields a d-dimensional mean estimator $\hat{y}$. Finally, we return $y := (\frac{1}{\sqrt{d}}HD)^{-1}\frac{1}{\sqrt{d}}\hat{y}$.

Lemma 5.4. *Given $\mathcal{D} = \{x_i\}_{i\in[n]} \subset \mathbb{R}^d$ where $\|x_i\|_2 \le C$ for all i, there is a one-round (ε, δ)-DP mean estimation protocol in the shuffle model that returns a y such that*

$$\Pr\left[\|y - f(\mathcal{D})\|_2 \le O\left(\frac{C}{\varepsilon n}\sqrt{d\log(nd)\log\frac{d}{\delta}}\right)\right] \ge 2/3.$$

Proof. Privacy of this protocol follows directly from advanced composition (Lemma 2.2). Below we analyze its accuracy. Conditioned upon the event that all coordinates are bounded by C', which happens with probability 5/6, the C'-clipping on each coordinate has no effects. Then

$$\begin{aligned}
\mathsf{E}[\|y - f(\mathcal{D})\|_2] &= \frac{1}{\sqrt{d}}\mathsf{E}\left[\left\|\hat{y} - \frac{1}{n}\sum_i \hat{x}_i\right\|_2\right] \\
&= \frac{1}{\sqrt{d}}\mathsf{E}\left[\sqrt{\sum_j\left(\hat{y}_j - \frac{1}{n}\sum_i \hat{x}_{i,j}\right)^2}\right] \\
&\le \frac{1}{\sqrt{d}}\sqrt{\sum_j \mathsf{E}\left[\left(\hat{y}_j - \frac{1}{n}\sum_i \hat{x}_{i,j}\right)^2\right]} \\
&\le \frac{1}{\sqrt{d}} \cdot \sqrt{d \cdot O\left(\left(\frac{C'}{\varepsilon' n}\right)^2\right)} = O\left(\frac{C}{\varepsilon n}\sqrt{d\log(nd)\log\frac{d}{\delta}}\right),
\end{aligned}$$

where the first inequality follows from Jensen's inequality and the second inequality is by Lemma 5.3. Then the theorem follows from the Markov inequality. □

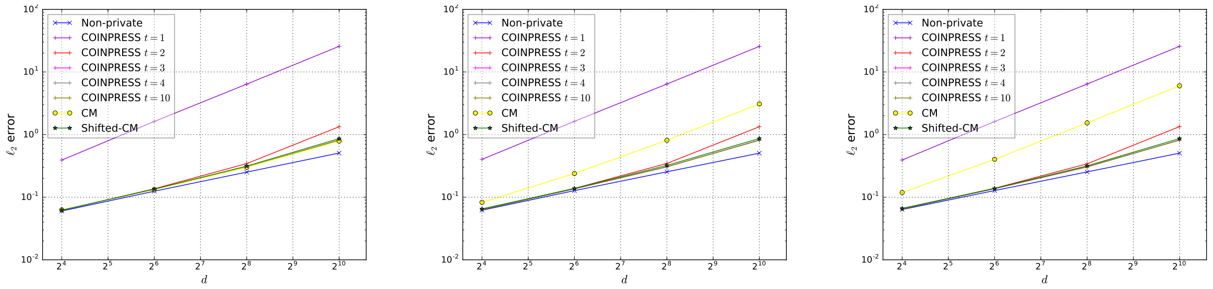
Replacing the clipped-mean estimator with the protocol above, the optimal choice for C becomes the $m = \left(n - \Theta\left(\frac{1}{\varepsilon}\sqrt{d\log(nd)\log\frac{d}{\delta}}\right)\right)$-th quantile. For the `NoisyRC` queries in the algorithm `PrivQuant`, we can use the following range counting mechanism [26] in the shuffle model:

Lemma 5.5 ([26])**.** *There is a one-round (ε, δ)-DP protocol in the shuffle model that answers all range counting queries within error $O\left(\frac{1}{\varepsilon}\log^2 u\sqrt{\log^3\frac{u}{\beta}\log\frac{\log(u/\beta)}{\delta}}\right)$ with probability $1 - \beta$.*

Putting things together, we obtain the following result.

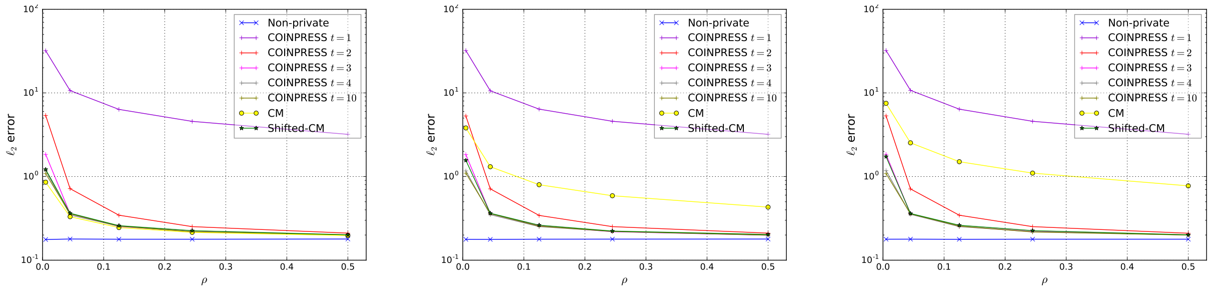
Theorem 5.6. *Set* $\tau = \frac{1}{\varepsilon}\log^{3.5}(du)\sqrt{\log\frac{d\log(du)}{\delta}}$ *and assume* $n = \Omega\left(\tau\sqrt{d\log\frac{d}{\delta}}\right)$. *There is a 3-round* (ε, δ)*-DP mean estimation mechanism in the shuffle model, achieving an* ℓ_2*-error of* $O\left(\left(\frac{1}{\varepsilon}\sqrt{d\log\frac{d}{\delta}} + \tau\right)\sqrt{\log(nd)}\right) \cdot w(\mathcal{D})/n$ *with probability* $2/3$.

6 Experiments



(a) $\mu = 0 \cdot \mathbf{1}_d$. (b) $\mu = 5 \cdot \mathbf{1}_d$. (c) $\mu = 10 \cdot \mathbf{1}_d$.

Figure 1: Comparison with `MVMRec`. ℓ_2 error vs. d for $\mathcal{N}(\mu, I_{d\times d})$, where $n = 4000, \rho = 0.5, R = 50\sqrt{d}$.



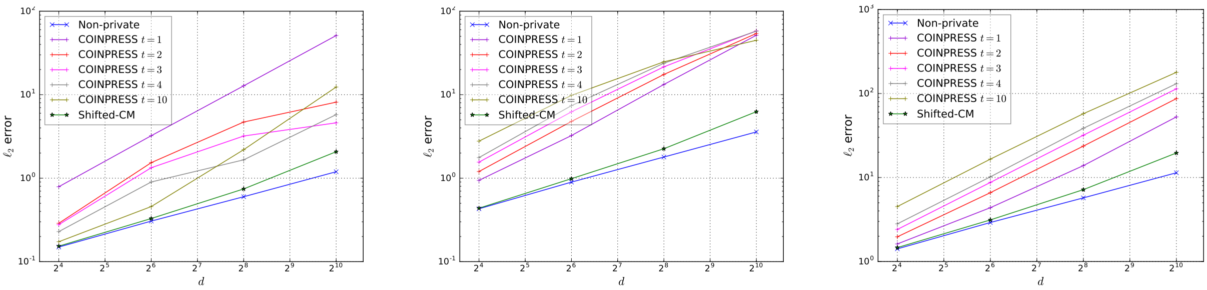
(a) $\mu = 0 \cdot \mathbf{1}_d$. (b) $\mu = 5 \cdot \mathbf{1}_d$. (c) $\mu = 10 \cdot \mathbf{1}_d$.

Figure 2: Comparison with `MVMRec`. ℓ_2 error vs. ρ for $\mathcal{N}(\mu, I_{d\times d})$, where $n = 4000, d = 128, R = 50\sqrt{d}$.

We performed both statistical and empirical mean estimation experiments to evaluate our method. For statistical mean estimation, we used multivariate Gaussian distributions with various μ and Σ. All algorithms are given the same R, $\sigma_{\min}$, $\sigma_{\max}$. We tried various R, while fixing $\sigma_{\min} = 0.1$ and $\sigma_{\max} = R/\sqrt{d}$. For empirical mean estimation, we used a real-world dataset, MNIST, which consists of 70,000 images of handwritten digits, where each image is represented by a vector of dimension $d = 784 = 28 \times 28$. We quantized the values to integers $[u]$ for $u = 2^{10}$. We measured the ℓ_2 error by taking the trimmed mean with trimming parameter 0.1 over 100 trials (as in [9]).

6.1 Results in the Central Model

For the central model, we compared with two versions of `COINPRESS` [9]. `MVMRec`, which is the mean estimation algorithm in [9], and `MVCRec-MVMRec`, where we first scale the data by an estimated Σ (obtained by `MVCRec` using half of the privacy budget ρ) and then apply `MVMRec`. Both `MVMRec` and `MVCRec` start with a given confidence ball of radius R and iteratively refine it. The number of iterations t is an important internal parameter. Following the suggestions in [9], we tried $t = 1, 2, 3, 4, 10$ for MVMRec; for MVCRec-MVMRec, we fixed $t = 10$ for MVMRec and tried $t_{\text{cov}} = 1, 2, 3, 4, 5$. We also note that MVCRec-MVMRec involves complex matrix operations (e.g., matrix inverse) and costs at least $\tilde{\Omega}(nd^2)$ time, in contrast to the $\tilde{O}(nd)$ time of MVMRec and our algorithm.

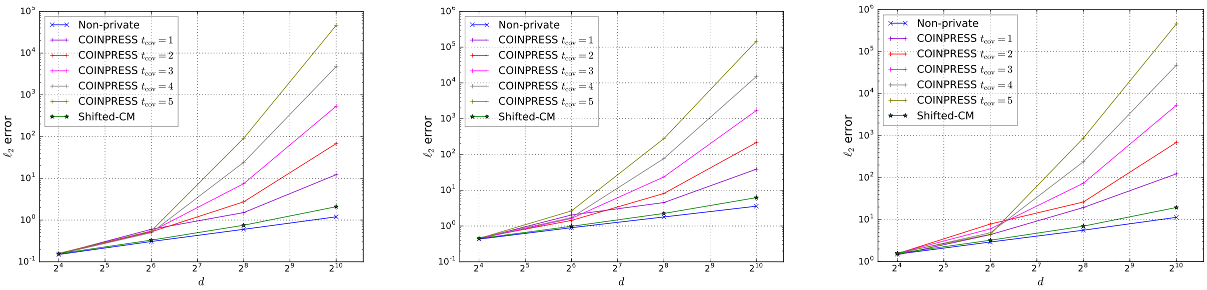


(a) $\kappa = 10$.

(b) $\kappa = 100$.

(c) $\kappa = 1000$.

Figure 3: Comparison with MVMRec. ℓ_2 error vs. d for $\mathcal{N}(0, \Sigma(\kappa))$, where $n = 4000, \rho = 0.5, R = 100\sqrt{d}$.

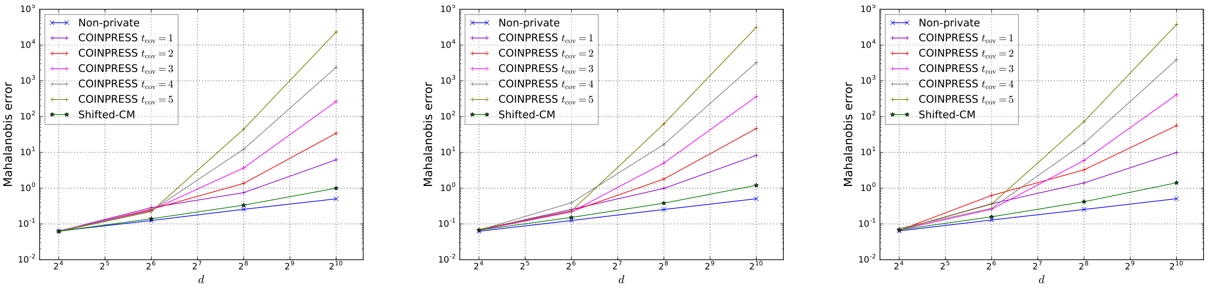


(a) $\kappa = 10$.

(b) $\kappa = 100$.

(c) $\kappa = 1000$.

Figure 4: Comparison with MVCRec-MVMRec. ℓ_2 error vs. d for $\mathcal{N}(0, \Sigma(\kappa))$, where $n = 4000, \rho = 0.5, R = 100\sqrt{d}$.



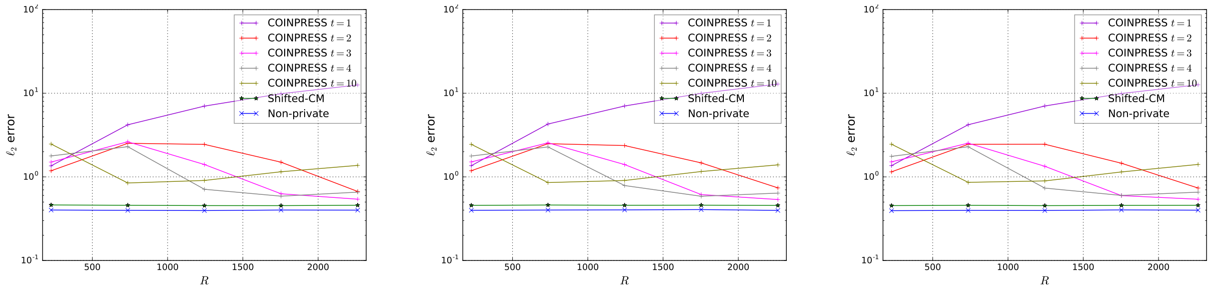
(a) $\kappa = 10$.

(b) $\kappa = 100$.

(c) $\kappa = 1000$.

Figure 5: Comparison with MVCRec-MVMRec. Mahalanobis error vs. d for $\mathcal{N}(0, \Sigma(\kappa))$, where $n = 4000, \rho = 0.5, R = 100\sqrt{d}$.

The results for statistical mean estimation are shown in Fig. 1 to 6, where the detailed parameter settings are given in the captions. The covariance matrix is $\Sigma(\kappa) = A\Lambda(\kappa)A^T$, where A is a random rotation matrix and $\Lambda(\kappa)$ is the diagonal matrix $\text{diag}([\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2])$, $\sigma_j^2 \sim_{\text{u.a.r.}} [1, \kappa]$ for all $j \in [d]$. From these results we can make the following observations. (1) The best choice of t is sensitive to the unknown Σ (see Fig. 3 and 4), making it difficult to tune in practice. (2) The error of our method (Shifted-CM) is always at least as good as `COINPRESS` with the best t across a variety of settings. (3) When Σ is not identity, our method outperforms `MVMRec` significantly (Fig. 3), and yields errors close to the non-private estimator. (4) Estimating Σ first (i.e., `MVCRec-MVMRec`) reduces the error in low dimensions, but the gain diminishes as d gets larger (Fig. 4), which agrees with the analysis in [29] that Σ estimation needs $\tilde{\Omega}(d^{3/2}/\sqrt{\rho})$ samples. (5) `MVCRec-MVMRec` does not really do better in terms of the Mahalanobis error (Fig. 5), again because Σ cannot be estimated well in high dimensions. (6) Both our method and `COINPRESS` are translation-invariant. This can be verified from Fig. 1, 2, and 6, where the results are not effected by μ. However, the approaches taken are different: `COINPRESS` uses an iterative process, while we shift the dataset to be centered around an approximate center point. In Fig. 1 and 2, we also include `CM`, our estimator without this shift operation, which is indeed affected by μ.

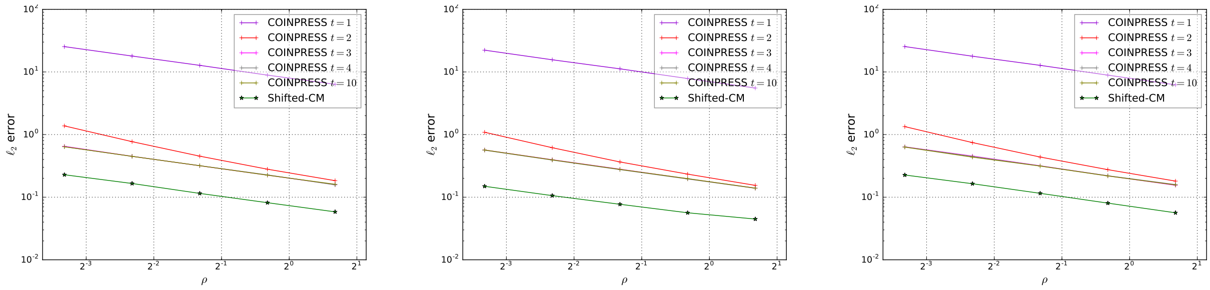


(a) $\mu = 0 \cdot \mathbf{1}_d$.

(b) $\mu = 5 \cdot \mathbf{1}_d$.

(c) $\mu = 10 \cdot \mathbf{1}_d$.

Figure 6: ℓ_2 error vs. R for $\mathcal{N}(\mu, \Sigma(10))$, where $n = 4000, \rho = 0.5, d = 128$.

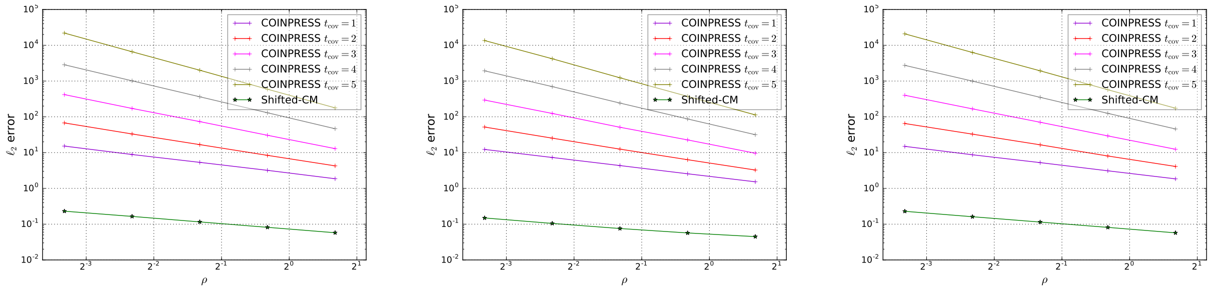


(a) digit 0.

(b) digit 1.

(c) digit 2.

Figure 7: Comparison with `MVMRec`. ℓ_2 error vs. ρ for various digits in MNIST, where $d = 784, R = 50\sqrt{d}$.



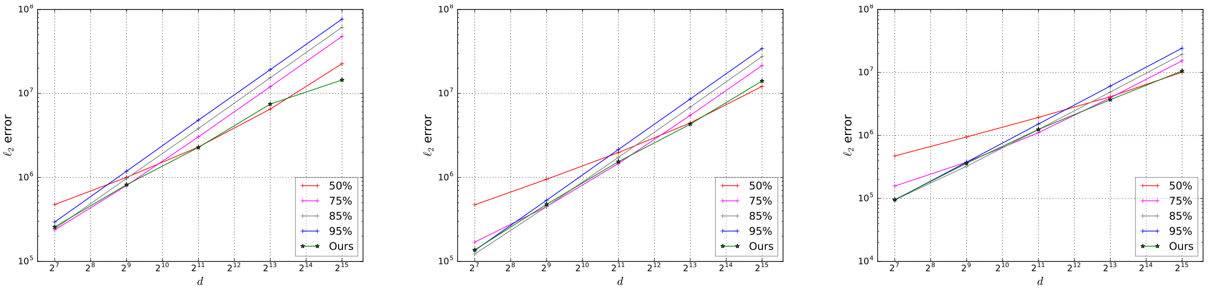
(a) digit 0.

(b) digit 1.

(c) digit 2.

Figure 8: Comparison with `MVCRec-MVMRec`. ℓ_2 error vs. ρ for various digits in MNIST.

For empirical mean estimation on the MNIST dataset, we see in Fig. 7 and 8 that our method outperforms COINPRESS for various privacy levels. This means that this dataset, as with most real-world datasets, does not follow Gaussian distribution with an identity Σ. The instance-optimality of our method is precisely the reason behind its robustness to different distributional assumptions, or the lack of.



(a) $\rho = 0.1$.

(b) $\rho = 0.5$.

(c) $\rho = 1$.

Figure 9: ℓ_2 error vs. d for clipping at various quantiles of the norms on a synthetic dataset.

One important component of our method is the optimal clipping threshold C, which is the $(n - \sqrt{2d/\rho})$-th quantile of the norms. Note that this depends on d and ρ. Prior work [4] used a fixed quantile (e.g., the median). To better see the relationship between the optimal C and d, ρ, we used a synthetic dataset $\mathcal{D} = \{i \cdot \mathbf{1}_d\}_{i \in [n]}$ with $n = 500$ and tried different quantiles as the clipping threshold. The results in Fig. 9 confirm our theoretical analysis: The optimal C indeed depends on d and ρ, while our choice attains the optimum.

6.2 Results in the Local Model

In the local model, there is no prior work on high-dimensional mean estimation. The existing method [25] works only for 1D data. To avoid disadvantaging their method, we used Gaussian distribution with an identity Σ (a Σ with different components would make their method even worse), and apply their method coordinate-wise. They adopt (ε, δ)-DP, so we compose across all dimensions by the advanced composition theorem. We also convert our ρ-zCDP guarantee to (ε, δ)-DP for a fair comparison and set $\delta = 10^{-9}$ in all the experiments.

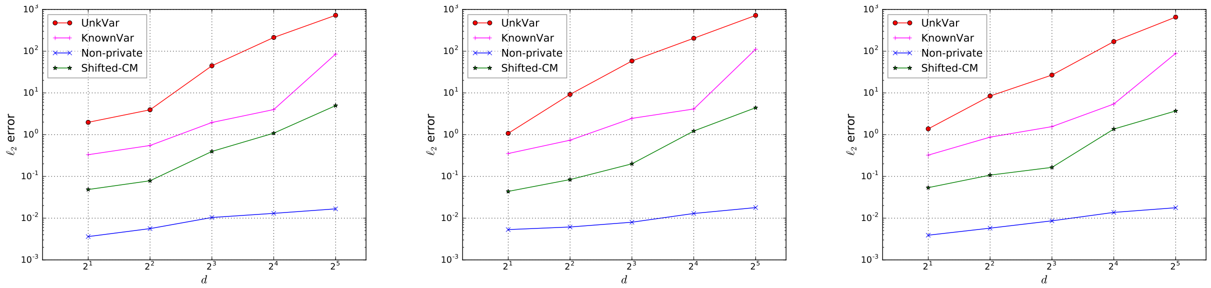
Their method has two versions: KnownVar and UnkVar. The former requires σ to be known, while the latter only requires $\sigma \in [\sigma_{\min}, \sigma_{\max}]$, the same as our method. The results are reported in Fig. 10 to 12, where we vary d, ε, and R, respectively, on a set of $n = 10^5$ samples generated from $\mathcal{N}(\mu, I_{d\times d})$. Our method even outperforms KnownVar, which is actually not a fair comparison.

However, compared with the central model, there is still a large gap from our LDP algorithm to the non-private mean, due to the inherent hardness of the local model. By our theoretical analysis, this gap can be greatly reduced in the shuffle model, once a practical implementation of the range counting protocol from [26] is available.

Acknowledgments

We would like to thank Gautam Kamath and Jonathan Ullman for helpful discussions about CoinPress [9].

This work is supported by HKRGC under grants 16201318, 16201819, and 16205420.

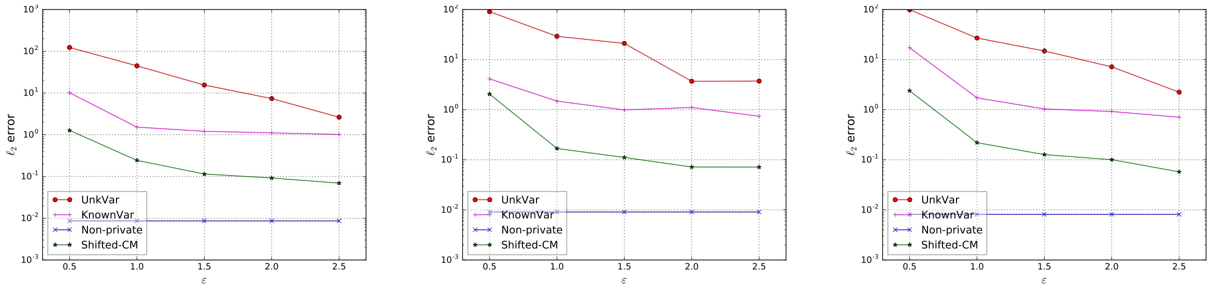


(a) $\mu = 0 \cdot \mathbf{1}_d$.

(b) $\mu = 5 \cdot \mathbf{1}_d$.

(c) $\mu = 10 \cdot \mathbf{1}_d$.

Figure 10: ℓ_2 error vs. d for $\mathcal{N}(\mu, I_{d\times d})$ in the local model, where $n = 10^5, \varepsilon = 1, R = 20\sqrt{d}$.

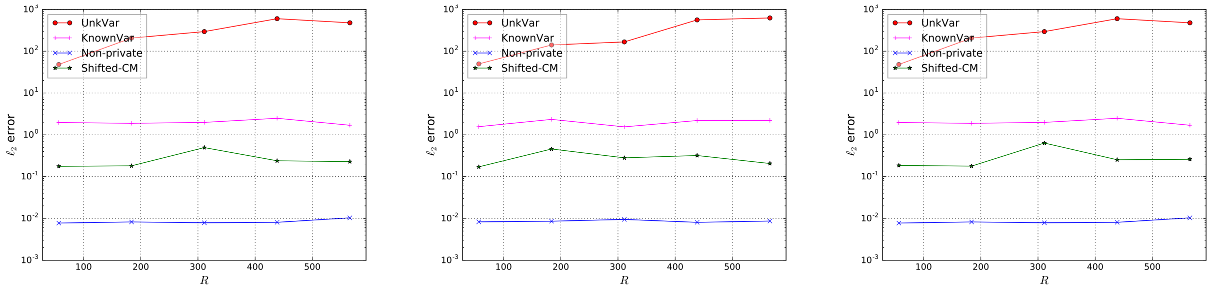


(a) $\mu = 0 \cdot \mathbf{1}_d$.

(b) $\mu = 5 \cdot \mathbf{1}_d$.

(c) $\mu = 10 \cdot \mathbf{1}_d$.

Figure 11: ℓ_2 error vs. ε for $\mathcal{N}(\mu, I_{d\times d})$ in the local model, where $n = 10^5, d = 8, R = 20\sqrt{d}$.



(a) $\mu = 0 \cdot \mathbf{1}_d$.

(b) $\mu = 5 \cdot \mathbf{1}_d$.

(c) $\mu = 10 \cdot \mathbf{1}_d$.

Figure 12: ℓ_2 error vs. R for $\mathcal{N}(\mu, I_{d\times d})$ in the local model, where $n = 10^5, \varepsilon = 1, d = 8$.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

[3] Kareem Amin, Alex Kulesza, Andres Munoz, and Sergei Vassilvtiskii. Bounding user contributions: A bias-variance trade-off in differential privacy. In *International Conference on Machine Learning*, pages 263–271. PMLR, 2019.

[4] Galen Andrew, Om Thakkar, H Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. *arXiv preprint arXiv:1905.03871*, 2019.

[5] Hilal Asi and John C Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in Neural Information Processing Systems*, 33, 2020.

[6] Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. Private summation in the multi-message shuffle model. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 657–676, 2020.

[7] Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 127–135, 2015.

[8] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

[9] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *Advances in Neural and Information Processing Systems*, 2020.

[10] Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. In *Advances in Neural Information Processing Systems*, 2021.

[11] Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 634–649. IEEE, 2015.

[12] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[13] T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Transactions on Information and System Security*, 2011.

[14] TH Hubert Chan, Elaine Shi, and Dawn Song. Optimal lower bound for differentially private multi-party aggregation. In *European Symposium on Algorithms*, pages 277–288. Springer, 2012.

[15] Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. Answering range queries under local differential privacy. *Proceedings of the VLDB Endowment*, 12(10):1126–1138, 2019.

[16] Graham Cormode and Shan Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

[17] Peter Davies, Vijaykrishna Gurunanthan, Niusha Moshrefi, Saleh Ashkboos, and Dan Alistarh. New bounds for distributed mean estimation and variance reduction. In *International Conference on Learning Representations*, 2021.

[18] Apple Differential Privacy Team. Learning with privacy at scale. `https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf`. December 2017.

[19] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *NIPS*, 2017.

[20] John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*, pages 1161–1191. PMLR, 2019.

[21] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy, data processing inequalities, and minimax rates. *arXiv preprint arXiv:1302.3203*, 2013.

[22] Cynthia Dwork, Moni Naor, Omer Reingold, and Guy N Rothblum. Pure differential privacy for rectangle queries via private partitions. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 735–751. Springer, 2015.

[23] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[24] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[25] Marco Gaboardi, Ryan Rogers, and Or Sheffet. Locally private mean estimation: z-test and tight confidence intervals. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2545–2554. PMLR, 2019.

[26] Badih Ghazi, Noah Golowich, Ravi Kumar, Rasmus Pagh, and Ameya Velingker. On the power of multiple anonymous messages. In *EuroCRYPT*, 2021.

[27] Badih Ghazi, Ravi Kumar, Pasin Manurangsi, Rasmus Pagh, and Amer Sinha. Differentially private aggregation in the shuffle model: Almost central accuracy in almost a single message. In *International Conference on Machine Learning*, 2021.

[28] Anna C Gilbert, Yannis Kotidis, S Muthukrishnan, and Martin J Strauss. How to summarize the universe: Dynamic maintenance of quantiles. In *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*, pages 454–465. Elsevier, 2002.

[29] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

[30] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory*, pages 2204–2235. PMLR, 2020.

[31] Gautam Kamath and Jonathan Ullman. A primer on private statistics. *arXiv preprint arXiv:2005.00010*, 2020.

[32] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[33] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[34] Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. *arXiv preprint arXiv:2102.11845*, 2021.

[35] Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 81–90. IEEE, 2010.

[36] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

[37] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

[38] Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

[39] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[40] Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer, 2017.