

Compressive Sensing Approaches for Sparse Distribution Estimation Under Local Privacy

Zhongzheng Xiong
zxxiong21@m.fudan.edu.cn
School of Data Science,
Fudan University
Shanghai, China

Jialin Sun
sunjl20@fudan.edu.cn
School of Data Science,
Fudan University
Shanghai, China

Xiaojun Mao
maoxj@sjtu.edu.cn
School of Mathematical Sciences,
Shanghai Jiao Tong University
Shanghai, China

Jian Wang
jian_wang@fudan.edu.cn
School of Data Science,
Fudan University
Shanghai, China

Ying Shan
yingssh@tencent.com
Tencent
Shenzhen, China

Zengfeng Huang*
huangzf@fudan.edu.cn
School of Data Science,
Fudan University
Shanghai, China

ABSTRACT

Recent years, local differential privacy (LDP) has been adopted by many web service providers like Google [23], Apple [33] and Microsoft [15] to collect and analyse users' data privately. In this paper, we consider the problem of discrete distribution estimation under local differential privacy constraints. Distribution estimation is one of the most fundamental estimation problems, which is widely studied in both non-private and private settings. In the local model, private mechanisms with provably optimal sample complexity are known. However, they are optimal only in the worst-case sense; their sample complexity is proportional to the size of the entire universe, which could be huge in practice. In this paper, we consider sparse or approximately sparse (e.g. highly skewed) distribution, and show that the number of samples needed could be significantly reduced. This problem has been studied recently [1], but they only consider strict sparse distributions and the high privacy regime. We propose new privatization mechanisms based on compressive sensing. Our methods work for approximately sparse distributions and medium privacy, and have optimal sample and communication complexity.

CCS CONCEPTS

• **Security and privacy** → **Privacy-preserving protocols**; • **Mathematics of computing** → **Density estimation**.

KEYWORDS

locally differential privacy, sparse distribution estimation, compressive sensing.

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9096-5/22/04...\$15.00

<https://doi.org/10.1145/3485447.3512220>

ACM Reference Format:

Zhongzheng Xiong, Jialin Sun, Xiaojun Mao, Jian Wang, Ying Shan, and Zengfeng Huang. 2022. Compressive Sensing Approaches for Sparse Distribution Estimation Under Local Privacy. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*, April 25–29, 2022, Virtual Event, Lyon, France. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3485447.3512220>

1 INTRODUCTION

Discrete distribution estimation [25, 27, 29] from samples is a fundamental problem in statistical analysis. In the traditional statistical setting, the primary goal is to achieve best trade-off between sample complexity and estimation accuracy. In many modern data analytical applications, the raw data often contains sensitive information, e.g. medical data of patients, and it is prohibitive to release them without appropriate privatization. Differential privacy is one of the most popular and powerful definitions of privacy [21]. Traditional centralized model assumes there is a trusted data collector. In this paper, we consider *locally differential privacy* (LDP) [8, 28, 40], where users privatize their data before releasing it so as to keep their personal data private even from data collectors. Recently, LDP has been deployed in real world online platforms by several technology organizations including Google [23], Apple [33] and Microsoft [15]. For example, Google deployed their RAPPOR system [23] in Chrome browser for analyzing web browsing behaviors of users in a privacy-preserving manner. LDP has become the standard privacy model for large-scale distributed applications and LDP algorithms are now being used by hundreds of millions of users daily.

We study the discrete distribution estimation problem under LDP constraints. The main theme in private distribution estimation is to optimize statistical and computational efficiency under privacy constraints. Given a privacy parameter, the goal is to achieve best tradeoff between estimation error and sample complexity. In the local model, the communication cost and computation time are also important complexity parameters. This problem has been widely studied in the local model recently [3, 3, 6, 20, 23, 25, 26, 30, 37, 40, 41]. Thus far, the worst-case sample complexity, i.e., the minimum number of samples needed to achieve a desired accuracy for the worst-case distribution, has been well-understood [3, 37, 41]. However, worst-case behaviors are often not indicative of their performance

in practice; real-world inputs often contain special structures that allow one to bypass such worst-case barriers.

In this paper, we consider sparse or approximately sparse distributions $p \in \mathbb{R}^k$, which are perhaps the most natural structured distributions. Let k be the ambient dimensionality of the distribution. The goal in this setting is to achieve *sublinear* (in k) sample complexity. This problem has been studied in [1] very recently. Their method first applies one-bit Hadamard response from [3], and then projects the final estimate to the set of sparse distributions; it is proved that this simple idea leads to sample complexity that only depends on the sparsity s . However, there are still several problems left unresolved. First, the theoretical results in [1] only hold for strictly sparse distributions, which is too restrictive for many applications. Second, they only consider the high privacy regime. A more subtle issue is that, from their algorithm and analyses, the number of samples needed is implicitly assumed to be larger than k . This is because one-bit HR needs to partition the samples into more than k groups of the same size; otherwise the estimation procedure is not well-defined. Therefore, their technique cannot achieve *sublinear* sample complexity even if the distribution is extremely sparse. It is unclear to us whether their projection-based techniques can be modified to resolve all these problems. In this paper, we take a different approach, which resolves the above issues in a unified way. Our contributions are summarized as follows.

- (1) We propose novel privatization schemes based on *compressive sensing* (CS). Our new algorithms have optimal sample and communication complexity simultaneously for sparse distribution estimation. As far as we know, these are the first LDP schemes that achieve this; and this is the first work to apply CS techniques in LDP distribution estimation.
- (2) Applying standard results in CS theory, our method is immediately applicable to estimating approximately sparse distributions.
- (3) We also generalize our techniques to handle medium privacy regimes using ideas from model-based compressive sensing.

Our main idea is to do privatization and dimensionality reduction simultaneously, and then perform distribution estimation in the lower dimensional space. This can reduce the sample complexity because the estimation error depends on the dimensionality of the distribution. The original distribution is then recovered from the low-dimensional one using tools from compressive sensing. We call this technique *compressive privatization* (CP).

1.1 Problem Definition and Results

We consider k -ary discrete distribution estimation. W.l.o.g., we assume the target distribution is defined on the universe $\mathcal{X} = [k] := [1, 2, \dots, k]$, which can be viewed as a k -dimensional vector $p \in \mathbb{R}^k$ with $\|p\|_1 = 1$. Let Δ_k be the set of all k -ary discrete distributions. Given n i.i.d. samples, X_1, \dots, X_n , drawn from the unknown distribution p , the goal is to provide an estimator \hat{p} such that $d(p, \hat{p})$ is minimized, where $d(\cdot)$ is typically the ℓ_1 or ℓ_2 norm.

Local privacy. In the local model, each X_i is held by a different user. Each user will only send a privatized version of their data to the central server, who will then produces the final estimate. A privatization mechanism is a randomized mapping Q that maps

$x \in \mathcal{X}$ to $y \in \mathcal{Y}$ with probability $Q(y|x)$ for some output set \mathcal{Y} . The mapping Q is said to be ϵ -locally differential private (LDP) [20] if for all $x, x' \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have

$$\frac{Q(y|x)}{Q(y|x')} \leq e^\epsilon.$$

LDP distribution estimation. Let $Y = (Y_1, Y_2, \dots, Y_n) \in \mathcal{Y}^n$ be the privatized samples obtained by applying Q on $X = (X_1, \dots, X_n)$. Given privacy parameter ϵ , the goal of LDP distribution estimation is to design an ϵ -LDP mapping Q and a corresponding estimator $\hat{p} : \mathcal{Y}^n \rightarrow \Delta_k$, such that $\mathbb{E}[d(p, \hat{p})]$ is minimized. Given ϵ and α , we are most interested in the number of samples needed (as a function of ϵ and α) to assure ϵ -LDP and $\mathbb{E}[d(p, \hat{p})] \leq \alpha$.

Sparsity. A discrete distribution $p \in \Delta_k$ is called s -sparse if the number of non-zeros in p is at most s . Let $[p]_s$ be the s -sparse vector that contains the top- s entries of p . We say p is approximately (s, λ) -sparse, if $\|p - [p]_s\|_1 \leq \lambda$.

Our results. For the high privacy regime, i.e., $\epsilon = O(1)$, existing studies [3, 23, 25, 37, 40, 41] have achieved optimal sample complexity, which is $\Theta\left(\frac{k^2}{\alpha^2 \epsilon^2}\right)$ for ℓ_1 norm error and $\Theta\left(\frac{k}{\alpha^2 \epsilon^2}\right)$ for ℓ_2 norm error. These worst-case optimal bounds have a dependence on k . Our result (informal) for the high privacy regime is summarized as follows; see Theorem 2.2 for exact bounds and results for approximately sparse distributions.

THEOREM 1.1 (INFORMAL). *For any $0 < \epsilon < 1$ and $\alpha > 0$, there is an ϵ -LDP scheme Q , which produces an estimator \hat{p} with error guarantee $d(p, \hat{p}) \leq \alpha$. If p is s sparse, then the sample complexity of Q is $O\left(\frac{s^2 \log(k/s)}{\epsilon^2 \alpha^2}\right)$ for ℓ_1 error and $O\left(\frac{s \log(k/s)}{\epsilon^2 \alpha^2}\right)$ for ℓ_2 error.*

We provide two different sample optimal privatization methods; the first one has one-bit communication cost and the other one is symmetric (i.e. all users perform the same privatization scheme) but at the cost of using logarithmic communication. Symmetric mechanisms could be beneficial in some distributed settings; it is proved that the communication cost overhead cannot be avoided [2]. Our CS-based technique can be extended to the medium privacy regime with $1 \leq \epsilon \leq \log s$ (see Section 4). The result is summarized in the following theorem. See Theorem 4.3 for exact bounds.

THEOREM 1.2 (INFORMAL). *For any $1 \leq \epsilon^e, 2^b \leq s$ and $\alpha > 0$, there is an ϵ -LDP scheme Q , which produces an estimator \hat{p} with error guarantee $d(p, \hat{p}) \leq \alpha$ with communication no more than b bits. If p is s sparse, then the sample complexity of Q is $O\left(\frac{s^2 \log k/s}{\min\{\epsilon^e, 2^b\} \alpha^2}\right)$ for ℓ_1 error and $O\left(\frac{s \log k/s}{\min\{\epsilon^e, 2^b\} \alpha^2}\right)$ for ℓ_2 error.*

This result provides a characterization on the relationship between accuracy, privacy, and communication, which is nearly tight. A tight and complete characterization for dense distribution estimation was obtained in [14].

1.2 Related work

Differential privacy is the most widely adopted notion of privacy [21]; a large body of literature exists (see e.g. [22] for a comprehensive survey). The local model has become quite popular recently [8, 28, 40]. The distribution estimation problem considered in this

paper has been studied in [3, 6, 14, 20, 23, 25, 26, 30, 37, 38, 40, 41]. Among them, [3, 37, 38, 41] have achieved worst-case optimal sample complexity over all privacy regime. [14] provides a tight characterization on the trade-off between estimation accuracy and sample size under fixed privacy and communication constraints. Their results are tight for all privacy regimes, but have not considered sparsity. Kairouz et al. [25] propose a heuristic called projected decoder, which empirically improves the utility for estimating skewed distributions. They also propose a method to deal with open alphabets, which also reduces the dimensionality of the original distribution first by using hash functions. However, hash functions are not invertible, so they use least squares to recover the original distribution, which has no theoretical guarantee on the estimation error even for sparse distributions. Recently, [1] studied the same problem as in this work. Their method combined one-bit Hadamard response with sparse projection onto the probability simplex. Their methods have provable theoretical guarantees but there are some technical limitations. They also proposed a method, which combined sparse projection with RAPPOR [23]. It achieves optimal sample complexity, but the communication complexity of RAPPOR is $O(k)$ bits for each user, where k is the domain size of the distribution. Recently, [24] lowered the communication complexity of RAPPOR to $O(\log k)$ by employing pseudo random generator. This result is still worse than ours, which only requires 1 bit for each user. The heavy hitter problem, which is closely related to distribution estimation, is also extensively studied in the local privacy model [2, 7, 9, 39]. [36] studies 1-sparse linear regression under LDP constraints. Statistical mean estimation with sparse mean vector is also studied under local privacy, e.g. [5, 19].

1.3 Preliminaries on Compressive Sensing

Let x be an unknown k -dimensional vector. The goal of compressive sensing (CS) is to reconstruct x from only a few linear measurements [11, 13, 16]. To be precise, let $B \in \mathbb{R}^{m \times k}$ be the measurement matrix with $m \ll k$ and $e \in \mathbb{R}^m$ be an unknown noise vector, given $y = Bx + e$, CS aims to recover a sparse approximation \hat{x} of x from y . This problem is ill-defined in general, but when B satisfies some additional properties, it becomes possible [13, 16]. In particular, the *Restricted Isometry Property* (RIP) is widely used.

DEFINITION 1 (RIP). *The matrix B satisfies (s, δ) -RIP property if for every s -sparse vector x ,*

$$(1 - \delta) \|x\|_2 \leq \|Bx\|_2 \leq (1 + \delta) \|x\|_2.$$

We will use the following results from [10]

LEMMA 1.3. *If B satisfies $(2s, 1/\sqrt{2})$ -RIP. Given $y = Bx + e$, there is a polynomial time algorithm, which outputs \hat{x} that satisfies $\|x - \hat{x}\|_2 \leq \frac{C}{\sqrt{s}} \|x - [x]_s\|_1 + D \|e\|_2$ for some constant C, D .*

For the medium privacy regime, we will use the notion of hierarchical sparsity and a model-based RIP condition [31, 32].

DEFINITION 2 (HIERARCHICAL SPARSITY [31]). *Let x be a $k_1 k_2$ -dimensional vector consists of k_1 blocks, each of size k_2 (e.g., $x = [x^{(1)}, \dots, x^{(k_1)}]$). Then x is (s, σ) -hierarchically sparse if at most s blocks have non-zero entries and each of these blocks is σ -sparse.*

DEFINITION 3 (HiRIP [31]). *A matrix $A \in \mathbb{R}^{m \times k}$, where $k = k_1 \times k_2$, is (s, σ) -HiRIP with constant δ if for all (s, σ) -hierarchically sparse vectors $x \in \mathbb{R}^{k_1 k_2}$, we have*

$$(1 - \delta) \|x\|_2 \leq \|Ax\|_2 \leq (1 + \delta) \|x\|_2. \quad (1)$$

THEOREM 1.4 (HiRIP OF KRONECKER PRODUCT [31]). *For any matrix $A \in \mathbb{R}^{M \times K}$ that is (s, δ_1) -RIP and any matrix $B \in \mathbb{R}^{m \times k}$ that is (σ, δ_2) -RIP, $A \otimes B \in \mathbb{R}^{Mm \times Kk}$ satisfies (s, σ) -HiRIP with constant $\delta_{s, \sigma} \leq \delta_1 + \delta_2 + \delta_1 \delta_2$.*

THEOREM 1.5 (RECOVERY GUARANTEE FOR HIERARCHICALLY SPARSE VECTORS [32]). *Suppose matrix $A \in \mathbb{R}^{m \times k}$ ($k = k_1 \times k_2$) is $(3s, 2\sigma)$ -HiRIP with constant $\frac{1}{\sqrt{3}}$. Given $y = Ax + e$ where $x \in \mathbb{R}^k$ is (s, σ) -hierarchically sparse, there is a polynomial time algorithm, which outputs \hat{x} satisfying $\|x - \hat{x}\|_2 \leq C \|e\|_2$ for some constant C .*

2 ONE-BIT COMPRESSIVE PRIVATIZATION

To estimate sparse distributions, Acharya et al. [1] simply add a sparse projection operation at the end of the one-bit HR scheme proposed in [2].

One-bit HR. The users are partitioned into K groups of the same size deterministically, with K being the smallest power of 2 larger than k . Let S_1, \dots, S_K be the groups. Since the partition can be arbitrary, we assume $S_j := \{i \in [n] \mid i \equiv j \pmod{K}\}$. Let H_K be the $K \times K$ Hadamard matrix and $H_{i,j}$ be the (i, j) -entry. In one-bit HR, each user i in group S_j with a sample X_i sends a bit $Y_i \in \{0, 1\}$ distributed as

$$\Pr(Y_i = 1) = \begin{cases} \frac{e^{\epsilon}}{e^{\epsilon} + 1}, & H_{j, X_i} = 1, \\ \frac{1}{e^{\epsilon} + 1}, & H_{j, X_i} = -1. \end{cases} \quad (2)$$

Let $t_j := P(Y_i = 1 \mid i \in S_j)$ for $j \in [K]$ and $\mathbf{t} := (t_1, \dots, t_K)$. The key observation is that

$$\frac{e^{\epsilon} + 1}{e^{\epsilon} - 1} (2\mathbf{t} - \mathbf{1}_K) = H_K \cdot p. \quad (3)$$

Let $\hat{\mathbf{t}} := (\hat{t}_1, \dots, \hat{t}_K)$ where $\hat{t}_j := \frac{1}{|S_j|} \sum_{i \in S_j} Y_i$ is the fraction of messages from S_j that are 1. Then $\hat{\mathbf{t}}$ is an unbiased empirical estimator of \mathbf{t} ; and $\hat{p} = \frac{e^{\epsilon} + 1}{K(e^{\epsilon} - 1)} H_K^T (2\hat{\mathbf{t}} - \mathbf{1}_K)$ is an unbiased estimate of p since $\frac{1}{K} H_K^T H_K = I$.

We note that to make the above estimation process well-defined, the number of samples n must be larger than K , since otherwise some group S_j will be empty and the corresponding \hat{t}_j is undefined. Moreover, the proof of [1] relies on the fact that each \hat{t}_j is the average of $|S_j|$ i.i.d. Bernoulli random variables, which implies $\hat{t}_j - t_j$ is sub-Gaussian with variance $\frac{1}{|S_j|}$. However, if the group S_j is empty, this doesn't hold anymore.

One bit compressive privatization. To resolve the above issue, our scheme doesn't apply one-bit HR but a variant of it. The intuition of our compressive privatization mechanism is that when the distributions are restricted to be sparse, by the theory of compressive sensing, one can use far fewer linear measurements to recover a sparse vector. In our CP method (shown in Algorithm 1), we do not require the response matrix to be invertible as the Hadamard matrix used in one-bit HR. Any matrix $A \in \{-1, +1\}^{m \times k}$ that satisfies the RIP condition will suffice. More specifically, given

Algorithm 1: 1-bit Compressive privatization**Result:** $\hat{p} \in \Delta_k$: an estimate of p **Input:** X_1, \dots, X_n i.i.d from p , privacy parameter ϵ , sparsity s , measurement matrix $A \in \mathbb{R}^{m \times k}$

- 1 For $x \in [m]$, let $B_x := \{y \in [k] : A(x, y) = 1\}$ be the columns where the x th row has 1.
- 2 Divide the n users into m sets S_1, \dots, S_m deterministically by assigning all $i \equiv j \pmod m$ to S_j for $i \in [n]$.
- 3 $\forall j \in [m]$ and $\forall i \in S_j$, the distribution of the one-bit message Y_i is

$$\Pr(Y_i = 1) = \begin{cases} \frac{e^\epsilon}{e^\epsilon + 1}, & X_i \in B_j \\ \frac{1}{e^\epsilon + 1}, & \text{otherwise} \end{cases} \quad (4)$$

Let $\hat{t} := (\hat{t}_1, \dots, \hat{t}_m)$ where $\forall j \in [m]$, $\hat{t}_j := \frac{1}{|S_j|} \sum_{i \in S_j} Y_i$ is the fraction of messages from S_j that are 1.

- 4 Apply Lemma 1.3, with $y = \frac{e^\epsilon + 1}{\sqrt{m(e^\epsilon - 1)}}(2\hat{t} - \mathbf{1}_m)$, $B = \frac{1}{\sqrt{m}}A$ and sparsity s ; let \tilde{p} be the output
- 5 Compute the projection of \tilde{p} onto Δ_k , denoted as \hat{p} .

target sparsity s , we require $\frac{1}{\sqrt{m}}A$ to satisfy $(s, 1/\sqrt{2})$ -RIP. The privatization scheme for each user is almost the same as in one-bit HR, with the Hadamard matrix being replaced by the matrix A above; and clearly this also satisfies ϵ -LDP. In this case, the relation between \mathbf{t} and p in (3) now becomes to

$$\frac{e^\epsilon + 1}{e^\epsilon - 1}(2\mathbf{t} - \mathbf{1}_m) = A \cdot p. \quad (5)$$

On the server side, since A is not necessarily invertible, we need to use sparse recovery algorithms to estimate p . More specifically, we reformulate (5) as

$$\frac{e^\epsilon + 1}{\sqrt{m}(e^\epsilon - 1)}(2\hat{t} - \mathbf{1}_m) = \frac{1}{\sqrt{m}}A \cdot p + \frac{2(e^\epsilon + 1)}{\sqrt{m}(e^\epsilon - 1)}(\hat{t} - \mathbf{t}). \quad (6)$$

Then we can directly apply Lemma 1.3 to compute a sparse vector \hat{p} , with ℓ_2 error, i.e. $\|p - \hat{p}\|_2$, proportional to $\frac{2(e^\epsilon + 1)}{\sqrt{m}(e^\epsilon - 1)}\|\hat{t} - \mathbf{t}\|_2$. Note $\mathbb{E}[\|\hat{t} - \mathbf{t}\|_2]$ is the MSE of the empirical estimator \hat{t} , which only depends on m rather than on k . Moreover, Lemma 1.3 can handle approximately sparse vectors, and thus the above argument is immediately applicable to the setting when p is only approximately sparse. The result is summarized in the following theorem.

THEOREM 2.1 (HIGH PRIVACY REGIME). *Given any matrix $A \in \{\pm 1\}^{m \times k}$ with $\frac{1}{\sqrt{m}}A$ satisfies $(s, 1/\sqrt{2})$ -RIP, assume $\epsilon = O(1)$ and p is s -sparse, then for a target error α , the sample complexity of our method is $O(\frac{m}{\epsilon^2 \alpha^2})$ for ℓ_2 error and $O(\frac{sm}{\epsilon^2 \alpha^2})$ for ℓ_1 error. The ℓ_2 result also holds for $(s, \sqrt{s}\alpha)$ -sparse p and the ℓ_1 result also holds for (s, α) -sparse p . The communication cost for each user is 1 bit.*

PROOF. We first consider the case for ℓ_2 error. Since Δ_k is convex, $\|\hat{p} - p\|_2 \leq \|\tilde{p} - p\|_2$. By Lemma 1.3, we know:

$$\mathbb{E}[\|p - \tilde{p}\|_2] \leq \frac{C}{\sqrt{s}}\|p - [p]_s\|_1 + D \cdot \frac{2(e^\epsilon + 1)}{\sqrt{m}(e^\epsilon - 1)}\mathbb{E}[\|\hat{t} - \mathbf{t}\|_2], \quad (7)$$

where C and D are absolute constants. Since \hat{t} is an empirical estimator of \mathbf{t} , we have

$$\mathbb{E}^2[\|\hat{t} - \mathbf{t}\|_2] \leq \mathbb{E}[\|\hat{t} - \mathbf{t}\|_2^2] = \sum_{y=1}^m \frac{1}{|S_y|^2} \sum_{j \in S_y} \text{Var}(Y_j) \leq \frac{m^2}{4n}, \quad (8)$$

where the first inequality is from Jensen's inequality and the last inequality is from that $Y_j \in \{0, 1\}$. Combining (7) and (8) yields that,

$$\mathbb{E}[\|p - \tilde{p}\|_2] \leq \frac{C}{\sqrt{s}}\|p - [p]_s\|_1 + D \cdot \frac{e^\epsilon + 1}{e^\epsilon - 1} \sqrt{\frac{m}{n}}. \quad (9)$$

When p is s sparse, the first term in (9) is 0. Thus, when $n \geq \frac{bm}{\alpha^2 \epsilon^2}$ for some large enough constant b , the expected ℓ_2 error is at most α . For $(s, \sqrt{s}\alpha)$ -sparse p , the first error term in (9) is bounded by $O(\alpha)$, thus the result still holds for approximately sparse case. For ℓ_1 error, we use L_1 projection in the final step, which means projection by minimizing L_1 distance. In this way, when p is s sparse, we have

$$\|p - \hat{p}\|_1 \leq \|p - \tilde{p}\|_1 + \|\tilde{p} - \hat{p}\|_1 \leq 2\|\tilde{p} - p\|_1 \leq 2\sqrt{2s}\|p - \tilde{p}\|_2 \quad (10)$$

where the first inequality is from triangle inequality, the second inequality is from the L_1 projection and the last inequality is from Cauchy-Schwartz and the fact that $\tilde{p} - p$ is $2s$ -sparse. Thus to achieve an ℓ_1 error of α , it's sufficient to get an estimate with $\alpha' = \alpha/\sqrt{s}$ error for ℓ_2 . The sample complexity is $O(sm/\epsilon^2 \alpha^2)$. For ℓ_1 error with p being (s, α) -sparse, now $p - \tilde{p}$ is $(2s, \alpha)$ -sparse. We have

$$\begin{aligned} \|p - \tilde{p}\|_1 &= \|[p - \tilde{p}]_{2s}\|_1 + \|(p - \tilde{p}) - [p - \tilde{p}]_{2s}\|_1 \\ &\leq \sqrt{2s}\|p - \tilde{p}\|_2 + \alpha. \end{aligned}$$

Then, the ℓ_1 result follows by a similar argument as for the exact sparse case. \square

2.1 Guarantees on Random Matrices

The measurement matrix we use is $B = \frac{1}{\sqrt{m}}A$, where the entries of $A \in \{-1, +1\}^{m \times k}$ are i.i.d. Rademacher random variables, i.e., takes $+1$ or -1 with equal probability. It is known that for $m \geq O(s \log \frac{k}{s})$, B satisfies $(s, 1/\sqrt{2})$ -RIP with probability $1 - e^{-m}$ [4]. By Theorem 2.1, we have the following theorem.

THEOREM 2.2. *For $m = O\left(s \log \frac{k}{s}\right)$ and $\epsilon = O(1)$, if the entries of A are i.i.d. Rademacher random variables, then with probability at least $1 - e^{-m}$, our method has sample complexity $O\left(\frac{s \log(k/s)}{\epsilon^2 \alpha^2}\right)$ for ℓ_2 error, and $O\left(\frac{s^2 \log(k/s)}{\epsilon^2 \alpha^2}\right)$ for ℓ_1 error. The ℓ_2 result holds for $(s, \sqrt{s}\alpha)$ -sparse p and the ℓ_1 result holds for (s, α) -sparse p . The communication cost for each user is 1 bit.*

Lower bound. [1] proves a lower bound of $n = \Omega\left(\frac{s^2 \log(k/s)}{\epsilon^2 \alpha^2}\right)$ on the sample complexity for ℓ_1 error with $\epsilon = O(1)$. This matches our bound up to a constant.

3 SYMMETRIC COMPRESSIVE PRIVATIZATION

The one-bit compressive privatization scheme is asymmetric, where users in different groups apply different privatization schemes. In

this section, we introduce a symmetric version of compressive privatization, which could be easier to implement in real applications.

To estimate an unknown distribution $p \in \mathbb{R}^k$, all previous symmetric LDP mechanisms essentially apply a probability transition matrix Q mapping p to q , where q is the distribution of the privatized samples. The central server get n independent samples from q , from which it computes an empirical estimate of q , denoted as \hat{q} , and then computes an estimator of p from \hat{q} by solving $Qp = \hat{q}$. The key is to design an appropriate Q such that it satisfies privacy guarantees and achieves low recovery error. The error of $\hat{p} = Q^{-1}p$ is dictated by the estimation error of \hat{q} and the spectral norm of Q^{-1} . In our symmetric scheme, we map p to a much lower dimensional q , and then \hat{q} with similar estimation error can be obtained with *much less number of samples*. However, now Q is not invertible; to reconstruct \hat{p} from \hat{q} , we use sparse recovery [12, 13].

Privatization. Our mechanism Q is a mapping from $[k]$ to $[m]$. For each $x \in [k]$, we pick a set $C_x \subseteq [m]$, which will be specified later, and let $n_x = |C_x|$. Our privatization scheme Q is given by the conditional probability of y given x :

$$Q(y|x) := \begin{cases} \frac{e^\epsilon}{n_x e^\epsilon + m - n_x} & \text{if } y \in C_x, \\ \frac{1}{n_x e^\epsilon + m - n_x} & \text{if } y \in [m] \setminus C_x. \end{cases} \quad (11)$$

Note $Q(y|x)$ is an m -ary distribution for any $x \in [k]$. For each $x \in [m]$, we define its incidence vector as $I_x \in \{-1, +1\}^m$ such that $I_x(j) = +1$ iff $j \in C_x$. Let $A \in \mathbb{R}^{m \times k}$ be the matrix whose x -th column is I_x . Each user i with a sample X_i generates Y_i according to (11) and then send Y_i to the server with communication cost $O(\log m)$ bits.

Sufficient conditions for Q . The difference between our mechanism, RR [40] and HR [3] is the choice of each C_x , or equivalently the matrix A . In RR, $C_x = \{x\}$ for all $x \in [k]$, while in HR, A is the Hadamard matrix. In our privatization method, any matrix A whose column sums are close to 0 and that satisfies RIP will suffice. More formally, given target error α , privacy parameter ϵ and target sparsity s , we require A to have the following 2 properties:

- **P1:** $(1 - \beta) \frac{m}{2} \leq n_i \leq (1 + \beta) \frac{m}{2}$ for all $i \in [k]$, with $\beta \leq \frac{\epsilon}{2}$ and $\beta \leq c\alpha$ for some c depending on the error norm.
- **P2:** $\frac{1}{\sqrt{m}}A$ satisfies (s, δ) -RIP, where s is the target sparsity and $\delta \leq 1/\sqrt{2}$.

In [3], Hadamard matrix is used to specify each set C_x . The proportion of +1 entries is exactly half in each column of H (except for the first column), and thus P1 is automatically satisfied with $\beta = 0$. Since Hadamard matrix is orthonormal, it is $(k, 0)$ -RIP.

3.1 Estimation Algorithm

We first show how to model the estimation of p as a standard compressive sensing problem. Recall n_i is the number of +1's in the i th column of A . Let $q \in \Delta_m$ be the distribution of a privatized sample, given that the input sample is distributed according to $p \in \Delta_k$. Then, for each $j \in [m]$, we have

$$\begin{aligned} q_j &= \sum_{i=1}^k p_i \cdot Q(Y = j|X = i) \\ &= \sum_{i: j \in C_i} \frac{e^\epsilon \cdot p_i}{n_i e^\epsilon + m - n_i} + \sum_{i: j \in [m] \setminus C_i} \frac{p_i}{n_i e^\epsilon + m - n_i}. \end{aligned}$$

By writing the above formula in the matrix form, we get

$$q = \left(\frac{e^\epsilon - 1}{2} A + \frac{e^\epsilon + 1}{2} J \right) Dp, \quad (12)$$

where $J \in \mathbb{R}^{m \times k}$ is the all-one matrix and D is the diagonal matrix with $d_i = \frac{1}{n_i e^\epsilon + m - n_i}$ in the i th diagonal entry. As mentioned above, the matrix $\frac{1}{\sqrt{m}}A$ to be used will satisfy RIP. We then rewrite (12) to the form of a standard noisy compressive sensing problem

$$\underbrace{\frac{(e^\epsilon + 1)}{(e^\epsilon - 1)} \left(\sqrt{m}q - \frac{1}{\sqrt{m}} \right)}_y = \underbrace{\frac{1}{\sqrt{m}}A}_{B} \underbrace{(D'p)}_x + \underbrace{\frac{e^\epsilon + 1}{\sqrt{m}(e^\epsilon - 1)} J(D' - I)p}_e.$$

where $\mathbf{1} \in \mathbb{R}^m$ is the all-one vector, $I \in \mathbb{R}^{k \times k}$ is the identity matrix and $D' = \frac{m(e^\epsilon + 1)}{2}D$. The exact q is also unknown, and we can only get an empirical estimate \hat{q} from the privatized samples. So, we need to add a new noise term that corresponds to the estimation error of \hat{q} , and the actual under-determined linear system is

$$\begin{aligned} \frac{(e^\epsilon + 1)}{(e^\epsilon - 1)} \left(\sqrt{m}\hat{q} - \frac{1}{\sqrt{m}} \right) &= \frac{1}{\sqrt{m}}AD'p + \underbrace{\frac{e^\epsilon + 1}{\sqrt{m}(e^\epsilon - 1)} J(D' - I)p}_{e_1} \\ &+ \underbrace{\frac{\sqrt{m}(e^\epsilon + 1)}{e^\epsilon - 1} (\hat{q} - q)}_{e_2}. \end{aligned} \quad (13)$$

Given the LHS of (13), $\frac{1}{\sqrt{m}}A$ and a target sparsity s , we reconstruct $\widehat{D'p}$ by applying Lemma 1.3. Then compute $\hat{p}' = D'^{-1}\widehat{D'p}$ (D' is known) and project it to the probability simplex. The pseudo code of the algorithm is presented in Algorithm 2.

3.2 Privacy Guarantee and Sample Complexity

In this section, we will provide the privacy guarantee and sample complexity of our symmetric privatization scheme. All the proofs are in the supplementary material.

LEMMA 3.1 (PRIVACY GUARANTEE). *If $(1 - \beta) \frac{m}{2} \leq n_i \leq (1 + \beta) \frac{m}{2}$ for all $i \in [k]$ and $0 \leq \beta < 1$, then the privacy mechanism Q from (11) satisfies $(\epsilon + 2\beta)$ -LDP.*

When the matrix A used in (11) satisfies $(1 - \epsilon) \frac{m}{2} \leq n_i \leq (1 + \epsilon) \frac{m}{2}$ for all $i \in [k]$, then Q is 3ϵ -LDP. We can rescale ϵ in the beginning by a constant to ensure ϵ -LDP, which will only affect the sample complexity by a constant factor. Next we consider the estimation error. By Lemma 1.3, the reconstruction error depends on the ℓ_2 norm of e_1 and e_2 in (13).

LEMMA 3.2. *For a fixed $\beta \in [0, 0.5)$, if $n_i \in (1 \pm \beta) \frac{m}{2}$ for all $i \in [k]$, then $\|e_1\|_2 \leq \frac{\beta}{1 - \beta} \leq 2\beta$.*

Algorithm 2: Symmetric compressive privatization**Result:** $\hat{p} \in \Delta_k$: an estimate of p **Input:** X_1, \dots, X_n i.i.d from p , privacy parameter ϵ , sparsity s , $A \in \mathbb{R}^{m \times k}$

- 1 $\forall x \in [k]$, Let $C_x := \{y : A(y, x) = 1\}$. Let n_x be the number of +1 in the x -th column of A . Then for each X_i , $i \in [n]$, the privatized sample Y_i is generated according to the following distribution

$$\Pr[Y_i = y | X_i = x] = \begin{cases} \frac{e^\epsilon}{n_x e^\epsilon + m - n_x} & y \in C_x, \\ \frac{1}{n_x e^\epsilon + m - n_x} & \text{otherwise} \end{cases}$$

- 2 $\hat{q} = (0, 0, \dots, 0) \in \mathbb{R}^m$, $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{R}^m$
- 3 **for** $i \leftarrow 1$ **to** m **do**
- 4 $\hat{q}[i] = \frac{\sum_{j=1}^n \mathbb{1}(Y_j = i)}{n}$
- 5 **end**
- 6 Apply Lemma 1.3, with $y = \frac{(e^\epsilon + 1)}{(e^\epsilon - 1)} (\sqrt{m}\hat{q} - \mathbf{1})$, $B = \frac{1}{\sqrt{m}}A$ and sparsity s ; let f be the output
- 7 Compute $\hat{p}' = D^{-1}f$, and then compute the projection of \hat{p}' onto the set Δ_k , denoted as \hat{p}
- 8 **return** \hat{p}

$$\text{LEMMA 3.3. } \mathbb{E}[\|e_2\|_2] \leq \frac{e^\epsilon + 1}{e^\epsilon - 1} \sqrt{\frac{m}{n}}.$$

Combining Lemma 3.2, 3.3 and Lemma 1.3, we can bound the estimation error.

THEOREM 3.4 (ESTIMATION ERROR). *If A satisfies two properties in Section 3, for some constant C ,*

$$\mathbb{E}[\|p - \hat{p}'\|_2] \leq \left(1 + \frac{\beta}{2}\right) \left(2D\beta + \frac{D(e^\epsilon + 1)}{e^\epsilon - 1} \sqrt{\frac{m}{n}}\right) + \left(1 + \frac{\beta}{2}\right)^2 \left(\frac{C}{\sqrt{s}} \|p - [p]_s\|_1\right).$$

The sample complexity to achieve an error α and ϵ -LDP for $0 \leq \epsilon \leq 1$ is summarized as follows.

COROLLARY 3.4.1. *If A satisfies two properties in section 3 with $\beta \leq c\alpha/4D$, with $c = 1$ for ℓ_2 error and $c = \frac{1}{\sqrt{s}}$ for ℓ_1 error, and p is s sparse, the sample complexity of our symmetric compressive privatization scheme is $O(\frac{m}{\epsilon^2 \alpha^2})$ for ℓ_2 error and $O(\frac{sm}{\epsilon^2 \alpha^2})$ for ℓ_1 error. The ℓ_2 result also holds for $(s, \sqrt{s}\alpha)$ -sparse p and the ℓ_1 result also holds for (s, α) -sparse p . The communication cost for each user is $\log m$ bits.*

This corollary can be directly derived by applying a similar proof as that of Theorem 2.1, so we omit the proof here.

3.3 Guarantees on Random Matrices

The response matrix we used here is $B = \frac{1}{\sqrt{m}} [A \frac{m}{2}; -A \frac{m}{2}]$, where $A \frac{m}{2}$ is a $\frac{m}{2} \times k$ Rademacher matrix. It can be easily shown that, if $m = \Omega(s \log \frac{k}{s})$, B is $(s, \frac{1}{\sqrt{2}})$ -RIP with probability $1 - e^{-m}$. At the same time, B satisfies P1 with $\beta = 0$. Thus by Corollary 3.4.1, we get the following result.

THEOREM 3.5. *For $m = \Omega(s \log \frac{k}{s})$, if A is defined as above, then with probability at least $1 - e^{-m}$, our symmetric compressive privatization scheme has sample complexity $O(\frac{m}{\epsilon^2 \alpha^2})$ for ℓ_2 error, and $O(\frac{sm}{\epsilon^2 \alpha^2})$ for ℓ_1 error. The ℓ_2 result holds for $(s, \sqrt{s}\alpha)$ -sparse p and the ℓ_1 result holds for (s, α) -sparse p . The communication cost for each user is $\log m = O(\log s + \log \log k)$ bits.*

Communication lower bound . Note that the sample complexity is the same as that of one-bit compressive privatization. But the communication complexity is $\log m$ bits. [2] proves that, for symmetric schemes, the communication cost is at least $\log k - 2$ for general distribution estimation. Thus, even if the sparse support is known, the communication cost is at least $\log s - 2$ bits. Our symmetric scheme requires $\log s + \log \log k$ bits, which is optimal up to a $\log \log k$ additive term.

4 RECURSIVE COMPRESSIVE PRIVATIZATION FOR MEDIUM PRIVACY REGIMES

For high privacy regime $\epsilon = O(1)$, we have provided symmetric and asymmetric privatization schemes that both achieve optimal sample and communication complexity. For medium privacy regime $1 < e^\epsilon < k$, the relationship between sample complexity, privacy, and communication cost become more complicated. Recently, [14] provides a clean characterization for any ϵ and communication budget b for dense distributions. Interestingly, they show that the complexity is determined by the more stringent constraint, and the less stringent constraint can be satisfied for free. In this section, we provide an analogous result for sparse distribution estimation for privacy regime where $1 \leq e^\epsilon \leq s$.

Our RCP scheme (shown in Algorithm 3) consists of three steps including random permutation, privatization and estimation. Let X be a element sampled from p , which is viewed as a one-hot vector. Let A be a measurement matrix and $Y = AX$. Since $\mathbb{E}[Y] = A\mathbb{E}[X] = Ap$, we want to get an estimator of p by recovering from the empirical mean of Y . In one-bit CP, A is a Rademacher matrix, while in RCP, we use the kronecker product of two RIP matrices. In other words, $A = A_1 \otimes A_2$, where A_1 and A_2 both satisfy RIP condition. It's known from kronecker compressive sensing [17] that A also satisfies s -RIP if both A_1 and A_2 satisfy s -RIP. However, s -RIP condition is too stringent for A_2 , since A_2 measures each block of p and the sparsity of each block could be much less than s on average.

We use hierarchical compressive sensing. To make the distribution p hierarchically sparse (see definition 2), we first randomly permute p in the beginning and let p' be the resulting distribution. Thus the sparsity of each block in p' is roughly s/L , where L is the number of blocks, and hence p' is nearly $(L, s/L)$ -hierarchically sparse.

Hierarchical sparsity after random permutation. Let P be a random permutation matrix, which is public information. Each user i with sample X_i first compute $X'_i = PX_i$. So $p' = Pp$. We divide p' into L consecutive blocks, then each of the L blocks of p' has sparsity around $\frac{s}{L}$ with high probability. Let \mathcal{E} be the event that

$$s_i \leq (1 + \beta) \frac{s}{L}, \text{ for all } i \in [L], \text{ for some } \beta > 0, \quad (16)$$

Algorithm 3: Recursive Compressive Privatization**Result:** $\hat{p} \in \Delta_k$: an estimate of p **Input:** X_1, \dots, X_n i.i.d. samples from p , privacy parameter ϵ , sparsity s , matrix $A_1 \in \mathbb{R}^{L \times L}$, $A_2 \in \mathbb{R}^{m \times \frac{k}{L}}$ where $K = 2^{\lceil \log_2 k \rceil}$ and $L = \min\{2^b, 2^{\lceil \log_2 e^\epsilon \rceil}\}$, public random permutation matrix $P \in \mathbb{R}^{K \times K}$. (We represent X_1, \dots, X_n as one-hot vectors.)

- 1 Divide the n users into m groups S_1, \dots, S_m :
 $S_j := \{i \in [n] \mid i \equiv j \pmod{m}\}$.
- 2 $\forall j \in [m]$ and $\forall i \in S_j$, pad $(K - k)$ zeroes to the end of X_i and get permuted sample $X'_i = PX_i$.

3 Define

$$Q_j(X'_i) = [(A_2)_j \cdot X'_i^{(1)}, \dots, (A_2)_j \cdot X'_i^{(L)}] \in \{-1, 0, 1\}^L.$$

The privatized output $\hat{Q}_j(X'_i)$ is defined as follows

$$\hat{Q}_j(X'_i) = \begin{cases} Q_j(X'_i), & \text{w.p. } \frac{e^\epsilon}{e^\epsilon + 2L - 1} \\ Q' \in \mathcal{Q} \setminus \{Q_j(X'_i)\}, & \text{w.p. } \frac{e^\epsilon}{e^\epsilon + 2L - 1} \end{cases} \quad (14)$$

where $\mathcal{Q} = \{\pm e_1, \pm e_2, \dots, \pm e_L\}$ is the collection of $2L$ standard basis vectors.

- 4 For each $j' \in [mL]$ such that $j' \equiv j \pmod{m}$ and $j' = j + (t - 1) \cdot m$, the server computes

$$\hat{q}_{j'} = \frac{m}{n} \left(\frac{e^\epsilon + 2L - 1}{e^\epsilon - 1} \right) \sum_{i \in S_j} (A_1)_t \cdot \hat{Q}_j(X'_i). \quad (15)$$

- 5 Let $\hat{q} := (\hat{q}_1, \dots, \hat{q}_{mL})$. Apply Theorem 1.5 with $y = \frac{1}{\sqrt{mL}} \hat{q}$, measurement matrix $\frac{1}{\sqrt{mL}} (A_1 \otimes A_2)$ and hierarchical sparsity $(L, (1 + \beta) \frac{s}{L})$; let \hat{p}' be the output. Let \tilde{p} be the first k elements of $P^{-1} \hat{p}'$.
- 6 Compute the projection of \tilde{p} onto Δ_k , denoted as \hat{p} .

where s_t is the sparsity of the t th block in p' . For notation convenience, we simply use X_i to denote the permuted one-hot sample vector of user i . By concentration inequalities, \mathcal{E} happens with high probability. One twist here is that we cannot apply standard Chernoff-Hoeffding bound, since the sparsity in each block is not a sum of i.i.d. random variables. But random permutation random variables are known to be negatively associated (see e.g. [35]), so the concentration bounds still holds. We have the following lemma, the proof of which is provided in the supplementary material.

LEMMA 4.1. *Event \mathcal{E} holds with probability at least $1 - Le^{-\frac{\beta^2 s}{2L}}$, and when \mathcal{E} happens, p' is $(L, (1 + \beta) \frac{s}{L})$ -hierarchically sparse.*

Privatization. In our privatization step, the response matrix is of the form $A = A_1 \otimes A_2$, where $A_2 \in \mathbb{R}^{m \times \frac{k}{L}}$ and $A_1 \in \mathbb{R}^{L \times L}$ are two ± 1 matrices. User i will send a privatized version of $Y_i = AX_i$. Let a_{ij} be the (i, j) -entry of A_1 , then

$$Y_i = AX_i = \begin{bmatrix} a_{11}A_2 & \cdots & a_{1L}A_2 \\ \vdots & \ddots & \vdots \\ a_{L1}A_2 & \cdots & a_{LL}A_2 \end{bmatrix} \begin{bmatrix} X_i^{(1)} \\ \vdots \\ X_i^{(L)} \end{bmatrix},$$

where the one-hot sample vector X_i is divided into L blocks. Since X_i has only one non-zero, the vector Y_i can be encoded with $m + \log L$

bits, where $\log L$ bits is to specified the block index ℓ that contains the non-zero of X_i and m bits is for $A_2 X_i^{(\ell)}$. However, this is still too large, so user will only pick one bit from $A_2 X_i^{(\ell)}$, which is the j th bit if $i \in S_j$. Then the user privatize the $1 + \log L$ bits using RR mechanism [40] with alphabet size 2^{2L} . More formally, let $Q_j(X_i) = [(A_2)_j \cdot X_i^{(1)}, \dots, (A_2)_j \cdot X_i^{(L)}] \in \{-1, 0, 1\}^L$, which is also a one-hot vector. Let ℓ be the block index such that $X_i^{(\ell)} \neq 0$, then the ℓ th bit in $Q_j(X_i)$ is the only non-zero entry, which has value $(A_2)_j \cdot X_i^{(\ell)}$. Then the user computes a privatization of $Q_j(X_i)$ (see (14) in Algorithm 3) with 2^{2L} -RR. Clearly the communication cost is $l = \log_2 L + 1$ bits.

Estimation via hierarchical sparse recovery. By Definition 2, p' is $(L, (1 + \beta) \frac{s}{L})$ -hierarchically sparse. To recover p' , by Theorem 1.4 and 1.5, $\frac{1}{\sqrt{L}} A_1$ and $\frac{1}{\sqrt{m}} A_2$ are required to satisfy $3L$ -RIP and $\frac{2(1+\beta)s}{L}$ -RIP condition respectively. Since A_1 is square, we can use the Hadamard matrix H_L . For A_2 , we use a Rademacher matrix with number of rows $m = \Theta\left((1 + \beta) \frac{s \log(k/(1+\beta)s)}{L}\right)$, so that $\frac{1}{\sqrt{m}} A_2$ satisfies $\frac{2(1+\beta)s}{L}$ -RIP. Let $q = Ap'$, which is equivalent to

$$\frac{1}{\sqrt{mL}} \hat{q} = \frac{1}{\sqrt{mL}} Ap' + \frac{1}{\sqrt{mL}} (\hat{q} - q). \quad (17)$$

In the estimation algorithm (step 5 in Algorithm 3), we use \hat{q} to recover p' . By Theorem 1.5, we have

$$\mathbb{E}[\|p - \hat{p}\|_2] \leq \mathbb{E}[\|p' - \hat{p}'\|_2] \leq \frac{1}{\sqrt{mL}} \mathbb{E}[\|\hat{q} - q\|_2] \quad (18)$$

Thus, the estimation error of \hat{p} is bounded by the error of \hat{q} . The following lemma gives an upper bound on the error of \hat{q} .

$$\text{LEMMA 4.2. } \forall j' \in [mL], \mathbb{E} \left[(\hat{q}_{j'} - q_{j'})^2 \right] \leq \frac{m}{n} \left(\frac{e^\epsilon + 2L - 1}{e^\epsilon - 1} \right)^2.$$

Note that we require $m = C' \cdot \frac{(1+\beta)s \log(k/(1+\beta)s)}{L}$ for some absolute constant C' . It can be seen that the error in Lemma 4.2 is minimized when $L = \Theta(e^\epsilon)$ and decreasing as L increases from 1 to $\Theta(e^\epsilon)$. Note that the communication cost is $\log L + 1$. Thus, if we are further given a communication budget b , we need to set $L \leq 2^{b-1}$. When $e^\epsilon < 2^b$, the best L is e^ϵ , which leads to optimal error. If $e^\epsilon \geq 2^b$, i.e., communication becomes the more stringent constraint, then set $L = 2^b$. In other words, $L = \min\{2^b, e^\epsilon\}$. Combining (18) and Lemma 4.2, we have the following results on the sample complexity for medium privacy $1 \leq e^\epsilon \leq s$.

THEOREM 4.3. *Given ϵ and a communication budget b , with $1 \leq e^\epsilon, 2^b \leq s$ and let $L = \min\{2^b, e^\epsilon\}$. For $m = \Theta\left(\frac{(1+\beta)s \log(k/(1+\beta)s)}{L}\right)$, with probability at least $1 - Le^{-\frac{\beta^2 s}{2L}} - e^{-m}$, our scheme is ϵ -LDP and has communication cost $\log L + 1$, and the sample complexity for ℓ_2 error is $O\left((1 + \beta) \frac{s \log(k/(1+\beta)s)}{L \alpha^2}\right)$; for ℓ_1 error, the sample complexity is $O\left((1 + \beta) \frac{s^2 \log(k/(1+\beta)s)}{L \alpha^2}\right)$.*

The parameter β is from Lemma 4.1. Please refer to the appendix for more discussion.

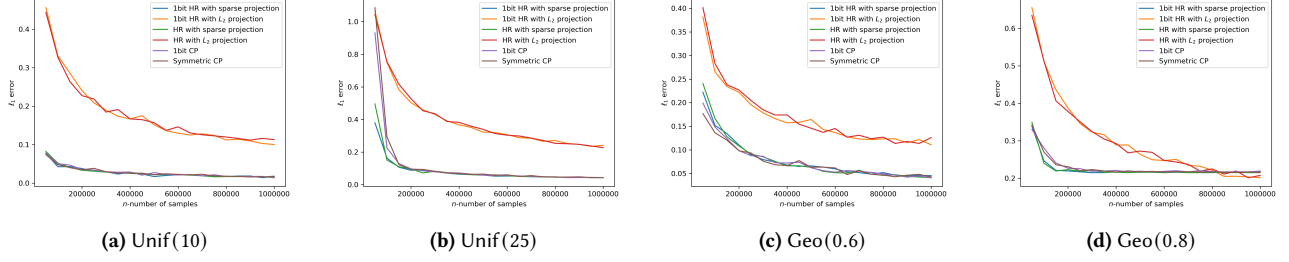


Figure 1: ℓ_1 -error for $k = 10000, m = 500, \epsilon = 1$. Sparse projection means L_2 projection onto simplex with sparsity constraint.

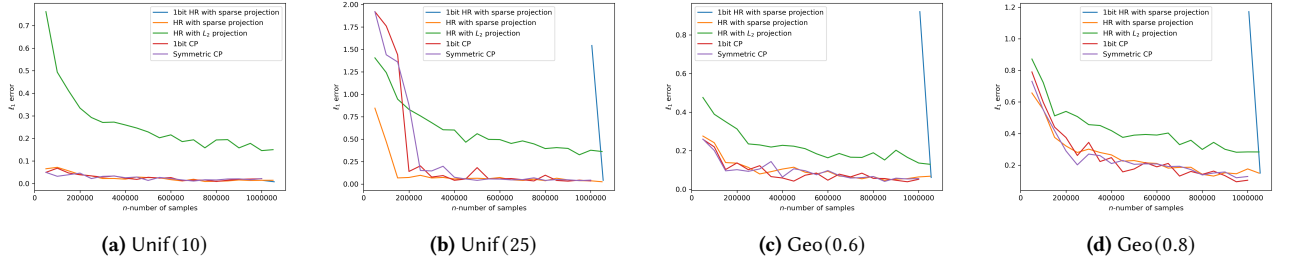


Figure 2: ℓ_1 -error for $k = 1000000, m = 500, \epsilon = 1$. For one bit HR, since the method requires the number of sample larger than k , thus the estimation result of one-bit HR starts from $n = 1000000$ in the figure.

5 EXPERIMENTS

We conduct experiments comparing our method with HR [3] and its one-bit version equipped with sparse projection. To implement our recovery process, we use the orthogonal matching pursuit (OMP) algorithm [34]. We test the performances on two types of (approximately) sparse distributions: 1) geometric distributions $\text{Geo}(\lambda)$ with $p(i) \propto (1 - \lambda)^i \lambda$; and 2) sparse uniform distributions $\text{Unif}(s)$ where $|\text{supp}(p)| = s$ and $p(i) = \frac{1}{s}$ for $i \in \text{supp}(p)$.

In our experiments, the dimensionality of the unknown distribution is $k \in \{10000, 1000000\}$, and the value of m in our method is set to 500. The default value of the privacy parameter is $\epsilon = 1$. Here we provide the results of different algorithms on $\text{Geo}(0.8)$, $\text{Geo}(0.6)$, $\text{Unif}(10)$ and $\text{Unif}(25)$.

We record the estimation errors of different methods with varying number of samples $n \in \{50000, 100000, \dots, 1000000\}$. Note that geometric distributions are not strictly sparse. For approximately sparse distributions, the sparsity parameter s is chosen such that the distributions are roughly $(s, 0.1)$ -sparse in our experiment. We assume that the value of s is provided to the recovery algorithm. We simulate 10 runs and report the average ℓ_1 errors. The results are shown in Figure 1 and Figure 2.

It can be seen from the numerical results that the performances of our compressive privatization approach are significantly better than the previous worst-case sample optimal methods like HR, which is aligned with our theoretical bounds. For small k , which is much smaller than sample size n , e.g. $k = 10000$, the one-bit HR with sparse projection is well-defined and the performance compared to our method is almost the same; this is not surprising as both method has the same (theoretical) sample complexity. Note,

HR with sparse projection is better than with non-sparse projection. On the other hand, when k is much larger e.g. $k = 1000000$, one-bit HR is not well-defined when the number of samples is less than k . In this case, we append zeros in the groups where there is no samples. However, the accuracy is much worse than our methods (see Figure 2). We note that HR with sparse projection still performs well in this case, but each user incurs $\log k$ bits of communication; our one-bit CP only needs one bit to achieve the same accuracy. For our symmetric CP method, the communication cost, which is $\log s + \log \log \frac{k}{s}$, is also lower than HR.

6 CONCLUSION

In this paper, we study sparse distribution estimation in the local differential privacy model. We propose a compressive sensing based method, which overcome the limitations of the projection based method in [1]. For high privacy regime, we provide asymmetric and symmetric schemes, both of which achieves optimal sample and communication complexity. We also extend compressive privatization to medium privacy regime, and obtain near-optimal sample complexity for any privacy and communication constraints.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China Grant No. 61802069, Shanghai Science and Technology Commission Grant No. 17JC1420200, and Science and Technology Commission of Shanghai Municipality Project Grant No. 19511120700.

REFERENCES

- [1] Jayadev Acharya, Peter Kairouz, Yuhan Liu, and Ziteng Sun. 2021. Estimating Sparse Discrete Distributions Under Privacy and Communication Constraints. In *Algorithmic Learning Theory*. PMLR, 79–98.
- [2] Jayadev Acharya and Ziteng Sun. 2019. Communication Complexity in Locally Private Distribution Estimation and Heavy Hitters. In *International Conference on Machine Learning*. 51–60.
- [3] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. 2018. Hadamard response: Estimating distributions privately, efficiently, and with little communication. *arXiv preprint arXiv:1802.04705* (2018).
- [4] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. 2008. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation* 28, 3 (2008), 253–263.
- [5] Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Özgür. 2020. Fisher information under local differential privacy. *IEEE Journal on Selected Areas in Information Theory* (2020).
- [6] Raef Bassily. 2019. Linear Queries Estimation with Local Differential Privacy. In *The 22nd International Conference on Artificial Intelligence and Statistics*. 721–729.
- [7] Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. 2017. Practical locally private heavy hitters. In *Advances in Neural Information Processing Systems*. 2288–2296.
- [8] Amos Beimel, Kobbi Nissim, and Eran Omri. 2008. Distributed private data analysis: Simultaneously solving how and what. In *Annual International Cryptology Conference*. Springer, 451–468.
- [9] Mark Bun, Jelani Nelson, and Uri Stemmer. 2018. Heavy hitters and the structure of local privacy. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 435–447.
- [10] T Tony Cai and Anru Zhang. 2013. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE transactions on information theory* 60, 1 (2013), 122–132.
- [11] Emmanuel J Candès, Justin Romberg, and Terence Tao. 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory* 52, 2 (2006), 489–509.
- [12] Emmanuel J Candès, Justin K Romberg, and Terence Tao. 2006. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59, 8 (2006), 1207–1223.
- [13] Emmanuel J Candès and Terence Tao. 2005. Decoding by linear programming. *IEEE transactions on information theory* 51, 12 (2005), 4203–4215.
- [14] Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür. 2020. Breaking the Communication-Privacy-Accuracy Trilemma. *arXiv preprint arXiv:2007.11707* (2020).
- [15] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting Telemetry Data Privately. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- [16] David L Donoho. 2006. Compressed sensing. *IEEE Transactions on information theory* 52, 4 (2006), 1289–1306.
- [17] Marco F Duarte and Richard G Baraniuk. 2011. Kronecker compressive sensing. *IEEE Transactions on Image Processing* 21, 2 (2011), 494–504.
- [18] Devdatt P Dubhashi and Desh Ranjan. 1996. Balls and bins: A study in negative dependence. *BRICS Report Series* 3, 25 (1996).
- [19] John Duchi and Ryan Rogers. 2019. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*. PMLR, 1161–1191.
- [20] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2013. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 429–438.
- [21] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.
- [22] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3-4 (2014), 211–407.
- [23] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. 1054–1067.
- [24] Vitaly Feldman and Kunal Talwar. 2021. Lossless Compression of Efficient Private Local Randomizers. *arXiv preprint arXiv:2102.12099* (2021).
- [25] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. 2016. Discrete distribution estimation under local privacy. *arXiv preprint arXiv:1602.07387* (2016).
- [26] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2014. Extremal mechanisms for local differential privacy. In *Advances in neural information processing systems*. 2879–2887.
- [27] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. 2015. On learning distributions from their samples. In *Conference on Learning Theory*. 1066–1100.
- [28] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2011. What can we learn privately? *SIAM J. Comput.* 40, 3 (2011), 793–826.
- [29] Erich L Lehmann and George Casella. 2006. *Theory of point estimation*. Springer Science & Business Media.
- [30] Adriano Pastore and Michael Gastpar. 2016. Locally differentially-private distribution estimation. In *2016 IEEE International Symposium on Information Theory (ISIT)*. Ieee, 2694–2698.
- [31] Ingo Roth, Axel Flinth, Richard Kueng, Jens Eisert, and Gerhard Wunder. 2018. Hierarchical restricted isometry property for Kronecker product measurements. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 632–638.
- [32] Ingo Roth, Martin Kliesch, Gerhard Wunder, and Jens Eisert. 2016. Reliable recovery of hierarchically sparse signals and application in machine-type communications. *arXiv preprint arXiv:1612.07806* (2016).
- [33] Apple Differential Privacy Team. 2017. Learning with privacy at scale. (2017).
- [34] Joel A Tropp. 2004. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory* 50, 10 (2004), 2231–2242.
- [35] David Wajc. 2017. Negative association: definition, properties, and applications. *Manuscript, available from https://goo.gl/j2ekqM* (2017).
- [36] Di Wang and Jinhui Xu. 2019. On sparse linear regression in the local differential privacy model. In *International Conference on Machine Learning*. PMLR, 6628–6637.
- [37] Shaowei Wang, Liusheng Huang, Pengzhan Wang, Yiwen Nie, Hongli Xu, Wei Yang, Xiang-Yang Li, and Chunming Qiao. 2016. Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025* (2016).
- [38] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally Differentially Private Protocols for Frequency Estimation. In *26th USENIX Security Symposium (USENIX Security 17)*. USENIX Association.
- [39] Tianhao Wang, Ninghui Li, and Somesh Jha. 2019. Locally differentially private heavy hitter identification. *IEEE Transactions on Dependable and Secure Computing* (2019).
- [40] Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* 60, 309 (1965), 63–69.
- [41] Min Ye and Alexander Barg. 2018. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory* 64, 8 (2018), 5662–5676.

A MISSING PROOF FROM SECTION 3

A.1 Proof of Lemma 3.1

PROOF. Observe that for any $x_1, x_2 \in [k]$, we have

$$\max_{y \in [m]} \frac{Q(y|x_1)}{Q(y|x_2)} \leq \frac{n_{x_2} e^\varepsilon + m - n_{x_2}}{n_{x_1} e^\varepsilon + m - n_{x_1}} e^\varepsilon.$$

By assumption,

$$\begin{aligned} \frac{n_{x_2} e^\varepsilon + m - n_{x_2}}{n_{x_1} e^\varepsilon + m - n_{x_1}} &\leq \frac{((1+\beta)\frac{m}{2})e^\varepsilon + m - ((1+\beta)\frac{m}{2})}{((1-\beta)\frac{m}{2})e^\varepsilon + m - ((1-\beta)\frac{m}{2})} \\ &= \frac{1+\beta \cdot \frac{e^\varepsilon-1}{e^\varepsilon+1}}{1-\beta \cdot \frac{e^\varepsilon-1}{e^\varepsilon+1}} \leq \frac{1+\beta/2}{1-\beta/2} \leq 1+2\beta. \end{aligned}$$

where the second inequality is from $\frac{e^\varepsilon-1}{e^\varepsilon+1} \leq \frac{1}{2}$ for $\varepsilon \in (0, 1)$ and the last inequality is from $0 \leq \beta \leq 1$. It follows that

$$\begin{aligned} \max_{x_1, x_2, y \in [k]} \frac{Q(y|x_1)}{Q(y|x_2)} &\leq \max_{x_1, x_2 \in [k]} \frac{n_{x_2} e^\varepsilon + m - n_{x_2}}{n_{x_1} e^\varepsilon + m - n_{x_1}} e^\varepsilon \\ &\leq (1+2\beta)e^\varepsilon = e^{\varepsilon+\ln(1+2\beta)} \leq e^{\varepsilon+2\beta}. \end{aligned}$$

The last inequality is from $\ln(1+x) \leq x$ for $x > -1$. \square

A.2 Proof of Lemma 3.2

PROOF. By definition

$$\begin{aligned} \|e_1\|_2 &= \left\| \frac{e^\varepsilon + 1}{\sqrt{m}(e^\varepsilon - 1)} J(D' - I)p \right\|_2 \\ &= \frac{e^\varepsilon + 1}{\sqrt{m}(e^\varepsilon - 1)} \sqrt{((D' - I)p)^T \cdot J^T J \cdot ((D' - I)p)} \\ &\leq \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \sum_i |d'_i - 1| p_i, \end{aligned}$$

where $d'_i = \frac{m(e^\varepsilon+1)/2}{n_i e^\varepsilon + m - n_i}$. By the assumption $n_i \in (1 \pm \beta)\frac{m}{2}$ for all $i \in [k]$, we have

$$\begin{aligned} |d'_i - 1| &= \frac{|m/2 - n_i|(e^\varepsilon - 1)}{n_i e^\varepsilon + m - n_i} \leq \frac{\beta m(e^\varepsilon - 1)/2}{n_i e^\varepsilon + m - n_i} \\ &\leq \frac{\beta m(e^\varepsilon - 1)/2}{(1-\beta)\frac{m}{2}e^\varepsilon + \frac{m}{2} - \frac{\beta m}{2}} = \frac{\beta(e^\varepsilon - 1)}{(1-\beta)(e^\varepsilon + 1)}. \end{aligned}$$

Thus, $\|e_1\|_2 \leq \frac{\beta}{1-\beta}$, which completes the proof. \square

A.3 Proof of Lemma 3.3

PROOF. Let Y_1, Y_2, \dots, Y_n be the privatized samples received by the server. We have, for each $i \in [m]$, $\hat{q}_i = \sum_{j=1}^n \frac{\mathbb{I}(Y_j=i)}{n}$, where \mathbb{I} is the indicator function. Thus $\mathbb{E}[\hat{q}_i] = q_i$ and $\text{Var}[\hat{q}_i] = \frac{q_i(1-q_i)}{n} \leq \frac{q_i}{n}$. It follows that

$$\mathbb{E}[\|\hat{q} - q\|_2] \leq \sqrt{\mathbb{E}[\|\hat{q} - q\|_2^2]} = \sqrt{\sum_i \text{Var}(\hat{q}_i)} \leq \sqrt{\frac{1}{n}}$$

where the first inequality is from Jensen's inequality. Multiplying $\frac{\sqrt{m}e^\varepsilon+1}{e^\varepsilon-1}$ on both sides of the inequality will conclude the proof. \square

A.4 Proof of Theorem 3.4

PROOF. Let $p' = D'p$. By definition of \hat{p}' , we have

$$\|p - \hat{p}'\|_2 = \|D'^{-1}p - D'^{-1}f\|_2 \leq \max_i \frac{1}{d'_i} \cdot \|p' - f\|_2$$

where $d'_i = \frac{m(e^\varepsilon+1)/2}{n_i e^\varepsilon + m - n_i}$ and f is the output of the recovery algorithm (step 7 in Algorithm 2). Since $\forall i \in [k]$, $n_i \leq (1+\beta)\frac{m}{2}$ for $0 \leq \beta \leq 1$, then we have

$$\begin{aligned} \|p - \hat{p}'\|_2 &\leq \max_i \frac{1}{d'_i} \cdot \|p' - f\|_2 \leq (1+\beta) \cdot \frac{e^\varepsilon - 1}{e^\varepsilon + 1} \|p' - f\|_2 \\ &\leq (1 + \frac{\beta}{2}) \|p' - f\|_2. \end{aligned}$$

By Lemma 1.3, we have

$$\begin{aligned} \|p' - f\|_2 &\leq \frac{C}{\sqrt{s}} \|p' - [p']_s\|_1 + D \|e_1 + e_2\|_2 \\ &\leq \max_i \frac{1}{d'_i} \cdot \frac{C}{\sqrt{s}} \|p - [p]_s\|_1 + D (\|e_1\|_2 + \|e_2\|_2). \end{aligned}$$

Note $\max_i \frac{1}{d'_i} \leq (1+\beta/2)$ as $n_i \leq (1+\beta)\frac{m}{2}$ for all i . By Lemma 3.2, 3.3, we get

$$\begin{aligned} \mathbb{E}[\|p - \hat{p}'\|_2] &\leq \left(1 + \frac{\beta}{2}\right) \left(2D\beta + \frac{D(e^\varepsilon + 1)}{e^\varepsilon - 1} \sqrt{\frac{m}{n}}\right) \\ &\quad + \left(1 + \frac{\beta}{2}\right)^2 \left(\frac{C}{\sqrt{s}} \|p - [p]_s\|_1\right), \end{aligned}$$

which proves the theorem. \square

B MISSING PROOF FROM SECTION 4

B.1 Proof of Lemma 4.1

PROOF. By symmetry, we only consider the sparsity of one specific block, say the first one. Let $k_1 = \frac{k}{L}$ be the size of a block. Let s_1 denote the sparsity, i.e. the number of non-zero entries, of the first block. Then, we have

$$s_1 = \sum_{j=1}^{k_1} \mathbf{1}\{p'_j \neq 0\}$$

where $\mathbf{1}\{p'_j \neq 0\}$ is an indicator to describe whether the j -th position of p' is nonzero. By direct calculation, we can get that $\mathbb{E}[\mathbf{1}\{p'_j \neq 0\}] = \frac{\binom{k-1}{s-1}/\binom{k}{s}}{\binom{k-1}{s-1}/\binom{k}{s}} = \frac{s}{k}$. Thus, $\mathbb{E}[s_1] = \frac{s k_1}{k} = \frac{s}{L}$. Since p' is a random permutation of p , $\mathbf{1}\{p'_1 \neq 0\}, \dots, \mathbf{1}\{p'_{k_1} \neq 0\}$ are negatively associated (NA) [35]. By Chernoff-Hoeffding bounds for NA variables [18, 35], we can get that

$$\begin{aligned} \Pr[s_1 \geq (1+\beta)\mathbb{E}[s_1]] &\leq \left(\frac{e^\beta}{(1+\beta)^{(1+\beta)}}\right)^{\mathbb{E}[s_1]} \\ &= e^{\mathbb{E}[s_1](\beta - (1+\beta)\ln(1+\beta))} \end{aligned} \quad (19)$$

It can be easily verified that

$$\beta - (1+\beta)\ln(1+\beta) \leq \begin{cases} -\frac{\beta^2}{4} & \beta \leq 4, \\ -\frac{\beta}{4} & \text{otherwise} \end{cases} \leq -\frac{1}{4} \min\{\beta^2, \beta\} \quad (20)$$

Combining (19), (20) and $\mathbb{E}[s_1] = \frac{s}{L}$ yields that

$$\Pr \left[s_1 \geq (1 + \beta) \frac{s}{L} \right] \leq e^{-\frac{\min\{\beta^2, \beta\}s}{4L}} \quad (21)$$

The proof is then completed by applying union bound over all L sections. \square

B.2 Proof of Lemma 4.2

PROOF. For any $j \in [m]$, we have

$$\mathbb{E}[\hat{Q}_j(X'_i)] = \mathbb{E}_p[\mathbb{E}_\varepsilon[\hat{Q}_j(x)|X'_i = x]] \quad (22)$$

When X'_i is fixed to be $x \in \mathcal{Q}$ where $\mathcal{Q} := \{\pm e_1, \dots, \pm e_L\}$ and we only consider the randomness from the privatization, we have

$$\begin{aligned} \mathbb{E}_\varepsilon[\hat{Q}_j(x)] &= \frac{e^\varepsilon}{e^\varepsilon + 2L - 1} Q_j(x) + \sum_{Q' \in \mathcal{Q} \setminus \{Q_j(x)\}} \frac{1}{e^\varepsilon + 2L - 1} \\ &= \frac{(e^\varepsilon - 1)}{e^\varepsilon + 2L - 1} \cdot Q_j(x) \end{aligned} \quad (23)$$

By the definition of Q_j , we can get

$$\mathbb{E}[Q_j(X'_i)] = \begin{bmatrix} (A_2)_j \cdot p'^{(1)} \\ (A_2)_j \cdot p'^{(2)} \\ \vdots \\ (A_2)_j \cdot p'^{(L)} \end{bmatrix} \quad (24)$$

Combining (22), (23) and (24) yields that

$$\mathbb{E}[\hat{Q}_j(X'_i)] = \frac{(e^\varepsilon - 1)}{e^\varepsilon + 2L - 1} \begin{bmatrix} (A_2)_j \cdot p'^{(1)} \\ (A_2)_j \cdot p'^{(2)} \\ \vdots \\ (A_2)_j \cdot p'^{(L)} \end{bmatrix} \quad (25)$$

Recall that $q = A \cdot p'$ and $A = A_1 \otimes A_2$. For $j' \equiv j \pmod{m}$ and $j' = j + (t-1)m$, by (25), we have

$$\begin{aligned} \mathbb{E}[\hat{q}_{j'}] &= \frac{m}{n} \cdot \frac{e^\varepsilon + 2L - 1}{e^\varepsilon - 1} \sum_{i \in S_{j'}} (A_1)_t \cdot \mathbb{E}[\hat{Q}_j(X'_i)] \\ &= (A_1)_t \cdot \left[(A_2)_j \cdot p'^{(1)}, \dots, (A_2)_j \cdot p'^{(L)} \right] = q_{j'} \end{aligned}$$

where the last equality is from the definition of kronecker product. Hence, $\hat{q}_{j'}$ is an unbiased estimator for $q_{j'}$. Thus,

$$\begin{aligned} \mathbb{E}[(q_{j'} - \hat{q}_{j'})^2] &= \text{Var}(q_{j'}) = \frac{m}{n} \left(\frac{e^\varepsilon + 2L - 1}{e^\varepsilon - 1} \right)^2 \text{Var} \left((A_1)_t \cdot \hat{Q}_j(X'_i) \right) \\ &\leq \frac{m}{n} \left(\frac{e^\varepsilon + 2L - 1}{e^\varepsilon - 1} \right)^2 \end{aligned}$$

where the inequality is from that $(A_1)_t \cdot \hat{Q}_j(X'_i)$ only takes value in $\{+1, -1\}$. The proof is completed. \square

B.3 Proof of Theorem 4.3

PROOF. From the analysis of estimation error in section 4, we know that

$$\mathbb{E}[\|p - \hat{p}\|_2] \leq \mathbb{E}[\|p' - \hat{p}'\|_2] \leq \frac{1}{\sqrt{mL}} \mathbb{E}[\|\hat{q} - q\|_2]$$

By Lemma 4.2, we can get

$$\begin{aligned} \mathbb{E}[\|p - \hat{p}\|_2] &\leq \mathbb{E}[\|p' - \hat{p}'\|_2] \leq \frac{1}{\sqrt{mL}} \mathbb{E}[\|\hat{q} - q\|_2] \\ &\stackrel{(a)}{\leq} \sqrt{\frac{1}{mL} \mathbb{E}[\|\hat{q} - q\|_2^2]} \\ &= \sqrt{\frac{1}{mL} \sum_{j' \in [mL]} \mathbb{E}[(\hat{q}_{j'} - q_{j'})^2]} \\ &\stackrel{(b)}{\leq} \sqrt{\frac{m}{n} \left(\frac{e^\varepsilon + 2L - 1}{e^\varepsilon - 1} \right)} \stackrel{(c)}{\leq} \sqrt{\frac{m}{n} \left(\frac{3e^\varepsilon - 1}{e^\varepsilon - 1} \right)} \end{aligned}$$

where (a) is from Jensen's inequality and (b) is from Lemma 4.2 and (c) is from $L = \min\{e^\varepsilon, 2^b\}$.

- (1) $\varepsilon = O(1)$. In this case, we can set $L = 1$ directly and the communication is 1 bit now. The event \mathcal{E} then holds with probability 1, hence $m = s \log(k/s)$. Since $\varepsilon = O(1)$, $\frac{3e^\varepsilon - 1}{e^\varepsilon - 1} = O(\frac{1}{\varepsilon})$. Thus $\mathbb{E}[\|p - \hat{p}\|_2] = O\left(\sqrt{\frac{s \log(k/s)}{ne^2}}\right)$, which is the same error bound as that in one-bit CP for high privacy. Note that A_1 is 1 now, $A = A_1 \otimes A_2 = A_2 \in \mathbb{R}^{m \times k}$ is a Rademacher matrix. Therefore, for high privacy, if we set $L = 1$, our scheme is exactly one-bit CP.
- (2) $\varepsilon = \omega(1)$. We mainly consider medium privacy case, where $\varepsilon = \omega(1)$ and $e^\varepsilon \leq s$. In this case, $\frac{3e^\varepsilon - 1}{e^\varepsilon - 1} = O(1)$. Hence, we have that $\mathbb{E}[\|p - \hat{p}\|_2] = O\left(\sqrt{\frac{(1+\beta)s \log(k/(1+\beta)s)}{nL}}\right)$ (note $m = O\left(\frac{(1+\beta)s \log(k/(1+\beta)s)}{L}\right)$). When $n \geq c \cdot \frac{(1+\beta)s \log(k/(1+\beta)s)}{L\alpha^2}$ for some large enough constant c , the expected ℓ_2 error is at most α . For ℓ_1 error, we have

$$\mathbb{E}[\|p - \hat{p}\|_1] \leq 2\sqrt{2s} \mathbb{E}[\|p' - \hat{p}'\|_2]$$

Thus to achieve an ℓ_1 error of α , it's sufficient to get an estimate with $\alpha' = \alpha/2\sqrt{2s}$ for $\mathbb{E}[\|p' - \hat{p}'\|_2]$. The sample complexity is $O\left(\frac{(1+\beta)s^2 \log(k/(1+\beta)s)}{L\alpha^2}\right)$. Since the event \mathcal{E} holds

with probability at least $1 - Le^{-\frac{\min\{\beta^2, \beta\}s}{4L}}$ and the RIP condition holds with probability at least $1 - e^{-m}$, by union bound, we can achieve the sample complexity above with probability $1 - Le^{-\frac{\min\{\beta^2, \beta\}s}{4L}} - e^{-m}$, where $m = \frac{(1+\beta)s \log(k/(1+\beta)s)}{L}$. When $\min\{\beta^2, \beta\} = \frac{4L \log(L/\delta)}{s}$, the error probability from \mathcal{E} is less than δ . If $L \log(L/\delta) = O(s)$, then $\min\{\beta^2, \beta\} = O(1)$ which means $\beta = O(1)$. In this case, with probability $1 - \delta - e^{-m}$, the sample complexity for ℓ_2 error is $O\left(\frac{s \log(k/s)}{L\alpha^2}\right)$ and for ℓ_1 error is $O\left(\frac{s^2 \log(k/s)}{L\alpha^2}\right)$. When $s \ll L \log(L/\delta)$ and $L \leq s$, $\min\{\beta^2, \beta\} = O(\log(s/\delta))$ which means $\beta = O(\log(s/\delta))$. For general distribution under medium privacy regime, the sample complexity for ℓ_1 error is at least $\Omega\left(\frac{k^2}{L\alpha^2}\right)$ [14], which implies a lower bound of $\Omega\left(\frac{s^2}{L\alpha^2}\right)$ for s -sparse distributions. Thus the sample complexity blows up by at most a logarithmic factor. \square