Compressive Sensing Approaches for Sparse Distribution Estimation Under Local Privacy

Zhongzheng Xiong
zzxiong21@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
yingshan@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.

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/authors. Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9096-5/22/04. $15.00 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 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, 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 privatization mechanism is a randomized mapping $Q$ to the central server, who will then produce the final estimate. Each user will only send a privatized version of their data $d$. Given the target distribution is defined on the universe $X = [k] := \{1, 2, \cdots, k\}$, which can be viewed as a $k$-dimensional vector $p \in \mathbb{R}^k$ with $\|p\|_1 = 1$. Let $\Delta_k$ be the set of all $k$-ary discrete distributions. Given $n$ i.i.d. samples, $X_1, \cdots, X_n$, drawn from the unknown distribution $p$, the goal is to provide an estimator $\tilde{p}$ such that $d(p, \tilde{p})$ is minimized, where $d(\cdot, \cdot)$ 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 produce the final estimate. A privatization mechanism is a randomized mapping $Q$ that maps $x \in X$ to $y \in \mathcal{Y}$ with probability $Q(y|x)$ for some output set $\mathcal{Y}$. The mapping $Q$ is said to be $\varepsilon$-locally differential private (LDP) [20] if for all $x, x' \in X$ and $y \in \mathcal{Y}$, we have

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

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

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 $(s, \lambda)$-sparse, if $\|p - |p|_s\|_1$ is smaller than $\lambda$.

Our results. For the high privacy regime, i.e., $\varepsilon = O(1)$, existing studies [3, 23, 25, 37, 40, 41] have achieved optimal sample complexity, which is $\Theta\left(\frac{k}{\alpha^2}\right)$ for $l_1$ norm error and $\Theta\left(\frac{k}{\alpha^2}\right)$ 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 $0 < \varepsilon < 1$ and $\alpha > 0$, there is an $\varepsilon$-LDP scheme $Q$, which produces an estimator $\hat{Q}$ with error guarantee $d(Q(p), \hat{Q}) \leq \alpha$. If $p$ is $s$-sparse, then the sample complexity of $Q$ is $O\left(\frac{s^2 \log (k/s)}{e^\varepsilon \alpha^2}\right)$ for $l_1$ error and $O\left(\frac{s \log (k/s)}{e^\varepsilon \alpha^2}\right)$ 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 $1 \leq \varepsilon \leq \log s$ (see Section 4). The result is summarized in the following theorem. See Theorem 4.3 for exact bounds.

**Theorem 1.2 (Informal).** For any $1 \leq \varepsilon^2, 2^b \leq s$ and $\alpha > 0$, there is an $\varepsilon$-LDP scheme $Q$, which produces an estimator $\hat{Q}$ with error guarantee $d(Q(p), \hat{Q}) \leq \alpha$ with communication no more than $b$ bits. If $p$ is $s$-sparse, then the sample complexity of $Q$ is $O\left(\frac{s^2 \log (k/s)}{\min\{e^\varepsilon, 2^b\} \alpha^2}\right)$ for $l_1$ error and $O\left(\frac{s \log (k/s)}{\min\{e^\varepsilon, 2^b\} \alpha^2}\right)$ 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 $l$-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 $\mathbf{x}$ be an unknown $k$-dimensional vector. The goal of compressive sensing (CS) is to reconstruct $\mathbf{x}$ from only a few linear measurements $\{y_i\}_{i=1}^n$. To be precise, let $\mathbf{B} \in \mathbb{R}^{m \times k}$ be the measurement matrix with $m \ll k$ and $\mathbf{e} \in \mathbb{R}^m$ 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 $(\delta, \sigma)$-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 $(2\sqrt{s}, 1/\sqrt{2})$-RIP. Given $y = Bx + e$, there is a polynomial time algorithm, which outputs $\hat{x}$ that satisfies $\|x - \hat{x}\|_2 \leq C \sqrt{s} \|x - [x]_s\|_1 + D \|e\|_2$ 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 $\mathbf{x}$ be a $k_1 k_2$-dimensional vector consists of $k_1$ blocks, each of size $k_2$ (e.g., $\mathbf{x} = [x^{(1)}, \cdots, x^{(k_1)}]$). Then $\mathbf{x}$ is $(s, \sigma)$-hierarchically sparse if at most $s$ blocks have non-zero entries and each of these blocks is $\sigma$-sparse.

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

$$\|x - \hat{x}\|_2 \leq (1 - \delta) \|x\|_2 \leq (1 + \delta) \|x\|_2.$$  

Theorem 1.4 (HiRIP of Kronecker product [31]). For any matrix $A \in \mathbb{R}^{m \times k}$ that is $(s, \sigma)$-RIP and any matrix $B \in \mathbb{R}^{m \times k}$ that is $(\sigma, \delta_2)$-RIP, $A \otimes B \in \mathbb{R}^{m \times kk}$ satisfies (s, $\sigma$)-HiRIP with constant $\delta_3$, where $\delta_3 = \delta_1 + \frac{\delta_2 + \delta_1 \delta_2}{2}$.

Theorem 1.5 (Recovery guarantee for hierarchically sparse vectors [32]). Suppose matrix $A \in \mathbb{R}^{m \times k}$ is $(3s, 2\sigma)$-HiRIP with constant $\frac{1}{\sqrt{3}}$. Given $y = A\mathbf{x} + e$ where $\mathbf{x} \in \mathbb{R}^k$ is $(s, \sigma)$-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 $S_1, \cdots, S_K$ be the groups. Since the partition can be arbitrary, we assume $S_j := \{i \in [n] | i \in j \ mod \ K\}$. 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 $X_j$ sends a bit $Y_i \in \{0, 1\}$ distributed as

$$\Pr(Y_i = 1) = \frac{e^\varepsilon + 1}{e^\varepsilon - 1} (2 \sqrt{t} - 1) = H_K \cdot p.$$  

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

$$\Pr(Y_i = 1) = \frac{e^\varepsilon + 1}{e^\varepsilon - 1} (2 \sqrt{t} - 1) = H_K \cdot p.$$  

Let $\tilde{t} := (\tilde{t}_1, \cdots, \tilde{t}_K)$ where $\tilde{t}_j := \frac{1}{|S_j|} \sum_{i \in S_j} Y_i$ is the fraction of messages from $S_j$ that are 1. Then $\tilde{t}$ is an unbiased estimator of $t$. Then $\tilde{t}$ is an unbiased estimator 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 $\tilde{t}_j$ is undefined. Moreover, the proof of [1] relies on the fact that each $\tilde{t}_j$ is the average of $|S_j|$ i.i.d. Bernoulli random variables, which implies $\tilde{t}_j - t_j$ 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 $A \in \{-1, 1\}^{n \times k}$ that satisfies the RIP condition will suffice. More specifically, given
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 \( \alpha \), the sample complexity of our method is \( O\left(\frac{m}{\epsilon^{\alpha} + \epsilon^2} \right) \) for \( \ell_2 \) error and \( O\left(\frac{sm}{\epsilon^{\alpha} + \epsilon^2} \right) \) for \( \ell_1 \) error. The \( \ell_2 \) result also holds for \((s, \sqrt{s})\)-RIP and the \( \ell_1 \) result also holds for \((s, \alpha)\)-RIP. The communication cost for each user is \( 1 \) bit.

Proof. We first consider the case for \( \ell_2 \) error. Since \( \Delta_k \) is convex, \( \|\hat{p} - p\|_2 \leq \|\tilde{p} - \hat{p}\|_2 \). By Lemma 1.3, we know:

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

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

\[
\mathbb{E}[\|\tilde{t} - t\|_2] \leq \mathbb{E}[\|\tilde{t} - t\|_2^2] = \sum_{i=1}^{m} \frac{1}{|S_i|} \sum_{j \in S_i} \text{Var}(Y_j) \leq \frac{m^2}{4n},
\]

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

\[
\mathbb{E}[\|p - \tilde{p}\|_2] \leq \frac{C}{\sqrt{2}}\|p - \hat{p}\|_1 + D \cdot \frac{\sqrt{2}m}{\epsilon^2 - 1} \sqrt{\frac{m}{n}}.
\]

When \( p \) is \( s \)-sparse, the first term in (9) is \( 0 \). Thus, when \( n \geq \frac{8m^2}{\epsilon^2 \alpha^2} \) for some large enough constant \( \alpha \), the expected \( \ell_2 \) error is at most \( \epsilon \). For \((s, \sqrt{s\alpha})\)-RIP, the first error term in (9) is bounded by \( O(\alpha) \), thus the result still holds for approximately sparse case. For \( \ell_1 \) 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

\[
\|p - \tilde{p}\|_1 \leq \|p - \hat{p}\|_1 + \|\tilde{p} - \hat{p}\|_1 \leq 2\|\tilde{p} - \hat{p}\|_1 \leq 2\sqrt{2}\|p - \tilde{p}\|_2.
\]

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 \( \ell_1 \) error of \( \alpha \), it’s sufficient to get an estimate with \( s' = \alpha / \sqrt{2} \) error for \( \ell_2 \). The sample complexity is \( O(sm/\epsilon^{\alpha} \alpha^2) \). For \( \ell_1 \) error with \( p \) being \((s, \alpha)\)-RIP, now \( \hat{p} - p \) is \((2s, \alpha)\)-RIP. We have

\[
\|p - \tilde{p}\|_1 \leq \|p - \hat{p}\|_2 + \|\tilde{p} - \hat{p}\|_1 \leq \|\tilde{p} - \hat{p}\|_2 + \sqrt{2}\|p - \tilde{p}\|_2.
\]

Then, the \( \ell_1 \) result follows by a similar argument as for the exact sparse case.

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 \( m \geq O\left(s \log \frac{k}{\epsilon} \right) \), \( B \) satisfies \((s, 1/\sqrt{2})\)-RIP with probability \( 1 - \epsilon^{-m} \) [4]. By Theorem 2.1, we have the following theorem.

Theorem 2.2. For \( m = O\left(s \log \frac{k}{\epsilon} \right) \) and \( \epsilon = O(1) \), if the entries of \( A \) are i.i.d. Rademacher random variables, then with probability at least \( 1 - \epsilon^{-m} \), our method has sample complexity \( O\left(\frac{s \log (k/\alpha)}{\epsilon^{\alpha} \alpha^2} \right) \) for \( \ell_2 \) error, and \( O\left(\frac{s \log (k/\epsilon)}{\epsilon^{\alpha} \alpha^2} \right) \) for \( \ell_1 \) error. The \( \ell_2 \) result holds for \((s, \sqrt{s})\)-RIP and the \( \ell_1 \) result holds for \((s, \alpha)\)-RIP. The communication cost for each user is \( 1 \) bit.

Lower bound. [1] proves a lower bound of \( n = \Omega(\frac{s^2 \log (k/\epsilon)}{\epsilon^{\alpha} \alpha^2}) \) on the sample complexity for \( \ell_1 \) 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
We first show how to model the estimation of $q$, whose column sums are close to 0. The key is to design an appropriate privatization, which could be easier to implement in real applications.

Our mechanism $\hat{q}$ 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 $x$:

$$Q(y|x) := \begin{cases} \frac{e^x}{\sum_{y \in C_x} e^y}, & \text{if } y \in C_x, \\ 0, & \text{if } y \in [m]\backslash C_x. \end{cases}$$

(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 $I_x(j)$ for all $j \in [k]$, 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 $\alpha$, privacy parameter $\epsilon$ and target sparsity $s$, we require $A$ to have the following 2 properties:

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

$$q_j = \sum_{i=1}^{k} p_i \cdot Q(Y = j|X = i)$$

$$= \sum_{i,j \in C_i} \frac{e^x}{\sum_{y \in C_x} e^y} + \sum_{j \in [k]\backslash C_i} \frac{p_i}{\sum_{y \in C_x} e^y + m - n_i}.$$ 

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

$$q = \begin{pmatrix} e^x - 1 \\ e^x + 1 \end{pmatrix} D p,$$

where $D$ is the all-one matrix and $D$ is the diagonal matrix with $d_i = \frac{1}{n_i e^x + m - n_i}$, in the $i$th diagonal entry. 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

$$(e^x + 1) \left( \frac{\sqrt{m} e^x - 1}{\sqrt{m}} \right) = \frac{1}{\sqrt{m}} AD'p + \frac{e^x + 1}{\sqrt{m}(e^x - 1)} (D' - I)p \epsilon$$

$$+ \frac{\sqrt{m}(e^x + 1)}{e^x - 1} (q - q').$$

(13)

Given the LHS of (13), $\frac{1}{\sqrt{m}} A$ and a target sparsity $s$, we reconstruct $D'p$ by applying Lemma 1.3. Then compute $\hat{p}' = D'^{-1}D'p$ ($D'$ 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 3.2**. For a fixed $\beta \in [0, 0.5]$, if $n_i \in (1 \pm \frac{\beta}{2})$ for all $i \in [k]$, then $\|e_1\|_2 \leq \frac{\beta}{1 - \beta} \leq 2\beta$.

**Lemma 3.1** (Privacy Guarantee). If $(1 - \beta) \frac{\alpha}{2} \leq n_i \leq (1 + \beta) \frac{\alpha}{2}$ 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 $(1 - \epsilon) \frac{\alpha}{2} \leq n_i \leq (1 + \epsilon) \frac{\alpha}{2}$ for all $i \in [k]$, then $Q$ is $3\epsilon$-LDP. We can rescale $\epsilon$ in the beginning by a constant to ensure $\epsilon$-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 $\ell_2$ norm of $e_1$ and $e_2$ in (13).
Algorithm 2: Symmetric compressive privatization

Result: $\hat{p} \in \Delta_k$: an estimate of $p$

Input: $X_1, \ldots, X_n$ i.i.d from $p$, privacy parameter $\epsilon$, sparsity $s$, $A \in \mathbb{R}^{n \times k}$

1. For all $x \in \mathcal{X}$, 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

$$\Pr[Y_i = y | X_i = x] = \left\{ \begin{array}{ll}
\frac{\epsilon^2}{n_x \epsilon^2 + m - n_x} & y \in C_x, \\
\frac{1}{n} & \text{otherwise}
\end{array} \right.$$  

2. $\tilde{q} = (0, 0, \ldots 0) \in \mathbb{R}^m, 1 = (1, 1, \ldots, 1) \in \mathbb{R}^m$

3. For $i = 1 \text{ to } m$

4. $\tilde{q}[i] = \frac{\sum_{j=1}^n (Y_j = i)}{n}$

5. End

6. Apply Lemma 1.3, with $y = \frac{(e^\epsilon + 1)}{(\epsilon - 1)} (\sqrt{m} - 1)$, $B = \frac{1}{\sqrt{m}} A$

7. Compute $\hat{p}' = D^{-1} \tilde{f}$, and then compute the projection of $\hat{p}'$ onto the set $\Delta_k$, denoted as $\hat{p}$

8. Return $\hat{p}$

Lemma 3.3. $\mathbb{E}[\|\epsilon_2\|_2] \leq \left(1 + \frac{2}{\epsilon - 1}\right) \sqrt{\frac{m}{n}}$.

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 $C$

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

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

Corollary 3.4.1. If $A$ satisfies two properties in Section 3 with $\beta \leq ca/4D$, with $c = 1$ for $\epsilon_2$ error and $c = \frac{1}{\sqrt{2}}$ for $\epsilon_1$ error, and $p$ is $s$ sparse, the sample complexity of our symmetric compressive privatization scheme is $O(\frac{m}{n})$ for $\epsilon_2$ error and $O(\frac{m}{\epsilon^2})$ for $\epsilon_1$ error.

Theorem 3.5. For $m = \Omega(s \log \frac{1}{\epsilon})$, if $A$ is defined as above, then with probability at least $1 - e^{-m}$, our symmetric compressive privatization scheme has sample complexity $O\left(\frac{m}{\epsilon^2} \right)$ for $\epsilon_2$ error, and $O\left(\frac{m}{\epsilon^2} \right)$ for some $\epsilon_1$ error. The $\epsilon_2$ result holds for $(s, \sqrt{s}a)$-sparse $p$ and the $\epsilon_1$ result holds for $(s, \alpha)$-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 $\log s - 2$. Our symmetric scheme requires $\log s + \log \log k$ bits, which is optimal up to a log $\log k$ additive term.

4. RECURSIVE COMPRESSION 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 $1 < \epsilon^2 < k$, the relationship between sample complexity, privacy, and communication cost become more complicated. Recently, [14] provides a clean characterization for any $\epsilon$ 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 $1 \leq \epsilon^2 \leq s$.

Our RCP scheme (shown in Algorithm 3) consists of three steps including random permutation, privatization, and estimation. Let $X$ be a sample scheme 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 condition. In this section, we provide an analogous result for sparse distribution estimation for privacy regime where $1 \leq \epsilon^2 \leq s$.

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 $E$ be the event that

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

for some $\beta > 0$.
Compressive Sensing Approaches for Sparse Distribution Estimation Under Local Privacy

Algorithm 3: Recursive Compressive Privatization

Result: ̂p ∈ Δk: an estimate of ̂p

Input: X1, · · · Xn i.i.d. samples from p, privacy parameter ε, sparsity s, matrix A1 ∈ Rn×L, A2 ∈ Rm×L, where K = 2 log (m, k) and L = min (2 log (m, k, ε−1), public random permutation matrix P ∈ Rn×k. (We represent X1, · · · , Xn as one-hot vectors.)

1. Divide the n users into m groups S1, · · · , Sm: Sj := {i ∈ [n] | i ≡ j mod m}.
2. ∀ j ∈ [m] and ∀ i ∈ Sj, pad (K − k) zeroes to the end of Xj and get permuted sample Xj,P = PXj.
3. Define

\[
\hat{Q}_j(X_j^j) = \{ (A_2)_j · X_j^{(1)}, \cdots, (A_2)_j · X_j^{(L)} \} ∈ \{-1, 0, 1\}^L.
\]

The privatized output \(\hat{Q}_j(X_j^j)\) is defined as follows

\[
\hat{Q}_j(X_j^j) = \left\{ \begin{array}{ll}
Q_j(X_j) & \text{if } Q_j \in \Theta(\hat{Q}_j(X_j^j)), \text{ w.p. } \frac{e}{4mL} \varepsilon + \frac{1}{\sqrt{mL}} \varepsilon \cdot \left(1 + \frac{\varepsilon}{\sqrt{mL}}\right)^2 \\
-\hat{Q}_j(X_j^j) & \text{otherwise, w.p. } \frac{1}{\sqrt{mL}} \varepsilon \cdot \left(1 + \frac{\varepsilon}{\sqrt{mL}}\right)^2
\end{array} \right.
\]

where \(\Theta(Q) = \{ \pm e_1, \pm e_2, \cdots, \pm e_L \}\) is the collection of 2L standard basis vectors.

4. For each j′ ∈ [mL] such that j′ ≡ j (mod m) and j′ = j + (t − 1) · m, the server computes

\[
\hat{q}_{j'} = \frac{1}{n} \left( e^{2} + \frac{1}{\varepsilon} - 1 \right) \sum_{i \in S_j} (A_1)_i · \hat{Q}_j(X_j^j).
\]

5. Let \(\hat{q} := (\hat{q}_1, \cdots, \hat{q}_{mL})\). 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 \((L, (1 + \beta)\frac{\varepsilon}{L})\); let \(p'\) be the output. Let \(\hat{p}\) be the first k elements of \(p'\).

6. Compute the projection of \(\hat{p}\) 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 \(X_j\) to denote the permuted one-hot sample vector of user i. By concentration inequalities, \(E\) 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.

\[\text{LEMMA 4.1. Event } E \text{ holds with probability at least } 1 - L e^{-\beta^2 \varepsilon^2} \text{ and when } E \text{ happens, } p' \text{ is } (L, (1 + \beta)\frac{\varepsilon}{L}) \text{-hierarchically sparse.}\]

Privatization. In our privatization step, the response matrix is of the form A = A1 ⊗ A2, where A2 ∈ Rn×L and A1 ∈ Rm×L are two ±1 matrices. User i will send a privatized version of Y1 = AX1. Let \(a_{ij}\) be the (i, j)-entry of A1, then

\[
Y_1 = AX_1 = \begin{bmatrix}
a_{11}A_2 & \cdots & a_{1L}A_2 \\
\vdots & \ddots & \vdots \\
a_{11A_2} & \cdots & a_{1L}A_2 \\
\end{bmatrix}
\begin{bmatrix}
X_1^{(1)} \\
\vdots \\
X_1^{(L)}
\end{bmatrix},
\]

where the one-hot sample vector \(X_j\) is divided into L blocks. Since \(X_j\) has only one non-zero, the vector \(Y_1\) can be encoded with \(m + \log L\) bits, where log L bits is to specified the block index \(t\) that contains the non-zero of \(X_j\) and \(m\) bits is for \(A_2X_j^{(t)}\). However, this is still too large, so user will only pick one bit from \(A_2X_j^{(t)}\), which is the \(j\)-th bit if \(j \in S_t\). Then the user privatize the \(1 + \log L\) bits using RR mechanism [40] with alphabet size \(2L\). More formally, let \(Q_j(X_j) = \{ (A_2)_j · X_j^{(1)}, \cdots, (A_2)_j · X_j^{(L)} \} ∈ \{-1, 0, 1\}^L\), which is also a one-hot vector. Let \(t\) be the block index such that \(X_j^{(t)} ≠ 0\), then the \(t\)-th bit in \(Q_j(X_j)\) is the only non-zero entry, which has value \((A_2)_j · X_j^{(t)}\). Then the user computes a privatization of \(Q_j(X_j)\) (see (14) in Algorithm 3) with \(2^{2\beta}\)-RR. Clearly the communication cost is \(l = \log_2 L + 1\) bits.

Estimation via hierarchical sparse recovery. By Definition 2, \(p'\) is \((L, (1 + \beta)\frac{\varepsilon}{L})\)-hierarchically sparse. To recover \(p'\), by Theorem 1.4 and 1.5, \(\frac{1}{\sqrt{mL}} A_1\) and \(\frac{1}{\sqrt{mL}} A_2\) are required to satisfy 3L-RIP and \(2^{(1+2\beta)}\)-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 \(m = \Theta((1 + \beta) \frac{\log \log (k/(1 + \beta) \varepsilon)}{\varepsilon})\), so that \(\frac{1}{\sqrt{mL}} A_2\) satisfies \(2^{(1+2\beta)}\)-RIP. Let \(q = A_p\), which is equivalent to

\[
\frac{1}{\sqrt{mL}} \hat{q} = \frac{1}{\sqrt{mL}} A_p \hat{q} + \frac{1}{\sqrt{mL}} (q - \hat{q}).
\]

(17)

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

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

(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 \(\hat{q}\).

\[\text{LEMMA 4.2. } \forall j' ∈ [mL], \mathbb{E}\left[ (\hat{q}_{j'} - q_j)^2 \right] ≤ m \left( e^{2} + \frac{1}{\varepsilon} - 1 \right)^2.\]

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

\[\text{THEOREM 4.3. Given } \varepsilon \text{ and a communication budget } b, \text{ with } 1 ≤ e' ≤ s \text{ and let } L = \min(2^b, e'). \text{ For } m = \Theta\left( \frac{(1 + \beta) \frac{\log \log (k/(1 + \beta) \varepsilon)}{\varepsilon}}{L} \right),
\]

with probability at least \(1 - L e^{-\beta^2 \varepsilon^2} - e^{-m}\), our scheme is \(\varepsilon\)-LDP and has communication cost \(\log L + 1\), and the sample complexity for \(l_2\) error is \(O\left( \frac{(1 + \beta) \frac{\log \log (k/(1 + \beta) \varepsilon)}{\varepsilon}}{L} \right)^{\frac{1}{\varepsilon}}\); for \(l_1\) error, the sample complexity is \(O\left( \frac{(1 + \beta) \frac{\log \log (k/(1 + \beta) \varepsilon)}{\varepsilon}}{L} \right)^{\frac{1}{\varepsilon}}\).

The parameter \(\beta\) is from Lemma 4.1. Please refer to the appendix for more discussion.
We simulate We conduct experiments comparing our method with HR [3] and 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, \( k \) for small approximately sparse distributions: 1) geometric distributions \( \text{Geo}(\lambda) \) with \( p(i) \sim (1 - \lambda)^{i-1}; \) 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 \( n \in \{50000, 100000, \ldots, 1000000\} \). 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 \( \ell_1 \) 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 \( \log s + \log \log \frac{1}{\epsilon} \), is also lower than HR.

**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 \( p(i) \sim (1 - \lambda)^{i-1}; \) 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 \( n \in \{50000, 100000, \ldots, 1000000\} \). 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 \( \ell_1 \) 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 \( \log s + \log \log \frac{1}{\epsilon} \), 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. 1951120700.
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
\[
\max_{y \in [m]} Q(y|x_1, x_2) \leq \frac{n_{x_2} e^f + m - n_{x_2}}{n_{x_2} e^f + m - n_{x_1}}.
\]
By assumption,
\[
\frac{n_{x_2} e^f + m - n_{x_2}}{n_{x_2} e^f + m - n_{x_1}} \leq \left(1 + \frac{\beta}{2}\right)\frac{e^f + m - (1 + \beta)\frac{m_f}{m}}{e^f + m - (1 - \beta)\frac{m_f}{m}} = \frac{1 + \beta \cdot \frac{e^f - 1}{e^f}}{1 - \beta \cdot \frac{e^f - 1}{e^f}} \leq 1 + \frac{\beta}{2} \cdot 1 - \frac{\beta}{2} \leq 1 + 2\beta.
\]
where the second inequality is from $\frac{e^f - 1}{e^f} = \frac{1}{2}$ for $e \in (0, 1)$ and the last inequality is from $0 \leq \beta \leq 1$. It follows that
\[
\max_{x_1, x_2, y \in [k]} \frac{Q(y|x_1)}{Q(y|x_2)} \leq \max_{x_1, x_2 \in [k]} \frac{n_{x_2} e^f + m - n_{x_2}}{n_{x_2} e^f + m - n_{x_1}} \leq (1 + 2\beta)e^f = e^{f + \ln(1 + 2\beta)} \leq e^{2\beta}.
\]
The last inequality is from $\ln(1 + x) \leq x$ for $x > -1$. □

A.2 Proof of Lemma 3.2

Proof. By definition
\[
\|e_1\|_2 = \sqrt{\frac{e^f + 1}{m(e^f - 1)}} \cdot J(D' - I)p
\]
\[
\leq \sqrt{\frac{e^f + 1}{m(e^f - 1)}} \cdot \sqrt{J(D' - I)p^T \cdot (J(D' - I)p)}
\]
\[
\leq \frac{e^f + 1}{e^f - 1} \sum_i |d'_i - 1| p_i,
\]
where $d'_i = \frac{m(e^f + 1/2)}{\tilde{n}_{e^f + m - n}}$. By the assumption $n_i \in (1 + \beta)\frac{m_f}{m}$ for all $i \in [k]$, we have
\[
|d'_i - 1| = \frac{m/2 - n_i(m(e^f - 1))}{n_i e^f + m - n_i} \leq \frac{\beta m(e^f - 1)/2}{n_i e^f + m - n_i}
\]
\[
\leq \frac{\beta m(e^f - 1)/2}{m \leq (1 - \beta)(e^f + 1)}
\]
Thus, $\|e_1\|_2 \leq \frac{\beta}{1 - \beta}$, which completes the proof. □

A.3 Proof of Lemma 3.3

Proof. Let $Y_1, Y_2, \ldots, Y_n$ be the privatized samples received by the server. We have, for each $i \in [m]$, $\hat{q}_i = \sum_{j=1}^n i(Y_j = i)/n$, where $\mathbf{1}$ 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 $\sqrt{\frac{m (e^f + 1)}{e^{f - 1}}}$ on both sides of the inequality will conclude the proof. □

A.4 Proof of Theorem 3.4

Proof. Let $p' = D^tp$. By definition of $p'$, we have
\[
\|p - p\|_2 = \|D^{-1} p - D^{-1} f\|_2 \leq \max_i \frac{1}{d_i'} \cdot \|p' - f\|_2
\]
where $d_i' = \frac{m(e^f + 1/2)}{n_{e^f + m - n}}$ and $f$ is the output of the recovery algorithm (step 7 in Algorithm 2). Since $\forall i \in [k], n_i \leq (1 + \beta)\frac{m_f}{m}$ for $0 \leq \beta \leq 1$, then we have
\[
\|p - p\|_2 \leq \max_i \frac{1}{d_i'} \cdot \|p' - f\|_2 \leq (1 + \beta \cdot \frac{e^f - 1}{e^f + 1}) \cdot \|p' - f\|_2
\]
\[
\leq (1 + \frac{\beta}{2}) \cdot \|p' - f\|_2.
\]
By Lemma 1.3, we have
\[
\|p' - f\|_2 \leq \frac{C}{\sqrt{\mathbb{E}[\|p' - [p]_1\|_1 + D \cdot \|e_1 + e_2\|_2]}}
\]
\[
\leq \max_i \frac{1}{d_i'} \cdot \frac{C}{\sqrt{\mathbb{E}[\|p - [p]_1\|_1 D (\|e_1 + e_2\|_2)]]}}.
\]
Note $\max_i \frac{1}{d_i'} \leq (1 + \beta/2)$ as $n_i \leq (1 + \beta)\frac{m_f}{m}$ for all $i$. By Lemma 3.2, 3.3, we get
\[
\mathbb{E}[\|p - p\|_2] \leq \left(1 + \frac{\beta}{2}\right) \left(2D\beta + \frac{D(e^f + 1)}{e^f - 1} \sqrt{\frac{m}{n}}\right)\frac{C}{\sqrt{\mathbb{E}[\|p - [p]_1\|_1 D (\|e_1 + e_2\|_2)]]}}.
\]
which proves the theorem. □

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 $k_1 = \frac{k}{n}$ 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
\[
\mathbb{E}[\|p'_j \neq 0\|] = \left(\frac{k-1}{s_1 - 1}\right) \cdot \frac{q_1}{k} = \frac{q_1}{k}.
\]
Thus, $\mathbb{E}[s_1] = \frac{k}{n}$, $\mathbb{E}[s_1] = \frac{s_1}{k}$. Since $p'$ is a random permutation of $p$, $\mathbf{1}(p'_1 \neq 0), \ldots, \mathbf{1}(p'_k \neq 0)$ are negatively associated (NA) [35]. By Chernoff-Hoeffding bounds for NA variables [18, 35], we can get that
\[
\Pr[s_1 \geq (1 + \beta)\mathbb{E}[s_1]] \leq \frac{e^{\beta}}{(1 + \beta + (1 + \beta)) \ln (1 + \beta)}
\]
\[
= e^{\mathbb{E}[s_1] + (1 + \beta) \ln (1 + \beta)}
\]
(19)
It can be easily verified that
\[
\beta - (1 + \beta) \ln (1 + \beta) \leq \left\{ \begin{array}{ll}
-\frac{\beta^2}{4} & \beta \leq 4, \\
-\frac{\beta}{4} & \text{otherwise}
\end{array} \right.
\]
(20)
Combining (19), (20) and \(\mathbb{E}[s_1] = \frac{L}{\epsilon} \) yields that
\[
\Pr\left[ s_1 \geq (1 + \beta) \frac{\epsilon}{L}\right] \leq e^{-\frac{\min(\beta, \beta) s}{\alpha}}.
\]
(21)
The proof is then completed by applying union bound over all \(L\) sections.

**B.2 Proof of Lemma 4.2**

**Proof.** For any \(j \in [m]\), we have
\[
\mathbb{E}[\hat{Q}_j(X_j')] = \mathbb{E}_p[\mathbb{E}[\hat{Q}_j(x)|X_j' = x]]
\]
When \(X_j'\) is fixed to be \(x \in Q\) where \(Q := \{\pm e_1, \cdots, \pm e_L\}\) and we only consider the randomness from the privatization, we have
\[
\mathbb{E}[\hat{Q}_j(x)] = \frac{\epsilon^j f}{\epsilon^j + 2L - 1} Q_j(x) + \sum_{Q' \in Q, (Q_j(X_j'))} \frac{1}{\epsilon^j + 2L - 1} = \frac{(\epsilon^j + 1)}{\epsilon^j + 2L - 1} \cdot Q_j(x)
\]
(23)
By the definition of \(Q_j\), we can get
\[
\mathbb{E}[Q_j(X_j')] = \begin{pmatrix}
(\alpha_1)_{j} \cdot p^{(1)} \\
(\alpha_2)_{j} \cdot p^{(2)} \\
\vdots \\
(\alpha_{2L})_{j} \cdot p^{(L)}
\end{pmatrix}
\]
(24)
Combining (22), (23) and (24) yields that
\[
\mathbb{E}[\hat{Q}_j(X_j')] = \frac{(\epsilon^j + 1)}{\epsilon^j + 2L - 1} \begin{pmatrix}
(\alpha_1)_{j} \cdot p^{(1)} \\
(\alpha_2)_{j} \cdot p^{(2)} \\
\vdots \\
(\alpha_{2L})_{j} \cdot p^{(L)}
\end{pmatrix}
\]
(25)
Recall that \(q = A \cdot p'\) and \(A = A_1 \otimes A_2\). For \(j' \equiv j \mod m\) and \(j' = j + (t - 1)m\), by (25), we have
\[
\mathbb{E}[\hat{Q}_j'] = \frac{m}{n} \cdot \frac{\epsilon^j + 2L - 1}{\epsilon^j - 1} \sum_{i \in S_j} (A_1)_i \cdot \mathbb{E}[\hat{Q}_j(X_j')]
\]
\[
= (A_1)_t \cdot \begin{pmatrix}
(\alpha_1)_{j'} \cdot p^{(1)} \cdot (\alpha_2)_{j'} \cdot p^{(2)} \cdots (\alpha_{2L})_{j'} \cdot p^{(L)}
\end{pmatrix} = q_{j'},
\]
where the last equality is from the definition of kronecker product. Hence, \(q_{j'}\) is an unbiased estimator for \(q_{j'}\). Thus,
\[
\mathbb{E}[(q_j' - \hat{q}_j')^2] = \text{Var}(q_j') \leq \frac{m^2}{n} \left( \frac{\epsilon^j + 2L - 1}{\epsilon^j - 1} \right)^2 \text{Var} \left((A_1)_t \cdot \hat{Q}_j(X_j')\right)
\]
\[
\leq \frac{m}{n} \left( \frac{\epsilon^j + 2L - 1}{\epsilon^j - 1} \right)^2,
\]
where the inequality is from that \((A_1)_t \cdot \hat{Q}_j(X_j')\) only takes value in \([+1, -1]\). The proof is completed.

**B.3 Proof of Theorem 4.3**

**Proof.** From the analysis of estimation error in section 4, we know that
\[
\mathbb{E}[\|\hat{p} - \hat{p'}\|_2] \leq \mathbb{E}[\|p' - \hat{p'}\|_2] \leq \frac{1}{\sqrt{mL}} \mathbb{E}[\|\hat{q} - q\|_2]
\]
(26)
By Lemma 4.2, we can get
\[
\mathbb{E}[\|\hat{p} - \hat{p'}\|_2] \leq \mathbb{E}[\|p' - \hat{p'}\|_2] \leq \frac{1}{\sqrt{mL}} \mathbb{E}[\|\hat{q} - q\|_2]
\]
\[
\leq \sqrt{\frac{1}{mL} \mathbb{E}[(\hat{q} - q)^2]}\]
\[
\leq \sqrt{\frac{1}{mL} \sum_{j' \in \{m\}} \mathbb{E}[(\hat{q}_j' - q_j')^2]}
\]
\[
\leq \sqrt{\frac{m}{n} \left( \frac{\epsilon^j + 2L - 1}{\epsilon^j - 1} \right) \leq \sqrt{\frac{m}{n} \left( 3\epsilon^j - 1 \right)}}
\]
where (a) is from Jensen’s inequality and (b) is from Lemma 4.2 and (c) is from \(L = \min\{\epsilon^j, 2^k\}\).

(1) \(\epsilon = O(1)\). In this case, we can set \(L = 1\) directly and the communication is 1 bit now. The event \(E\) then holds with probability 1, hence \(m = s \log(k/s)\). Since \(\epsilon = O(1), \frac{\epsilon^j}{e^j} = O(\frac{1}{L})\). Thus \(\mathbb{E}[\|p - \hat{p}\|_2] = O \left( \frac{s \log(k/s)}{mL} \right)\), which is the same error bound as that in one-bit CP for high privacy. Note that \(A_1\) is 1 now, \(A = A_1 \otimes A_2 = A_2 \in \mathbb{R}^{n\times k}\) is a Rademacher matrix. Therefore, for high privacy, if we set \(L = 1\), our scheme is exactly one-bit CP.

(2) \(\epsilon = o(1)\). We mainly consider medium privacy case, where \(\epsilon = o(1)\) and \(\epsilon^j < s\). In this case, \(3\epsilon^j - 1 < 0(1)\). Hence, we have that \(\mathbb{E}[\|p - \hat{p}\|_2] = O(\frac{1}{L})\).

Thus to achieve an \(\epsilon\) error of \(\alpha\), it’s sufficient to get an estimate with \(a' = \frac{\alpha}{2}\sqrt{2s}\) for \(\mathbb{E}[\|p' - \hat{p}'\|_2]\). The sample complexity is \(O \left( \frac{(1 + \beta) s \log(k)}{\epsilon^j t_L} \right)\). Since the event \(E\) holds with probability at least \(1 - L e^{-\frac{\min(\beta, \beta) s}{\alpha}}\) and the RIP condition holds with probability at least \(1 - e^{-mn}\), by union bound, we can achieve the sample complexity above with probability \(1 - L e^{-\frac{\min(\beta, \beta) s}{\alpha}} - e^{-mn}\), where \(m = \frac{(1 + \beta) s \log(k)}{\epsilon^j t_L}\). When \(\min(\beta, \beta) = \frac{4L \log(L/L_D)}{s}\), the error probability from \(E\) is less than \(\frac{1}{T}\). If \(L \log(L/L_D) = O(s)\), then \(\min(\beta, \beta) = O(1)\) which means \(\beta = O(1)\). In this case, with probability 1, \(\min(\beta, \beta) = O(1)\), the sample complexity for \(t_1\) error is \(O \left( \frac{s \log(k/s)}{\epsilon^j t_L} \right)\) and for \(t_1\) error is \(O \left( \frac{s \log(k/s)}{\epsilon^j t_L} \right)\). When \(s \ll L \log(L/L_D)\) and \(L < s\), \(\min(\beta, \beta) = O(\log(s/\delta))\) which means \(\beta = O(\log(s/\delta))\).

For general distribution under medium privacy regime, the sample complexity for \(t_1\) error is at least \(\Omega(\frac{s^2}{\epsilon^j L t_L})\), which implies a lower bound of \(\Omega(\frac{s^2}{\epsilon^j L t_L})\) for \(s\)-sparse distributions. Thus the sample complexity blows up as at most a logarithmic factor.

\[\square\]