Hadamard Response: Estimating Distributions Privately, Efficiently, and with Little Communication

Jayadev Acharya
Cornell University

Ziteng Sun
Cornell University

Huanyu Zhang
Cornell University

Abstract

We study the problem of estimating k-ary distributions under ε-local differential privacy. n samples are distributed across users who send privatized versions of their sample to a central server. All previously known sample optimal algorithms require linear (in k) communication from each user in the high privacy regime ($\varepsilon = O(1)$), and run in time that grows as $n \cdot k$, which is prohibitive for a large domain size k.

We propose *Hadamard Response (HR)*, a local privatization scheme that requires no shared randomness and is symmetric with respect to the users. HR has order optimal sample complexity for all ε, a communication of at most $\log k + 2$ bits per user, and nearly linear running time of $\tilde{O}(n + k)$.

HR is based on Hadamard matrices, and is simple to implement. The statistical performance relies on the coding theoretic aspects of Hadamard matrices, ie, the large Hamming distance between the rows. Computational efficiency is achieved by using the Fast Walsh-Hadamard transform.

We compare our approach with Randomized Response (RR), RAPPOR, and subset-selection mechanisms (SS), both theoretically, and experimentally. For $k = 10000$, our algorithm runs about 100x faster than SS, and RAPPOR.

1 Introduction

Estimating the underlying probability distribution from data samples is a quintessential statistical problem. Given samples from an unknown distribution p, the goal is to obtain an estimate $\hat{p}$ of p. The problem has a rich, and vast literature (see e.g. [6, 39, 17, 18], and many others), with the primary goal of statistical efficiency, namely minimizing the sample complexity for estimation, which is the first resource we consider.

1. Utility. What is the **sample complexity** of estimation?

In many applications, data contains sensitive information, and preserving the privacy of individuals is paramount. Without proper precautions, sensitive information can be inferred as evidenced by well publicized data leaks over the past decade, including de-anonymization of public health records in Massachusetts [41], de-anonymization of Netflix users [37] and de-anonymization of individuals participating in the genome wide association study [29].

Private data release and computation on data has been studied in several fields, including statistics, machine learning, database theory, algorithm design, and cryptography (See e.g., [45, 14, 22, 46, 23, 42, 13]). *Differential Privacy (DP)* [24] has emerged as one of the most popular notions of privacy (see [24, 46, 26, 9, 36, 32], references therein, and the recent book [25]). DP has been adopted by several companies including Google, and Apple [21, 27].

A popular privacy definition is *local differential privacy (LDP)* [45, 23], where users do not trust the data collector, and privatize their data before releasing. We study distribution estimation under LDP. Distribution estimation with privacy is an important problem. For example, understanding the drug usage habits of the entire population (the distribution) is crucial for policy design. Understanding the internet traffic distribution is important for adplacement. In both these applications, preserving individual privacy is essential.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.
The authors are listed in alphabetical order.

2. Privacy. How much information about a user is leaked by the scheme?

There are inherent trade-offs between utility and privacy. Sample privacy trade-offs have been recently studied for various problems, including distribution estimation [23, 31, 47, 43, 20, 35].

However, two crucial resources have not been considered in private distribution estimation, computation, and communication. In applications where the underlying dimensionality is high, or the number of samples is large, it is imperative to have computationally efficient algorithms. Internet companies collect information about user's browsing history over a large number of users and websites, and large departmental stores collect purchase statistics over a large number of users and products. In these problems, algorithms with high computational overhead are prohibitive, even if they have optimal sample complexity. There has been recent interest in computationally efficient distribution estimation in the non-private setting (see e.g., [15, 1, 33, 12, 16, 40, 3]).

3. Computational Complexity. What is the running time of the algorithm?

In distributed applications, communication (both with and without privacy) is critical. For example, a large fraction of internet traffic is on hand-held devices with limited uplink capacity due to limited battery power, limited uplink bandwidth, or expensive data rates. Similarly, in large scale distributed machine learning problems, communication from processors to the server is the bottleneck since local computations are fast. Communication limited distributed distribution estimation has been studied in the non-private setting(e.g., [48, 4, 19, 2, 28]).

In the context of private estimation tasks, the problem of finding the heavy hitters, and learning properties under local differential privacy under the assumption of public randomness, where the server can send communication to the clients to reduce communication from user end has received much attention recently [8, 7, 5, 30, 44, 11]. However, these algorithms require shared randomness, as well as asymmetric schemes, where each user can use a different privatization mechanism. [7] uses a Hadamard transform, but they use it to form orthogonal basis and reduce storage, which is different from us.

4. Communication Complexity. How many bits are communicated?

In this work, we consider discrete distribution estimation under the aforementioned four resources. We provide the first algorithm that is simultaneously sample order optimal for any privacy value, has logarithmic communication per symbol, and runs in linear time in the input and output size.

1.1 Organization.

In Section 2 we describe the problem set-up. In Section 2.1, 2.2 and 2.3, we describe prior privatization schemes, and our results. In Section 3, we provide a family of ε-LDP privatization schemes. Based on these, in Section 4, we specialize and design schemes that are optimal in the most interesting regime of high privacy. Finally in Section 5, we will briefly describe how to extend these schemes to general ε. The full detail will be provided in Section A in the supplementary file.

2 Preliminaries

Local Differential Privacy (LDP). Suppose x is a private information that takes values in a set $\mathcal{X}$ with k elements (wlog let $\mathcal{X} = [k]:=\{0, 1, \ldots, k-1\}$). A privatization mechanism is a randomized mapping Q from $[k]$ to an output set $\mathcal{Z}$ (which can be arbitrary), that maps $x \in \mathcal{X}$ to $z \in \mathcal{Z}$ with probability $Q(z|x)$. The output z of this mapping, called the privatized sample, is then released. Q is ε-locally differentially private (ε-LDP) [23] if for all $x, x' \in \mathcal{X}$,

$$\sup_{z\in\mathcal{Z}} \frac{Q(z|x)}{Q(z|x')} \le e^{\varepsilon}. \tag{1}$$

Small values of $\varepsilon(\varepsilon = O(1))$ are more stringent and are the high privacy regime, and large values of ε are the low privacy regime. When $\mathcal{X}$ and $\mathcal{Z}$ are both discrete, the mechanism Q is described by a stochastic matrix of size $|\mathcal{X}| \times |\mathcal{Z}|$ whose (x, z)th entry is $Q(z|x)$. Q is ε-LDP if the ratio of *any two entries* in a column of this matrix is at most e^{ε}.

Randomness and Symmetry. A scheme that requires shared/public randomness requires the generation of shared randomness at the server, which needs to be communicated to the users. Symmetric schemes are those where each user uses the same privatization scheme [38]. In this paper, we consider schemes that are symmetric and require no shared randomness. Other such schemes include RAPPOR, Randomized Response, and subset selection methods, described later. We note that the literature on heavy hitter estimation has mostly considered schemes with shared randomness [8, 7, 11], and it will be interesting to see if our methods can provide improved algorithms for the heavy hitter problem.

We note that [7] also uses Hadamard matrix during the encoding phase. We emphasize that we use

completely different encoding methods based on different properties of Hadamard matrices. They use Hadamard matrix to sample one bit in the frequency domain while we sample an entire column index (represented by $\log k$ bits) from a row vector of the Hadamard matrix. They use rows of Hadamard matrix to provide orthogonal channels for the users. While any orthogonal channels will work for their problem, they use Hadamard matrix to reduce storage at the users since it is easy to compute without storing it. However, we use Hadamard matrix to define subsets of columns with large symmetric difference, which provides better statistical performance, and use Fast Hadamard Transform in the decoding for improving computational performance. This will become clearer after we describe our scheme fully.

Another advantage of this work is that our method can be generalized to general regimes of ε while schemes for heavy hitter detection [8, 7] only work for $\varepsilon < 1$.

LDP distribution estimation. Let $\Delta_k = \left\{p(0), \ldots, p(k-1) : p(x) \geq 0, \sum_{x=0}^{k-1} p(x) = 1\right\}$ be the set of all distributions over $[k]$. Let $X_1, \ldots, X_n$ be independent samples drawn from an *unknown* $p \in \Delta_k$, where X_i is the private (sensitive) data with the ith user. Each user maps X_i through an ε-LDP Q, to obtain Z_i. The task at the server, upon observing the privatized samples $Z_1, \ldots, Z_n$, is to output $\hat{p} : \mathcal{Z}^n \to \Delta_k$, an estimate of p. Let $d : \Delta_k \times \Delta_k \to \mathbb{R}_+$ be a distance measure between distributions in Δ_k. Private distribution estimation task is the following:

Given $\alpha > 0$, $\varepsilon > 0$, $d : \Delta_k \times \Delta_k \to \mathbb{R}$, design an ε-LDP Q, and a corresponding estimation $\hat{p}$, such that $\forall p \in \Delta_k$, with probability at least 0.9, $d(\hat{p}, p) < \alpha$.

The *sample complexity* is the least n for which such an ε-LDP scheme Q, and a corresponding $\hat{p}$ exists. The *communication complexity* is the number of bits to send Z_i to the server. The *computational complexity* is the total time to estimate $\hat{p}$ from $Z_1, \ldots, Z_n$ at the server and to privatize X_i using Q at each user.

We will use ℓ_1, and ℓ_2 distance in this paper. For $r \geq 0$, the ℓ_r distance between $p, q \in \Delta_k$ is $\ell_r(p, q) := (\sum_x |p(x) - q(x)|^r)^{1/r}$. In non-private setting, the sample complexity of distribution estimation under these distances is known even including precise constants [10, 34].

ε	k-RR	RAPPOR	k -SS	ε-HR
$(0, 1)$	$\frac{k^3}{\varepsilon^2\alpha^2}$	$\frac{k^2}{\varepsilon^2\alpha^2}$	$\frac{k^2}{\varepsilon^2\alpha^2}$	$\frac{k^2}{\varepsilon^2\alpha^2}$
$(1, \log k)$	$\frac{k^3}{e^{2\varepsilon}\alpha^2}$	$\frac{k^2}{e^{\varepsilon/2}\alpha^2}$	$\frac{k^2}{e^{\varepsilon}\alpha^2}$	$\frac{k^2}{e^{\varepsilon}\alpha^2}$
$(\log k, 2\log k)$	$\frac{k}{\alpha^2}$	$\frac{k^2}{e^{\varepsilon/2}\alpha^2}$	$\frac{k}{\alpha^2}$	$\frac{k}{\alpha^2}$
$(2\log k, +\infty)$	$\frac{k}{\alpha^2}$	$\frac{k}{\alpha^2}$	$\frac{k}{\alpha^2}$	$\frac{k}{\alpha^2}$

Table 1: Sample complexity, up to constant factors, under ℓ_1 distance for the different methods. The sample complexity under ℓ_2 distance is exactly a factor k smaller in each cell above.

2.1 The privatization mechanisms

We will now briefly describe RR, RAPPOR, the most popular ε-LDP schemes using no interaction and public randomness. We will also mention SS, and our proposed HR. For a detailed description of RAPPOR and SS, please refer to Section D in the supplementary file.

k-Randomized Response (RR). The k-RR mechanism [45, 31] is an ε-LDP Q_{RR} with $\mathcal{Z} = \mathcal{X} = [k]$, such that

$$Q_{\text{RR}}(z|x) := \begin{cases} \frac{e^\varepsilon}{e^\varepsilon + k - 1} & \text{if } z = x, \\ \frac{1}{e^\varepsilon + k - 1} & \text{otherwise.} \end{cases} \tag{2}$$

k-RAPPOR. The randomized aggregatable privacy-preserving ordinal response (RAPPOR) is an ε-LDP mechanism which was proposed in [23, 27]. Its simplest implementation k-RAPPOR [31] maps $\mathcal{X} = [k]$ to $\mathcal{Z} = \{0, 1\}^k$. It first does a one hot encoding to the input $x \in [k]$ to obtain $\mathbf{y} \in \{0, 1\}^k$, such that $\mathbf{y}_j = 1$ for $j = x$, and $\mathbf{y}_j = 0$ for $j \neq x$. The privatized output of k-RAPPOR is a k-bit vector obtained by independently flipping each bit of $\mathbf{y}$ with probability $\frac{1}{e^{\varepsilon/2}+1}$.

Subset Selection techniques. [43, 47] propose an ε-LDP scheme that maps $x \in [k]$ to subsets of $[k]$ of size $\lceil k/(e^\varepsilon + 1) \rceil$. The scheme is described in detail in Section D.

Hadamard Response. We propose Hadamard Response (HR), an ε-LDP scheme with $\mathcal{Z} = [K]$, for some $k \leq K \leq 4k$. The algorithm is described in Section 4 for high privacy, and in Section 5 and A for general privacy.

2.2 Previous Results

To estimate distributions in Δ_k to ℓ_1 distance α under ε-LDP, the sample, communication and time re-

ε	k-RR	RAPPOR	k -SS	ε-HR
$(0, 1)$	$\log k$	k	k	$\log k$
$(1, \log k)$	$\log k$	$\frac{k}{e^{\varepsilon/2}}$	$\frac{k}{e^{\varepsilon}}$	$\log k$
$(\log k, 2\log k)$	$\log k$	$\frac{k}{e^{\varepsilon/2}}$	$\log k$	$\log k$
$(2\log k, +\infty)$	$\log k$	$\log k$	$\log k$	$\log k$

Table 2: Communication requirements for distribution estimation techniques.

k-RR	k-RAPPOR	Subset selection	ε-HR
$n + k$	$n + k + \frac{nk}{e^{\varepsilon/2}}$	$n + k + \frac{nk}{e^{\varepsilon}}$	$n + k$

Table 3: Time bounds for distribution estimation. The running times are described in Section D. These are upper bounds up to logarithmic factors.

quirements of the various schemes are given in Table 1, 2 and 3 respectively..

The sample complexity is given in **Table 1**. The entries in *green* boxes are sample-order optimal, namely there is a matching lower bound [47]. Note that RR is sample-optimal in the low privacy regime (last two rows), and is *highly sub-optimal* in the high privacy regime ($\varepsilon = O(1)$). RAPPOR is optimal for high-privacy, but sub-optimal for medium privacy. SS, and our proposed HR are sample-order-optimal for all ε. The sample complexity arguments for RR, RAPPOR, and SS can be found in [31, 47].

Table 2 describes the communication requirements of various schemes. However, it is not clear how to measure the communication requirements, since for a given privatization scheme, there might be communication protocols requiring fewer bits than others. For example, RAPPOR is described as giving k bits as its output, but perhaps these k bits can be compressed further requiring much smaller communication. We get around such concerns by observing that, once the input distribution p and the privatization mechanism Q is fixed, the output distribution of the privatized sample Z is fixed. By Shannon's source coding theorem, to *faithfully* send Z to the server requires at least $H(Z)$ bits of communication. The entries in the table are derived by considering the input distribution to be near uniform, and evaluating the entropy of the output of the mechanisms. For RR, $\log k$ bits of communication follows from $\mathcal{Z} = [k]$. Note that in this paper all logarithms are in base 2. The communication requirements for RAPPOR, and SS are derived in Section D (Theorems 9, and Theorem 10 respectively).

Table 3 describes the total running time lower bounds for faithfully implementing the known schemes. The argument is that at the server, the computation complexity is at least the number of bits that need to be read, which is the amount of communication from the users. If there are n users, then $n \cdot H(Z)$ serves as our time complexity bound, and these form the entries in the table.

2.3 Motivation and Our Results

Our work is motivated by the first three columns of the tables, which captures the apparent sample-communication-computation trade-offs present in the existing schemes. We elaborate this point in the most interesting regime of high privacy. For simplicity, fix $\varepsilon = 1$, and $\alpha = 0.1$ (chosen arbitrarily!), and treat them as fixed constants in this paragraph. In this setting, from Table 1, note that the optimal sample complexity is $\Theta(k^2)$, achieved by RAPPOR, and SS, while RR has a sub-optimal sample complexity of $\Theta(k^3)$. Now consider the communication requirements. $\mathcal{Z} = [k]$ for RR, requiring only $\log k$ bits. A straight-forward computation shows that any input distribution to the RAPPOR mechanism induces an output distribution over $\{0, 1\}^k$ with entropy at least $\Omega(k)$, thus requiring $\Omega(k)$ bits to *faithfully* send the privatized samples to the server. SS also requires $\Omega(k)$ bits in this regime. These are formally shown in Theorem 9 and Theorem 10 in Section D. As for the running time at the server end, a bound of $\Omega(k^3)$ for all these three methods follows from the total communication to the server (#samples $\times$ #bits per sample), which is a factor k larger than the $\Theta(k^2)$ optimal sample complexity bound.

Our main result is the following, which is formally stated in Theorem 2, and Theorem 5.

Theorem 1. *We propose a simple algorithm for ε-LDP distribution estimation that for all parameter regimes, is sample optimal, runs in near-linear time in the number of samples, and has only a logarithmic communication complexity in the domain size, for both the ℓ_1, and ℓ_2 distance.*

Going back to the high privacy regime, considered before, this shows that our scheme has a running time of $\tilde{O}(k^2)$, which is nearly linear in the optimal sample complexity under ℓ_1 distance.

3 A family of ε-LDP schemes

We first propose a general family of LDP schemes, and then carefully choose schemes from this family that are sample-optimal, communication and computationally efficient for distribution estimation.

The scheme involves the following steps:

1. Choose an integer K, and let the output alphabet be $\mathcal{Z} = [K]$.
2. Choose a positive integer $s \le K$.
3. For each $x \in \mathcal{X} = [k]$, pick $C_x \subseteq [K]$ with $|C_x| = s$.
4. The privatization scheme from $[k]$ to $[K]$ is then given by:

$$Q(z|x) := \begin{cases} \frac{e^\varepsilon}{se^\varepsilon + K - s} & \text{if } z \in C_x, \\ \frac{1}{se^\varepsilon + K - s} & \text{if } z \in \mathcal{Z} \setminus C_x. \end{cases} \tag{3}$$

This scheme satisfies (1), and is ε-LDP. This privatization scheme chooses a set C_x for each x and assigns the elements in C_x a higher probability than those not in C_x. We also note that RR is a special case of this construction when $K = k$, $s = 1$, and $C_x = \{x\}$. We know from the last section that RR is sub-optimal in the high privacy regime. Our general inspiration comes from coding theory, and we select s, and C_x carefully in order to send more information across Q than RR.

In Section 4 we give an optimal scheme in the high privacy regime, and extend it to the general case in Section 5 and A.

4 Optimal scheme for high privacy

Privatization scheme. If for $x \neq x'$, $C_x = C_{x'}$, then we cannot tell them apart. Therefore, the hope is that the farther apart C_x and $C_{x'}$ are, the easier it is to tell them apart. With this in mind, we specify a particular choice of parameters for our scheme, which turns out to be sample-optimal in the high privacy regime. In particular, our privatization scheme will satisfy the following:

An optimal privatization for high privacy

Choose K, and C_x's such that (We will show in Section 4.1 how to satisfy these conditions.):

C1. K is between k and $2k$, and $s = K/2$, namely for all $x \in [k]$, $|C_x| = \frac{K}{2}$.

C2. For any $x, x' \in [k]$, and $x \neq x'$, $|\Delta(C_x, C_{x'})| = |(C_x \setminus C_{x'}) \cup (C_{x'} \setminus C_x)| = \frac{K}{2}$.

Use (3) for privatization.

Performance. We will show that for $\varepsilon = O(1)$, this privatization is sample-order-optimal, namely there is a corresponding estimator $\hat{p} : [K]^n \to \Delta_k$ that is sample-optimal. Before describing the estimation procedure, we provide the statistical guarantees.

Theorem 2. *For any privatization scheme satisfying **C1**, **C2**, there is a corresponding estimation scheme $\hat{p} : [K]^n \to \Delta_k$, such that*

$$\mathbb{E}\left[\ell_2^2(\hat{p}, p)\right] \le \frac{4k(e^\varepsilon + 1)^2}{n(e^\varepsilon - 1)^2}, \text{ and} \tag{4}$$

$$\mathbb{E}\left[\ell_1(\hat{p}, p)\right] \le \sqrt{\frac{4k^2(e^\varepsilon + 1)^2}{n(e^\varepsilon - 1)^2}}. \tag{5}$$

The sample optimality, and small communication for high privacy is an immediate corollary.

Corollary 3. *When $\varepsilon = O(1)$, the sample complexity of this scheme for estimation to ℓ_1 distance α is $O(k^2/\varepsilon^2\alpha^2)$ samples, and for ℓ_2^2 distance is $O(k/\varepsilon^2\alpha^2)$. Further, the communication from each user is at most $\log(k) + 1$ bits. This is sample-optimal for both ℓ_1 (Table 1) and ℓ_2^2 (see [47]).*

Proof. Applying Markov's inequality in Theorem 2, and substituting $e^\varepsilon + 1 = \Theta(1)$, and $e^\varepsilon - 1 = \Theta(\varepsilon)$ when $\varepsilon = O(1)$ gives the sample complexity bounds. The communication bounds are from $\log K \le \log(k) + 1$. □

Estimation. Suppose $Q_{K,\varepsilon}$ is an ε-LDP scheme satisfying **C1**, and **C2**. For an input distribution p over $[k]$, let $p(C_x)$ be the probability that the privatized sample $Z \in C_x$. Using $|C_x| = K/2$, and **C2**, it follows that $|C_x \setminus C_{x'}| = K/4$, and $|C_x \cap C_{x'}| = K/4$. Therefore,

$$\begin{aligned} p(C_x) &= p(x)\left(\sum_{z \in C_x} Q_{K,\varepsilon}(z|x)\right) + \sum_{x' \neq x} p(x') \cdot \\ &\qquad \left(\sum_{z \in C_x \setminus C_{x'}} Q_{K,\varepsilon}(z|x') + \sum_{z \in C_x \cap C_{x'}} Q_{K,\varepsilon}(z|x')\right) \\ &= p(x) \cdot |C_x| \cdot \frac{e^\varepsilon}{(se^\varepsilon + K - s)} + \\ &\qquad \sum_{x' \neq x} p(x')\left(\frac{|C_x \setminus C_{x'}| \cdot 1}{se^\varepsilon + K - s} + \frac{|C_x \cap C_{x'}| \cdot e^\varepsilon}{se^\varepsilon + K - s}\right) \end{aligned} \tag{6}$$

$$= \frac{1}{2} + \frac{e^\varepsilon - 1}{2(e^\varepsilon + 1)} p(x), \tag{7}$$

where (6) follows from (3), and (7) by plugging $s = K/2$, and from **C2**. We can rewrite this as

$$p(x) = \frac{2(e^\varepsilon + 1)}{e^\varepsilon - 1}\left(p(C_x) - \frac{1}{2}\right). \tag{8}$$

This forms the basis of our estimation. From the privatized samples, we estimate $p(C_x)$, and from that we estimate p. The entire scheme is given below.

An optimal distribution estimation scheme for high privacy

Input: k, ε, privatized samples $Z_1, \ldots, Z_n$

1. For each $x \in [k]$, estimate $p(C_x)$ with its empirical probability:

$$\widehat{p(C_x)} := \sum_{j=1}^{n} \frac{\mathbb{I}\{Z_j \in C_x\}}{n}. \tag{9}$$

2. Estimate $\hat{p}$ as:

$$\hat{p}(x) := \frac{2(e^\varepsilon+1)}{e^\varepsilon-1}\left(\widehat{p(C_x)} - \frac{1}{2}\right). \tag{10}$$

Next we will prove Theorem 2 by showing the performance of this estimation scheme.

Proof of Theorem 2.[1] Let $p(C), \widehat{p(C)}$, be the vector of probabilities of $p(C_x)$'s and $\widehat{p(C_x)}$'s respectively. From (8) and (10),

$$\mathbb{E}\left[\ell_2^2(\hat{p}, p)\right] = \frac{4(e^\varepsilon+1)^2}{(e^\varepsilon-1)^2}\mathbb{E}\left[\ell_2^2(\widehat{p(C)}, p(C))\right].$$

From (9), $\mathbb{E}\left[\widehat{p(C_x)}\right] = \mathbb{E}\left[\mathbb{I}\{Z_j \in C_x\}\right] = p(C_x)$. Therefore,

$$\mathbb{E}\left[\ell_2^2(\widehat{p(C)}, p(C))\right] = \mathbb{E}\left[\sum_{x\in[k]} (\widehat{p(C_x)} - p(C_x))^2\right]$$
$$= \sum_{x\in[k]} \mathbb{E}\left[(\widehat{p(C_x)} - p(C_x))^2\right] = \sum_{x\in[k]} \mathrm{Var}(\widehat{p(C_x)}).$$

By the independence of Z_i's, $\widehat{p(C_x)}$ is the average of n independent Bernoulli random variables each with expectation $p(C_x)$. Hence,

$$\sum_{x\in[k]} \mathrm{Var}(\widehat{p(C_x)}) = \sum_{x\in[k]} \frac{1}{n}\cdot p(C_x)(1-p(C_x))$$
$$\le \frac{1}{n}\sum_{x\in[k]} p(C_x) \le \frac{k}{n}.$$

Plugging this bound in the previous expression gives the bound on ℓ_2^2 distance of the theorem.

$$\mathbb{E}\left[\ell_2^2(\hat{p}, p)\right] \le \frac{4k(e^\varepsilon+1)^2}{n(e^\varepsilon-1)^2}. \tag{11}$$

[1] A technicality here is that $\hat{p}(x)$'s can be negative, but we can project $\hat{p}$ onto the simplex with the same order performance. We therefore only analyze the performance of $\hat{p}$ described in (10).

Using $k\cdot\ell_2^2(\hat{p}, p) \ge \ell_1(\hat{p}, p)^2$ with (11) gives the desired bound on $\mathbb{E}\left[\ell_1(\hat{p}, p)\right]$. □

4.1 Computational complexity and Hadamard matrices.

We showed the sample, and communication complexity guarantees. However, two questions are still unanswered:

- How to choose K, and design C_x's that satisfy **C1, C2**?
- What is the time complexity of privatization and estimation?

We now address these questions. We start with the computational requirements of the proposed scheme, assuming **C1,C2**.

Computation at users. Given C_x's, each user needs to implement (3). This requires uniform sampling from C_x's, as well as from $[K] \setminus C_x$. We will design schemes to do this in time $O(\log K)$.

Computation at the server. The server needs to implement (9) and (10). Note that (10) can be implemented in time $O(k)$ after implementing (9). However, a straightforward implementation of (9) requires $n\cdot k$ time, since for each x we iterate over all the samples, giving running time of $O(n\cdot k)$. In particular, in the high privacy regime (say with $\varepsilon = 1$, and $\alpha = 0.1$) the sample complexity is $O(k^2)$ and the time requirement will be $O(k^3)$. We now show how to design a privatization to satisfy **C1, C2**, and for which we can implement (9) in time only $\tilde{O}(n+k)$.

Hadamard Response (HR) for high privacy. Suppose K is a power of two, and $H_K \in \{\pm 1\}^{K\times K}$ is the Hadamard matrix of size $K\times K$ designed by Sylvester's construction as follows. Let $H_1 = [1]$, and for $m = 2^j$, for $j \ge 1$,

$$H_m := \begin{bmatrix} H_{m/2} & H_{m/2} \\ H_{m/2} & -H_{m/2} \end{bmatrix}.$$

Some standard properties of Hadamard matrices that we use are the following:

(i) The number of $+1$'s in each row except the first is $K/2$,
(ii) Any two rows agree (and disagree) on exactly $K/2$ locations,
(iii) Vector multiplication with H_K is possible in time $O(K\log K)$ with Fast Walsh Hadamard transform,

(iv) We can uniformly sample from the $+1$'s (and the -1's) in any row in time $O(\log K)$.

We now describe the parameters for the privacy mechanism:

1. Choice of K: Let $K = 2^{\lceil \log_2(k+1) \rceil} \geq k+1$, the smallest power of 2 larger than k. To satisfy **C1**, we will choose $s = K/2$.
2. Choice of C_x's: Map the symbols $[k] = \{0, \ldots, k-1\}$ to rows of H_K as follows: map 0 to the second row, 1 to the third row, and so on. In other words, x is mapped to row $x+1$. Given any x, we choose $C_x \subset [K]$ to be the column indices with a '$+1$' in the $(x+1)$th row of H_K.

By Property (i) and (ii) of H_K, both **C1**, and **C2** are satisfied. This implies a privatization scheme with optimal sample and communication complexity in the high privacy regime.

Fast computation with HR. By Property (iv), we can efficiently implement the privatization scheme at the users. We will now provide an efficient implementation of (9). Let $q = (q(0), \ldots, q(K-1))$ be the vector of the empirical distribution of $Z_1, \ldots, Z_n$ over $[K] = \{0, \ldots, K-1\}$, namely

$$q(z) = \sum_{i=1}^{n} \frac{\mathbb{I}\{Z_i = z\}}{n}.$$

We can compute q in linear time with a single pass over $Z_1, \ldots, Z_n$. Consider the matrix vector product $\mathbf{c} = H_K \cdot q$. For $x \in [k]$, the $(x+1)$th entry of $H_K \cdot q$ is $\sum_{z=0}^{K-1} H_K(x+1, z) \cdot q(z)$. Now note that the $+1$'s in the $(x+1)$th column correspond to C_x by construction, therefore

$$\begin{aligned}\sum_{z=0}^{K-1} H_K(x+1,z) \cdot q(z) &= \sum_{z \in C_x} q(z) - \sum_{z \in [K] \setminus C_x} q(z) \\ &= 2\widehat{p(C_x)} - 1 = \left(\frac{e^\varepsilon - 1}{e^\varepsilon + 1}\right)\hat{p}(x),\end{aligned}$$

where the last line follows from observing that $\sum_{z \in C_x} q(z) = \widehat{p(C_x)}$ from (9), and from (10). Therefore the estimator $\hat{p}$ is simply entries of a Hadamard vector product, appropriately normalized. By property (iii), this can be done in time $O(K \log K) = O(k \log k)$. This computational advantage is captured in the following theorem:

Theorem 4. *HR is an ε-LDP mechanism satisfying Theorem 2 that has a running time $\tilde{O}(n+k)$.*

5 General privacy regimes.

For general values of ε, we will still use the general structure of schemes given by (3). However, we will choose the values of s to be dependent on ε (which will be close to k/e^ε for general ε). After fixing this s, we design the sets C_x's with size s by forming block-structured matrices with Hadamard matrices at the diagonals. The general construction, along with the encoding, estimation, and analysis is provided in Section A in the supplementary file. We simply state the main result for general ε here.

The two parameters we need to describe the theorem are B, and b, where b can be thought of as a proxy for s in our scheme. Their precise values are given in Section A, but we only need that $B = \Theta(\min\{e^\varepsilon, 2k\})$, and $b = \Theta(k/B + 1)$ to state the result below.

Theorem 5. *There is an ε-LDP estimate $\hat{p}$ such that*

$$\mathbb{E}\left[\ell_2^2(\hat{p}, p)\right] = O\left(\frac{(k + (e^\varepsilon - 1)b)(B + e^\varepsilon)}{n(e^\varepsilon - 1)^2}\right),$$

$$\mathbb{E}\left[\ell_1(\hat{p}, p)\right] = O\left(\sqrt{\frac{k}{n}\frac{(k + (e^\varepsilon - 1)b)(B + e^\varepsilon)}{2(e^\varepsilon - 1)^2}}\right).$$

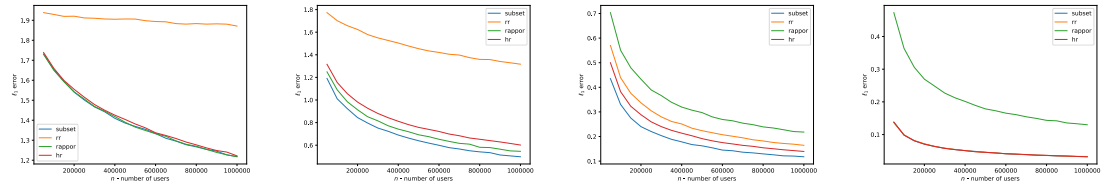
The running time of the algorithm is $\tilde{O}(n+k)$, and communication is at most $2 + \log k$ bits.

Plugging the values of b, and B, and apply Markov's inequality, we can obtain *all the sample complexity* bounds for HR, namely the last column of the sample complexity in Table 1.

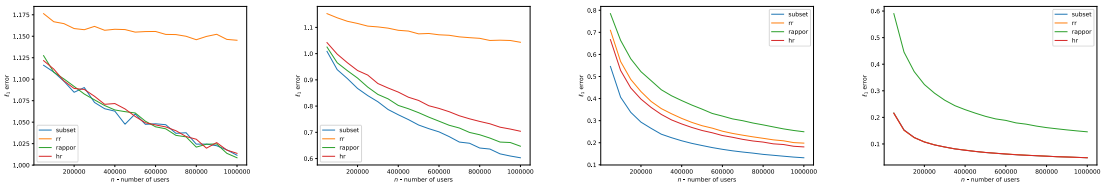
6 Experiments.

We experimentally compare our algorithm with RR, RAPPOR and SS. Our code is available at https://github.com/zitengsun/hadamard_response. We set $k \in \{100, 1000, 5000, 10000\}$, $n \in \{50000, 100000, 150000, ..., 1000000\}$, and $\varepsilon \in \{0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. We consider geometric distributions $Geo(\lambda)$, where $p(i) \propto (1-\lambda)^i \lambda$, Zipf distributions $Zipf(k, t)$ where $p(i) \propto (i+1)^{-t}$, two-step distributions, and uniform distributions. For every setting of (k, p, n, ε), and for each scheme, we simulate 30 runs, and compute the averaged ℓ_1 error, and averaged decoding time at the server.

In a nutshell, we observe that in each regime, the statistical performance of HR is comparable to the best possible. Moreover, the decoding time of HR is similar to that of RR. In comparison to RAPPOR and SS, our running times can be orders of magnitude smaller, particularly for large k, and small ε.

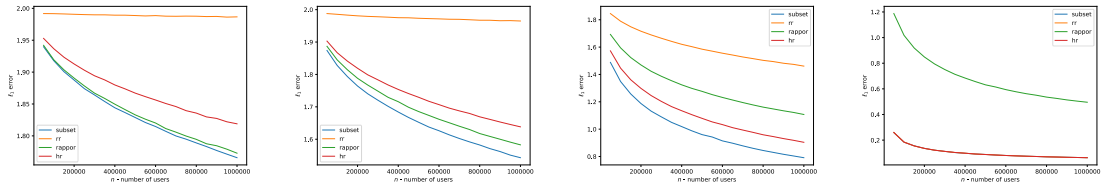


(a) $\varepsilon = 0.5, p \sim Geo(0.8)$ (b) $\varepsilon = 2, p \sim Geo(0.8)$ (c) $\varepsilon = 5, p \sim Geo(0.8)$ (d) $\varepsilon = 7, p \sim Geo(0.8)$



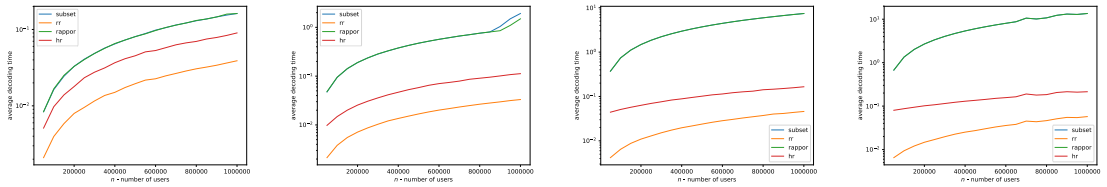
(e) $\varepsilon = 0.5, p \sim U[k]$ (f) $\varepsilon = 2, p \sim U[k]$ (g) $\varepsilon = 5, p \sim U[k]$ (h) $\varepsilon = 7, p \sim U[k]$

Figure 1: ℓ_1-error comparison between four algorithms $k = 1000$, $p \sim Geo(0.8)$ and $p \sim U[k]$



(a) $\varepsilon = 1$ (b) $\varepsilon = 2$ (c) $\varepsilon = 5$ (d) $\varepsilon = 10$

Figure 2: ℓ_1-error comparison between four algorithms $k = 10000$ and $p \sim Geo(0.8)$



(a) $k = 100$ (b) $k = 1000$ (c) $k = 5000$ (d) $k = 10000$

Figure 3: Decoding time comparison between four algorithms for $\varepsilon = 1$ and $p \sim Geo(0.8)$. Note that the decoding times are in logarithmic scale.

We remark that we implement RAPPOR, and SS such that their running time is almost linear in the time needed to read the already compressed communication from the users.

We describe some of our experimental results here. Figure 1 plots the ℓ_1 error for distribution estimation under geometric distribution and uniform distribution for $k = 1000$. Note that for $\varepsilon = 0.5$, and $\varepsilon = 7$, our performance matches with the best schemes. In all the plots SS has the best statistical performance, however that can come at the cost of higher communication, and computation. For larger k such as $k = 10000$, the performance is shown in figure 2. In Figure 1 (d), (h) and Figure 2 (d), you can only see two curves because when ε is high, HR, RAPPOR and SS perform almost the same.

The running time of our algorithm is theoretically a factor $k/\log k$ smaller than RAPPOR and subset selection. This is evident from figure 3, which shows that for large k the running times of RAPPOR and SS are orders of magnitude more than HR, and RR. For example, for $k = 10000$, our algorithm runs 100x faster than SS, and RAPPOR.

Acknowledgements

The authors thank Peter Kairouz, Ananda Theertha Suresh, and Aaron Wagner for many helpful discussions, and technical support during this research.

References

[1] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

[2] J. Acharya, C. L. Canonne, and H. Tyagi. Distributed simulation and distributed inference. *CoRR*, abs/1804.06952, 2018.

[3] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1278–1289, Philadelphia, PA, USA, 2017. SIAM.

[4] M. F. Balcan, A. Blum, S. Fine, and Y. Mansour. Distributed learning, communication complexity and privacy. In *Conference on Learning Theory*, pages 26–1, 2012.

[5] S. Banerjee, N. Hegde, and L. Massoulié. The price of privacy in untrusted recommendation engines. In *Communication, control, and computing (Allerton), 2012 50th annual Allerton conference on*, pages 920–927. IEEE, 2012.

[6] R. Barlow, D. Bartholomew, J. Bremner, and H. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.

[7] R. Bassily, K. Nissim, U. Stemmer, and A. G. Thakurta. Practical locally private heavy hitters. In *Advances in Neural Information Processing Systems*, pages 2285–2293, 2017.

[8] R. Bassily and A. Smith. Local, private, efficient protocols for succinct histograms. In *STOC*, pages 127–135. ACM, 2015.

[9] A. Blum, K. Ligett, and A. Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013.

[10] D. Braess and T. Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

[11] M. Bun, J. Nelson, and U. Stemmer. Heavy hitters and the structure of local privacy. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 435–447. ACM, 2018.

[12] S. O. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, 2014.

[13] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[14] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, 15:429–444, 1977.

[15] S. Dasgupta. Learning mixtures of gaussians. In *Annual Symposium on Foundations of Computer Science (FOCS)*, 1999.

[16] C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *COLT*, 2014.

[17] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, 1985.

[18] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

[19] I. Diakonikolas, E. Grigorescu, J. Li, A. Natarajan, K. Onak, and L. Schmidt. Communication-efficient distributed learning of discrete distributions. In *NIPS*, pages 6394–6404. Curran Associates, Inc., 2017.

[20] I. Diakonikolas, M. Hardt, and L. Schmidt. Differentially private learning of structured discrete distributions. In *NIPS*, pages 2566–2574, 2015.

[21] Differential Privacy Team, Apple. Learning with privacy at scale, December 2017.

[22] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210, New York, NY, USA, 2003. ACM.

[23] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *FOCS*, pages 429–438. IEEE, 2013.

[24] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284. Springer, 2006.

[25] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[26] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *FOCS*, pages 51–60, 2010.

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

[28] Y. Han, P. Mukherjee, A. Özgür, and T. Weissman. Distributed statistical estimation of high-dimensional and nonparametric distributions with communication constraints. In *ISIT*, 2018.

[29] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genetics*, 4(8):1–9, 2008.

[30] J. Hsu, S. Khanna, and A. Roth. Distributed private heavy hitters. In *International Colloquium on Automata, Languages, and Programming*, pages 461–472. Springer, 2012.

[31] P. Kairouz, K. Bonawitz, and D. Ramage. Discrete distribution estimation under local privacy. *arXiv preprint arXiv:1602.07387*, 2016.

[32] P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

[33] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two gaussians. In *STOC*, 2010.

[34] S. Kamath, A. Orlitsky, V. Pichapathi, and A. T. Suresh. On learning distributions from their samples. *In preparation*, 2015.

[35] J. Lei. Differentially private m-estimators. In *Advances in Neural Information Processing Systems*, pages 361–369, 2011.

[36] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103. IEEE, 2007.

[37] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy.IEEE Symposium on*, pages 111–125, 2008.

[38] O. Sheffet. Locally private hypothesis testing. *CoRR*, abs/1802.03441, 2018.

[39] B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

[40] A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *NIPS*, pages 1395–1403. Curran Associates, Inc., 2014.

[41] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[42] M. J. Wainwright, M. I. Jordan, and J. C. Duchi. Privacy aware learning. In *Advances in Neural Information Processing Systems*, pages 1430–1438, 2012.

[43] S. Wang, L. Huang, P. Wang, Y. Nie, H. Xu, W. Yang, X. Li, and C. Qiao. Mutual information optimally local private discrete distribution estimation. *CoRR*, abs/1607.08025, 2016.

[44] T. Wang and J. Blocki. Locally differentially private protocols for frequency estimation. In *Proceedings of the 26th USENIX Security Symposium*, 2017.

[45] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[46] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[47] M. Ye and A. Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *CoRR*, abs/1702.00610, 2017.

[48] Y. Zhang, M. J. Wainwright, and J. C. Duchi. Communication-efficient algorithms for statistical optimization. In *NIPS*, pages 1502–1510. 2012.