Communication Complexity in Locally Private Distribution Estimation and Heavy Hitters

Jayadev Acharya [1] **Ziteng Sun** [1]

Abstract

We consider the problems of distribution estimation, and heavy hitter (frequency) estimation under privacy, and communication constraints. While the constraints have been studied separately, optimal schemes for one are sub-optimal for the other. We propose a sample-optimal ε-locally differentially private (LDP) scheme for distribution estimation, where each user communicates one bit, and requires *no* public randomness. We also show that Hadamard Response, a recently proposed scheme for ε-LDP distribution estimation is also utility-optimal for heavy hitters estimation. Our final result shows that unlike distribution estimation, without public randomness, any utility-optimal heavy hitter estimation algorithm must require $\Omega(\log n)$ bits of communication per user.

1. Introduction

Inferring efficiently from available data forms the core of modern data science. In many applications, being able to infer information from the available data is perhaps the most critical step. However, in many cases, these data samples contain sensitive information about the various users/players, who would like to protect their information from being leaked. For example, medical data may contain sensitive information about individuals' disease and health record, which can be inferred without proper design of the collection scheme, a key issue highly publicized following the publication of (Sweeney, 2002; Homer et al., 2008).

Private data release and computation has been studied in various domains, such as statistics, machine learning, database theory, algorithm design, and cryptography (See e.g., (Warner, 1965; Dalenius, 1977; Dinur & Nissim, 2003; Wasserman & Zhou, 2010; Wainwright et al., 2012; Chaudhuri et al., 2011)). *Differential Privacy (DP)* (Dwork et al., 2006) has emerged as one of the most popular notions of privacy (see (Dwork et al., 2006; Wasserman & Zhou, 2010; Blum et al., 2013; McSherry & Talwar, 2007; Kairouz et al., 2017), references therein, and the recent book (Dwork & Roth, 2014)). DP has been adopted by companies including Google, Apple and Microsoft (Differential Privacy Team, Apple, 2017; Erlingsson et al., 2014; Ding et al., 2017).

A popular privacy definition is *local differential privacy (LDP)*, which was perhaps first proposed in (Warner, 1965), and more recently in (Beimel et al., 2008; Kasiviswanathan et al., 2011), where users do not trust the data collector, and privatize their data before releasing. LDP is a stringent privacy constraint that requires a noise to be added at each sample, and thus provides privacy to all users, even if the data collector is compromised. Often, LDP guarantees come at the cost of increased data requirement for various canonical inference tasks.

Communication, along with privacy has become a valuable resource in many applications. For example, in mobile devices, and small sensors with a limited power/limited uplink capacity, the communication budget can overshadow the local computations performed at each of them. This has led to a growing interest in understanding various inference tasks under limited communication, where the users do not have enough communication to even transmit their data (Braverman et al., 2016; Dagan & Shamir, 2018), and recent works have established optimal bounds, and algorithms for fundamental problems such as distribution estimation (Diakonikolas et al., 2017; Han et al., 2018; Acharya et al., 2018a;b). In particular, communication constraints have been proven to increase the data requirements for distribution estimation and hypothesis testing.

1.1. Notations and Set Up

We consider the following distributed setting. The underlying domain of interest is a known discrete set $\mathcal{X}$ of size k. Without loss of generality, let $\mathcal{X} = [k] := \{1, \ldots, k\}$. There are n users, where user i observes $X_i \in [k]$, and then sends a message $Y_i \in \mathcal{Y}$, the output domain, to the central server (data collector, referee) $\mathcal{R}$, who upon observ-

[*] Equal contribution [1] School of Electrical and Computer Engineering, Cornell University. Correspondence to: Ziteng Sun <zs335@cornell.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

ing the messages $Y^n := Y_1, \ldots, Y_n$ wants to solve some pre-specified inference task. At user i, the process of generating message Y_i from input X_i can be characterized via a channel (a randomized mapping) $W_i : [k] \to \mathcal{Y}$, where $W_i(x, y)$ is the probability that $Y_i = y$ given that $X_i = x$.

We now instantiate LDP, and communication constraints as special cases of this model.

1. Local Differential Privacy. A scheme is ε-Locally Differentially Private (LDP), if $\forall x, x' \in \mathcal{X}$, and $\forall y \in \mathcal{Y}$,

$$\frac{W_i(x, y)}{W_i(x', y)} \le e^{\varepsilon}, \quad \forall i = 1, \ldots, n. \tag{1}$$

2. Communication Constraints. Given a communication budget ℓ. The scheme is *ℓ-bit communication limited* if $\mathcal{Y} = \{0, 1\}^{\ell}$, and therefore, $W_i : \mathcal{X} \to \{0, 1\}^{\ell}$, namely the output messages are at most ℓ bits.

We consider two inference tasks that $\mathcal{R}$ wants to solve.

1. Discrete Distribution Estimation. p is an unknown distribution over the input domain $[k]$, namely $p \in \Delta_k$. Each X_i is an independent draw from p. $\mathcal{R}$ outputs a distribution from a mapping $\hat{p} : \mathcal{Y}^n \to \Delta_k$ to minimize the expected minimax ℓ_1 risk, namely, we want to design the W_i's and $\hat{p}$ to minimize the following optimization problem

$$r(\ell_1, \Delta_k) := \min_{\hat{p}} \min_{W_1, \ldots, W_n} \max_{p \in \Delta_k} \mathbb{E}\left[\|p - \hat{p}\|_1\right]. \tag{2}$$

When $W_1, \ldots, W_n$ satisfy (1), we denote $r(\ell_1, \Delta_k)$ by $r_{\texttt{DP}}(\ell_1, \Delta_k, \varepsilon)$, and when $W_1, \ldots, W_n$ are communication limited by at most ℓ bits, it is denoted by $r_{\texttt{CL}}(\ell_1, \Delta_k, \ell)$

2. Frequency/Heavy Hitter Estimation. Unlike distribution estimation, in this case there is no distributional assumption on the X_i's (i.e., $X^n := X_1, \ldots, X_n$ can be any element in $[k]^n$), and the goal is to estimate the empirical distribution of the symbols. In particular, for $x \in [k]$, let N_x be the number of appearances of x in X^n. The objective is to estimate the N_x/n's from the messages under LDP constraints. This problem has been studied under the ℓ_∞ norm objective, namely the goal is to design $W^n := W_1, \ldots, W_n$, and $\hat{p}$ to minimize

$$r_{\texttt{DP}}(\ell_\infty, k, \varepsilon) := \min_{\hat{p}} \min_{W^n} \max_{X^n} \mathbb{E}\left[\max_x \left|\hat{p}(x) - \frac{N_x}{n}\right|\right], \tag{3}$$

where the expectation is over the randomness over messages induced by the channels, and the estimator $\hat{p}$.

We will consider simultaneous message passing (SMP) communication protocols, where each user sends their message at the same time. And within these, we study both protocols that have access to public randomness, versus those that do not.

1. Private-coin Schemes: In private-coin schemes, the players choose their channels W_i's independently, without coordination between each other. Formally, $U_1, \ldots, U_n$ are n mutually independent random variables distributed across users $1, \ldots, n$ respectively. User i chooses W_i as a function of U_i. $\mathcal{R}$ knows the distribution of U_i, but not the instantiation of U_i used to choose W_i. We now show that for ε-LDP, and/or communication-limited schemes, private-coin schemes can be assumed to be deterministic, namely the channels $W_1, \ldots, W_n$ are all fixed and known to $\mathcal{R}$.

Lemma 1. *Private-coin schemes are equivalent to deterministic schemes under LDP, and communication-limited constraints.*

Proof. For any user i, the expected channel is given by $\mathbb{E}[W_i]$, where the expectation is over the randomness in U_i. Now the channels satisfying (1), and those with output at most ℓ bits both form a convex set, and therefore $\mathbb{E}[W_i]$ also satisfies the condition, and we can use a deterministic scheme with the channel given by $\mathbb{E}[W_i]$. □

2. Public-coin Schemes: In public-coin protocols, the users and referee all have access to a common random variable U. The users select their channels as a function of U, namely $W_i = f_i(U)$. $\mathcal{R}$ solves the inference task using messages Y^n, and U.

3. Symmetric Schemes: These are schemes, where each uses the same privatization scheme. In particular, for private-coin symmetric schemes, W_i's are the same for all users, denoted by W.

We note that private-coin schemes are easier to implement than public-coin schemes, since they do not require additional communication from the server specifying U. Even within private-coin schemes, symmetric schemes are easier to implement, since all users perform the same operation.

While we consider SMP, there are more general protocols called interactive/adaptive schemes which we do not consider. These operate in rounds, and in each round some users send their messages. The other players can choose their channels upon observing these messages (Duchi et al., 2013b; Qin et al., 2016; Duchi et al., 2017). The role of interaction in the local model has been studied in (Kasiviswanathan et al., 2011; Smith et al., 2017; Duchi et al., 2017) and a separation has been proved between the power of interactive and non-interactive schemes. However, interactive schemes are particularly prone to delays, and require more sophisticated implementation (Smith et al., 2017).

Our goal is to analyze the trade-offs between utility and communication and performance of various privatization schemes for the two problems of LDP distribution estimation, and heavy hitter estimation. Throughout the paper, we consider the high privacy regime, where $\varepsilon = O(1)$.

1.2. Prior Work

Distribution estimation is a classical task, and in the centralized setting, where $\mathcal{R}$ observes the true samples X^n (Devroye & Lugosi, 2001),

$$r(\ell_1, \Delta_k) = \Theta\left(\sqrt{\frac{k}{n}}\right). \tag{4}$$

Distribution estimation under ε-LDP has been well studied in the past few years (Duchi et al., 2013a; Erlingsson et al., 2014; Wang et al., 2016; Kairouz et al., 2016; Ye & Barg, 2018; Acharya et al., 2019; Bassily, 2019). (Duchi et al., 2013a; Kairouz et al., 2016; Ye & Barg, 2018; Acharya et al., 2019) have given private-coin, symmetric SMP schemes which for $\varepsilon = O(1)$ (our regime of interest) achieve

$$r_{\texttt{DP}}(\ell_1, \Delta_k, \varepsilon) = \Theta\left(\sqrt{\frac{k^2}{n\varepsilon^2}}\right). \tag{5}$$

This ℓ_1 risk is optimal over all protocols, even while allowing public-coins, or even interactive schemes (Duchi et al., 2013b; Ye & Barg, 2018; Acharya et al., 2018b). Note that compared to the centralized setting, the risk is a factor of $\sqrt{k}/\varepsilon$ higher (up to constant factors), which shows the significant drop in the performance under LDP.

In terms of communication requirements (number of bits to describe Y_i's) for the sample-optimal private-coin schemes, (Duchi et al., 2013a; Ye & Barg, 2018) require $\Omega(k)$ bits per user, and Hadamard Response (HR) of (Acharya et al., 2019) requires $\log k + 2$ bits. All these schemes are private-coin. (Bassily & Smith, 2015) showed the following remarkable result, which can help understand some of our contributions better. We rephrase their results, and the precise statement can be found their paper.

Lemma 2 ((Bassily & Smith, 2015)). *Any private-coin scheme with arbitrary communication requirements can be converted into a public-coin scheme that requires only one bit of communication from each user with almost no loss in performance.*

Therefore, with public randomness, we can obtain schemes that require one bit of communication from each user, and are sample-optimal. In this paper, we study the problem of whether public randomness is necessary to reduce communication. The answer turns out to be different for distribution estimation and heavy hitter detection. While we can reduce the communication to one bit without public randomness for distribution estimation, this is not possible for heavy hitter detection.

Distribution estimation has also been studied recently under very low communication budget (Diakonikolas et al., 2017; Han et al., 2018; Acharya et al., 2018a) , where each user sends only $\ell < \log k$ bits to $\mathcal{R}$. In particular, now it is established that by only using private coin protocols,

$$r_{\texttt{CL}}(\ell_1, \Delta_k, \ell) = \Theta\left(\sqrt{\frac{k^2}{n\min\{2^\ell, k\}}}\right). \tag{6}$$

Further, these results are tight even with public coins. Note that for $\ell = 1$, when each user can send only one bit,

$$r_{\texttt{CL}}(\ell_1, \Delta_k, 1) = \Theta\left(\sqrt{\frac{k^2}{n}}\right). \tag{7}$$

(5), and (7) show the parallel between LDP, and communication constraints for ℓ_1 risk of distribution estimation.

The problem of frequency/heavy hitter detection under ε-LDP has also received great attention (Hsu et al., 2012; Erlingsson et al., 2014; Bassily & Smith, 2015; Qin et al., 2016; Bassily et al., 2017; Ding et al., 2017; Bun et al., 2018). The state of the art techniques (Bassily & Smith, 2015; Bassily et al., 2017; Bun et al., 2018) require public randomness to guarantee privacy and reduce communication. The optimal ℓ_∞ risk was established in (Bassily & Smith, 2015) as

$$r_{\texttt{DP}}(\ell_\infty, k, \varepsilon) := \Theta\left(\frac{1}{\varepsilon}\sqrt{\frac{\log k}{n}}\right). \tag{8}$$

The focus of the recent works has been to study the computation, and communication requirements for this problem. They all propose algorithms that require $O(1)$ communication per user by using public randomness as mentioned above. Perhaps more intrestingly, they are also able to obtain algorithms whose running time at $\mathcal{R}$ is linear in n, and logarithmic in k. However, all these schemes require public-coin protocols, which implies communication from the server to the users. We also note that for example in (Bassily et al., 2017) the randomness can actually be simulated at each player who can then transmit it, causing an increased communication cost.

1.3. Our Results and Techniques

Recall that for distribution estimation under LDP constraints, all known private-coin schemes require $\Omega(\log k)$ communication bits per message, and are symmetric.

Our paper was motivated by the following question:

Does there exist a private-coin ε-LDP distribution estimation scheme with only one bit of communication per user?

Now consider the special case of $\varepsilon = 1$. If such a scheme exists, then from (5), and (7), it would be optimal simultaneously under both LDP, and communication constraints. Further recall that private-coin protocols are easier to implement, and do not require additional communication from

Communication Randomness	$O(1)$ bits	$O(\log k)$ bits
Symmetric, Private Randomness	$\Omega(1)$ (Theorem 6)	$\Theta\left(\sqrt{\frac{k^2}{n\varepsilon^2}}\right)$ (Acharya et al., 2019)
Private Randomness	$\Theta(\sqrt{\frac{k^2}{n\varepsilon^2}})$ (Corollary 1)	$\Theta\left(\sqrt{\frac{k^2}{n\varepsilon^2}}\right)$
Public Randomness	$\Theta\left(\sqrt{\frac{k^2}{n\varepsilon^2}}\right)$	$\Theta\left(\sqrt{\frac{k^2}{n\varepsilon^2}}\right)$

Table 1. ℓ_1 risk for distribution learning under different communication budget and available randomness.

Communication Randomness	$O(1)$ bits	$O(\log k)$ bits
Symmetric, Private Randomness	$\Omega(1)$	$\Theta\left(\sqrt{\frac{\log k}{n\varepsilon^2}}\right)$ (Theorem 7)
Private Randomness	$\omega(1)$ (Theorem 8)	$\Theta\left(\sqrt{\frac{\log k}{n\varepsilon^2}}\right)$
Public Randomness	$\Theta\left(\sqrt{\frac{\log k}{n\varepsilon^2}}\right)$ (Bassily & Smith, 2015)	$\Theta\left(\sqrt{\frac{\log k}{n\varepsilon^2}}\right)$

Table 2. ℓ_∞ risk for frequency estimation under different communication budget and available randomness.

$\mathcal{R}$ to specify the public randomness. Our first result shows that such a scheme exists. (See Theorem 5 for a precise statement).

Theorem 1. *There is a private-coin ε-LDP distribution estimation scheme, with optimal ℓ_1 risk and requiring one bit of communication per user.*

Our result builds on the Hadamard Response (HR) mechanism, and instead of sending information about rows of the matrix as done in (Acharya et al., 2019), we send only binary information about the columns. Additionaly, by utilizing the distributional assumption under the samples, we assign different users to send information about different columns. The scheme, and analysis is given in Section 2.

The scheme achieving Theorem 1 is asymmetric, and we next show any symmetric private scheme with $\ell < \log k$ bits of communication per user cannot achieve a non-trivial ℓ_1 risk (See Theorem 6 for a precise statement).

Theorem 2. *Let $\ell < \log k$. For any private-coin symmetric scheme that sends ℓ bits per user, there exists a distribution $p \in \Delta_k$ such that $\mathbb{E}[\|\hat{p} - p\|_1] \geq 1$.*

This result implies among all symmetric private-coin schemes, HR has optimal communication (up to two bits) of $\log k + 2$ bits.

We then consider the heavy hitter estimation, for which all known optimal algorithms use public randomness. We show that HR scheme, which is a symmetric scheme with no public randomness and only $\log k + 2$ bits of communication per user, has the optimal ℓ_∞ risk for heavy hitter estimation. (See Theorem 7 for a precise statement)

Theorem 3. *HR has optimal ℓ_∞ risk for heavy hitter estimation.*

However, we remark that the computation requirements of HR is $O(k \log k + n)$, which can be much worse than the guarantees in (Bassily et al., 2017) for $k \gg n$.

Finally, we consider the communication requirements for heavy hitter estimation. Even though the problems of distribution and heavy hitter estimation are similar, unlike distribution estimation for which there is an optimal private-coin ε-LDP protocol with one bit communication, we prove that any optimal private-coin scheme (even asymmetric) must send $\Omega(\log n)$ bits to solve the frequency estimation problem (See Theorem 8 for a precise statement).

Theorem 4. *Suppose $\ell = o(\log n)$. There is no private-coin heavy hitter estimation scheme with optimal performance that communicates ℓ bits.*

For a complete summary of results, see Table 1 and 2. In each table, the problem becomes easier as we go down in rows and go right in columns.

1.4. Organization

We first give our private-coin distribution estimation scheme with one bit communication in Section 2. In Section 3, we show that any symmetric private-coin scheme must transmit $\log k$ bits per user. In Section 4, we prove the optimality of Hadamard Response and finally we show that without public randomness, heavy hitter estimation is impossible with communication complexity $o(\log n)$.

2. Private-coin LDP Distribution Estimation with One Bit Communication

We propose a deterministic scheme, namely the W_i's are fixed apriori, for LDP distribution estimation that has the optimal ℓ_1 risk, and where the output of each W_i is binary, i.e., one bit of communication per user. The approach is the following. Each user is assigned to a deterministic set $B \subset [k]$. Upon observing a sample $X \sim p$, they output $Y \in \{0, 1\}$, according to the following distribution

$$\Pr(Y = 1) = \begin{cases} \frac{e^\varepsilon}{e^\varepsilon+1}, \text{if } X \in B, \\ \frac{1}{e^\varepsilon+1}, \text{otherwise.} \end{cases} \tag{9}$$

In other words, each user sends the indicator of whether their input belongs to a particular subset of the domain. The choice of the subsets is inspired by the Hadamard Response (HR) scheme described in (Acharya et al., 2019). A brief introduction of HR can be found in Section 4 where we show that HR is utility-optimal for heavy hitter estimation.

Recall Sylvester's construction of Hadamard matrices:

Definition 1. Let $H_1 \overset{\text{def}}{=} [1]$, and for $m = 2^j$, for $j \geq 1$,

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

Let $K = 2^{\lceil \log_2(k+1) \rceil}$ be the smallest power of 2 larger than k. Let H_K be the $K \times K$ Hadamard matrix. For simplicity of working with H_K, we assume that the underlying distribution is over $[K]$ by appending p with zeros, giving $p_K = (p(1), \ldots, p(k), 0, \ldots, 0)$. For $y = 1, \ldots, K$, let B_i be the set of all $x \in [K]$, such that $H_K(x, y) = 1$, namely the row indices that have '1' in the ith column. We associate the subsets for each user as follows. We deterministically divide the n users numbered $1, \ldots, n$ into K subsets $S_1, S_2, \ldots, S_K$, such that

$$S_y := \{j \in [n] | j \equiv y \pmod{K}\}. \tag{10}$$

For each user j, let $y_j \in [K]$ be the such that $j \in S_{y_j}$. The jth user then sends its binary output Y_i according to the distribution in (9), with $B = B_{y_j}$, and $X = X_j$.

For any $y = 1, \ldots, K$, the users in S_y have the same output distribution. Let s_y be the probability $Y_j = 1$ for $j \in S_y$. Let $p(B_y) = \Pr(X \in B_y | X \sim p)$. Note that

$$\begin{aligned} s_y &= p(B_y) \cdot \frac{e^\varepsilon}{e^\varepsilon + 1} + (1 - p(B_y)) \frac{1}{e^\varepsilon + 1} \\ &= \frac{1}{e^\varepsilon + 1} + p(B_y) \cdot \frac{e^\varepsilon - 1}{e^\varepsilon + 1}. \end{aligned} \tag{11}$$

Let $p_B := (p(B_1), p(B_2), \ldots, p(B_K))$. Then we obtain

$$\mathbf{s} := (s_1, \ldots, s_K) = \frac{1}{e^\varepsilon + 1} \mathbf{1}_K + \frac{e^\varepsilon - 1}{e^\varepsilon + 1} p_B. \tag{12}$$

This relates $p(B_y)$ with s_y, and now we relate $p(x)$ with $p(B_y)$'s. Recall that $B_1 = [K]$, the entire set. Since B_y's are defined by the rows of Hadamard matrix, we obtain the following (Acharya et al., 2019),

$$p_B = \frac{H_K \cdot p_K + 1_K}{2}. \tag{13}$$

We can now relate the results and describe our estimate.

1. Use an empirical estimate $\widehat{\mathbf{s}}$ for $\mathbf{s}$ as
$$\widehat{s_y} := \frac{1}{|S_y|} \sum_{j \in S_y} Y_j. \tag{14}$$
2. Motivated by (11) estimate p_B as
$$\widehat{p_B} = \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \left(\widehat{\mathbf{s}} - \frac{\mathbf{1}_K}{e^\varepsilon + 1} \right). \tag{15}$$
3. Estimate for the original distribution using (13) as
$$\widehat{p_K} := H_K^{-1} \cdot (2\widehat{p_B} - 1_K) = \frac{1}{K} H_K \cdot (2\widehat{p_B} - 1_K). \tag{16}$$
4. Output $\widehat{p}$, the projection of the first k coordinates of the K dimensional $\widehat{p_K}$ on the simplex $\triangle_k$.

Theorem 5. *Let $\widehat{p}$ be the output of the scheme above when the underlying distribution is p. Then,*

$$\mathbb{E}\left[\|\widehat{p} - p\|_2^2\right] \leq \min\left\{ \frac{2k(e^\varepsilon + 1)^2}{n(e^\varepsilon - 1)^2}, 8\sqrt{\frac{(e^\varepsilon + 1)^2 \log k}{n(e^\varepsilon - 1)^2}} \right\}.$$

Proof. First note that $\widehat{\mathbf{s}}$ is an unbiased estimator of $\mathbf{s}$, (13), (11) and (14), are all linear. Therefore, $\widehat{p_K}$ is an unbiased estimator of p_K. Hence,

$$\mathbb{E}\left[\|\widehat{p_K}(1:k) - p_K(1:k)\|_2^2\right] = \sum_{x=1}^{k} \mathrm{Var}\left(\widehat{p_K(x)}\right).$$

From (16), $\widehat{p_K(x)}$ is a weighted sum of $\{(2\widehat{p_B(y)} - 1)\}_{y=1}^K$ with coefficients either $+\frac{1}{K}$ or $-\frac{1}{K}$. Hence $\forall x \in [K]$,

$$\begin{aligned} \mathrm{Var}\left(\widehat{p_K(x)}\right) &\leq \frac{4}{K^2} \sum_{y=1}^{K} \mathrm{Var}\left(\widehat{p_B(y)}\right) \\ &= \frac{4}{K^2} \left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1} \right)^2 \sum_{y=1}^{K} \mathrm{Var}\left(\widehat{s_y}\right). \end{aligned}$$

By (14), $\widehat{s_y}$ is an average of $|S_y|$ independent Bernoulli random variables. From (10), we also have that $|S_y| \geq \lfloor \frac{n}{k} \rfloor \geq \frac{n}{2K}$, whenever $n > K$. Hence $\forall y \in [K]$,

$$\mathrm{Var}\left(\widehat{s_y}\right) = \frac{1}{|S_y|^2} \sum_{j \in S_y} \mathrm{Var}\left(Y_j\right) \leq \frac{K}{2n}.$$

Combining these, we get: $\forall x \in [K]$,

$$\mathrm{Var}\left(\widehat{p_K(x)}\right) = \frac{4}{K^2}\left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)^2 \sum_{y=1}^{K} \mathrm{Var}\left(\widehat{s_y}\right) \le \frac{2(e^\varepsilon+1)^2}{n(e^\varepsilon-1)^2}.$$

Then the final estimate $\widehat{p}$ is the projection of p_K on the first k coordinates onto the simplex $\triangle_k$. Since $\triangle_k$ is convex,

$$\begin{aligned}\mathbb{E}\left[\|\widehat{p}-p\|_2^2\right] &\le \mathbb{E}\left[\|\widehat{p_K}(1:k)-p_K(1:k)\|_2^2\right] \\ &\le \sum_{i=1}^{k}\left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)^2 \frac{2}{n} \le \frac{2k(e^\varepsilon+1)^2}{n(e^\varepsilon-1)^2}.\end{aligned}$$

Moreover, we have the following lemma:

Lemma 3. *(Corollary 2.3 (Bassily, 2019)) Let $L \subset R^d$ be a symmetric convex body of k vertices $\{\mathbf{a_j}\}_{j=1}^k$, and let $\mathbf{y} \in L$ and $\bar{\mathbf{y}} = \mathbf{y} + \mathbf{z}$ for some $\mathbf{z} \in R^d$. Let $\hat{\mathbf{y}} = \arg\min_{\mathbf{w}\in L}\|\mathbf{w}-\bar{\mathbf{y}}\|_2^2$. Then, we must have:*

$$\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2 \le 4\max_{j\in[k]}\{\langle \mathbf{z}, \mathbf{a_j}\rangle\} \tag{17}$$

Notice that according to (15), (14) and (11), $\{\widehat{p_B}(y) - p_B(y)\}_{y=1}^K$ are empirical averages of independent zero mean Bernoulli random variables scaled by constant $\frac{e^\varepsilon+1}{e^\varepsilon-1}$ and they are mutually independent. Hence, they are sub-Gaussian with variance proxy $\frac{K}{2n}\left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)^2$.

Additionally, by (11) and (16), we know each of $\{\widehat{p_K(x)} - p_K(x)\}$ is a linear combination of $\{\widehat{p_B}(y) - p_B(y)\}_{y=1}^K$ with coefficient either $+\frac{2}{K}$ or $-\frac{2}{K}$. Hence $\{\widehat{p_K(x)} - p_K(x)\}$'s are also sub-Gaussian with variance proxy $\frac{2}{n}(\frac{e^\varepsilon+1}{e^\varepsilon-1})^2$ (see Corollary 1.7 (Rigollet, 2015)).

Hence using Lemma 3, we have:

$$\begin{aligned}\mathbb{E}\left[\|\widehat{p}-p\|_2^2\right] &\le 4\mathbb{E}\left[\max_{x=1}^{k}|\widehat{p_K(x)} - p_K(x)|\right] \\ &\le 8\sqrt{\frac{(e^\varepsilon+1)^2\log k}{n(e^\varepsilon-1)^2}}.\end{aligned}$$

The last step is due to a well-known bound on expectation of maximum of sub-Gaussian random variables (see Theorem 1.16 (Rigollet, 2015)). □

Corollary 1. *Let $\widehat{p}$ be the distribution estimated by the scheme described above. Then for any input p,*

$$\mathbb{E}\left[\|\widehat{p}-p\|_1\right] \le \sqrt{\frac{2k^2(e^\varepsilon+1)^2}{n(e^\varepsilon-1)^2}}.$$

Proof. By Cauchy-Schwarz inequality, $\mathbb{E}\left[\|\widehat{p}-p\|_1\right] \le \sqrt{k\mathbb{E}\left[\|\widehat{p}-p\|_2^2\right]}$. Plugging in Theorem 5 gives the bound. □

Notice here that $e^\varepsilon - 1 = O(\varepsilon)$ when $\varepsilon = O(1)$. Hence we have $\mathbb{E}\left[\|\widehat{p}-p\|_1\right] = O(\sqrt{\frac{k^2}{n\varepsilon^2}})$, which is order optimal.

3. Lower Bound on Communication Complexity of Symmetric Schemes

We show that any private-coin symmetric distribution estimation scheme must communicate at least $\log k$ bits.

Theorem 6. *For any private-coin scheme without shared randomness that transmits $\ell < \log k$ bits per user, there exists a distribution $p_0 \in \triangle_k$ such that for $X^n \sim p_0$,*

$$\mathbb{E}\left[\|\hat{p}(Y^n) - p_0\|_1\right] \ge 1, \tag{18}$$

where Y^n are the messages sent to $\mathcal{R}$ after privatizing X^n.

Proof. Assume that $\mathcal{Y} = [2^\ell]$ is the output alphabet. By Lemma 1, and symmetry, let W be an ℓ-bit communication channel used by each user. We can describe W as a transition probability matrix (TPM) $W \in \mathbb{R}^{k\times 2^\ell}$:

$$W(x,y) := \Pr\left(Y = y | X = x\right).$$

When the input distribution is p, the distribution of the output message is $q = W^T p$. Notice that W^T is an $2^\ell \times k$ matrix, which is underdetermined since $2^\ell < k$. Therefore, there exists a non-zero vector $\mathbf{e}$ such that $W^T\mathbf{e} = 0$. Further, since W is a TPM, each row of W sums to one, and therefore $W^T\mathbf{e} = 0$ implies that $\sum_{x=1}^k \mathbf{e}(x) = 0$.

By scaling appropriately, we can ensure that $\|\mathbf{e}\|_1 = 2$, which ensures that the positive entries sum to one, and negative entries sum to -1. Now consider the distributions specified by these entries, namely let $p_1 = \max\{\mathbf{e}, 0\}$ and $p_2 = \max\{-\mathbf{e}, 0\}$. Then these two distributions have *disjoint support*, however,

$$W^T p_1 = q = W^T p_2,$$

showing that their output message distributions are identical and they cannot be distinguished. Since $\|p_1 - p_2\|_1 = 2$, when we get $Y^n \sim q^n$, for at least one of these distributions, the expected ℓ_1 error is 1, proving the result. □

Note that Theorem 6 holds for all symmetric schemes, not just ε-LDP schemes, which means the result also extends to non-private setting, proving the importance of asymmetry in communication efficient distribution estimation. Further, with just two more bits, using $\log k + 2$ bits, HR is private-coin, symmetric, and does optimal distribution estimation.

4. Hadamard Response is Optimal for Heavy Hitter Estimation

We first describe the scheme briefly, and prove the optimality. We refer the reader to (Acharya et al., 2019) for details.

Recall that K is the smallest power of 2 larger than k. Let H_K be the $K \times K$ Hadamard matrix. The output alphabet of the messages is $\mathcal{Y} := [K]$ For each input symbol $x \in \{1, \ldots, k\}$, let C_x be the symbols $y \in [K]$ such that there is a 1 in the yth column of the $(x+1)$th row of H_K. The reason we start with the second row is because the first row of H_K is all one's. Since H_K is Hadamard,

1. $\forall x \in [k], |C_x| = \frac{K}{2}$, and
2. $\forall x \neq x' \in [k], |C_x \cap C_{x'}| = \frac{K}{4}$.

HR is the following symmetric privatization scheme for all user with output $y \in [K]$, $x \in [k]$,

$$W(x,y) = \begin{cases} \frac{2e^\varepsilon}{K(1+e^\varepsilon)} & \text{if } y \in C_x, \\ \frac{2}{K(1+e^\varepsilon)} & \text{otherwise.} \end{cases} \tag{19}$$

Consider an arbitrary input X^n, with N_x being the number of appearances of x's in X^n. Let $N_{C_x} := \sum_{i \in [n]} \mathbf{1}\{Y_i \in C_x\}$ be the number of output symbols that are in C_x. Then, we have

$$\begin{aligned} \mathbb{E}\left[N_{C_x}\right] &= \sum_{i \in [n]} \mathbb{E}\left[\mathbf{1}\{Y_i \in C_x\}\right] = \sum_{i \in [n]} \Pr\left(Y_i \in C_x\right) \\ &= \sum_{i \in [n]} \left(\mathbf{1}\{X_i = x\} \frac{e^\varepsilon}{1+e^\varepsilon} + \mathbf{1}\{X_i \neq x\} \frac{1}{2} \right) \\ &= \frac{e^\varepsilon - 1}{2(e^\varepsilon + 1)} N_x + \frac{n}{2}. \end{aligned} \tag{20}$$

Hence,

$$\hat{p}(x) = \frac{2(e^\varepsilon+1)}{n(e^\varepsilon-1)} \Big(N_{C_x} - \frac{n}{2} \Big). \tag{21}$$

is an unbiased estimator for $\frac{N_x}{n}$. The performance of the estimator is stated in Theorem 7.

Theorem 7. *For any dataset X^n, the encoding scheme in (19) combined with the estimation scheme in (21) satisfies that:*

$$\mathbb{E}\left[\max_{x \in [k]} \left| \hat{p}(x) - \frac{N_x}{n} \right| \right] \leq \frac{4(e^\varepsilon+1)}{e^\varepsilon - 1} \sqrt{\frac{\log k}{n}}. \tag{22}$$

Proof. According to (20), we know the estimator is unbiased. Since each N_{C_x} is a sum of n independent Bernoulli random variables, $\hat{p}(x)$'s are sub-Gaussian with varaince proxy $\frac{4(e^\epsilon+1)^2}{n(e^\epsilon-1)^2}$. Hence, by Theorem 1.16 from (Rigollet, 2015), we get the result in (22).

□

In (Bassily & Smith, 2015), a matching lower bound of $\Omega(\frac{1}{\varepsilon}\sqrt{\frac{\log k}{n}})$ when $\varepsilon = O(1)$ is proved for LDP heavy hitter estimation algorithms. The above theorem shows that the proposed algorithm has optimal performance. We remark that this scheme has communication complexity of $\log k$ bits per user, and the total computation complexity is $O(k \log k + n)$. The dependence on k is usually undesirable in this problem, and therefore more sophisticated schemes are designed, which require higher communication complexity or shared randomness.

5. Constant Bits of Communication is not Optimal for Heavy Hitter Estimation

The previous section showed that with $\ell = \log k + 2$ bits of communication per user we can solve heavy hitters problem optimally. In this section, we assume that $\ell < \log k - 2$, and prove that there is no private-coin heavy hitter detection scheme that communicates $o(\log n)$ bits per user and is optimal.

Theorem 8. *Let $\ell < \log k - 1$. For all private-coin response schemes $(\{W_i\}_{i=1}^n, \hat{p})$ with only private randomness and ℓ bits of communication, there exists a dataset $X_1, \ldots, X_n$ with $n > 12(2^\ell+1)^2$, and $x_0 \in [k]$ such that:*

$$\mathbb{E}\left[\|\hat{p}(Y^n)(x_0) - \frac{N_x(X^n)}{n}\|_\infty \right] \geq \frac{1}{2^{\ell+2}+4},$$

where $Y_i = W_i(X_i)$ for $i \in [n]$.

Proof. We will use the probabilistic method to show the existence of such a dataset. To do so, we design a dataset generating process, and show that the expected ℓ_∞ loss over the process and randomness induced by the channels is large, which is smaller than the expected ℓ_∞ loss for the worst dataset.

Similar to Section 3, recall that each W_i can be represented by a $k \times 2^\ell$ transition probability matrix (TPM) where for user i, $W_i(x,y) = \Pr\left(Y_i = y | X_i = x\right)$. Consider distributions $p_1, \ldots, p_n$ over $[k]$, and suppose the data at user i, X_i is generated from p_i. Then q_i, the output distribution of Y_i is given by $W_i^T p_i$. We will restrict to distributions p_i's to have support over the first $2^\ell + 1$ symbols. Namely, for all $2^\ell + 1 < x \leq k$, $p_i(x) = 0$. Similar to the proof of Theorem 6, since the output dimension is 2^ℓ, for each i, there exists a non-zero vector $\mathbf{e}_i \in \mathbb{R}^k$, such that $\mathbf{e}_i(x) = 0$ for $2^\ell + 1 < x$, and $W_i^T \mathbf{e}_i = 0$. Further, recall that since W_i is a TPM, $\sum_{x=1}^k \mathbf{e}_i(x) = 0$. Therefore, upon normalizing, assume $\|e_i\|_1 = 2$. Let

$$p_i = \max\{\mathbf{e}_i, 0\}, p_i' = \max\{-\mathbf{e}_i, 0\}.$$

Then p_i and p_i' are valid distributions over $[k]$ and effective support only $\{1, \ldots, 2^\ell + 1\}$, and $\|p_i - p_i'\|_1 = 2$. Similarly construct p_i, p_i' for each $i = 1, \ldots, n$. Then,

$$2n = \sum_{x=1}^{k} \sum_{i=1}^{n} |p_i(x) - p_i'(x)| = \sum_{x=1}^{2^\ell+1} \sum_{i=1}^{n} |p_i(x) - p_i'(x)|,$$

where we use that p_i, and p'_i are supported onlyover the first $2^\ell + 1$ symbols. Hence there exists $x_0 \in [2^\ell + 1]$, such that

$$\sum_{i=1}^{n} |p_i(x_0) - p'_i(x_0)| \ge \frac{2n}{2^\ell + 1}.$$

Without loss of generality, assume $\forall i, p_i(x_0) \le p'_i(x_0)$. Then the above equation becomes

$$\sum_{i=1}^{n} p'_i(x_0) - \sum_{i=1}^{n} p_i(x_0) \ge \frac{2n}{2^\ell + 1}. \tag{23}$$

Now consider two datasets generated as follows. X^n satisfies $\forall i \in [n], X_i \sim p_i$ and X'^n satisfies $\forall i \in [n], X'_i \sim p_i$. Moreover since

$$W_i^T p_i = q_i = W_i^T p'_i,$$

the output distribution Y^n is identical for X'^n, and X^n.

Let $N_{x_0}(X^n)$ and $N_{x_0}(X'^n)$ be the number of appearances of x_0 in X^n and X^n. Then by (23),

$$\mathbb{E}\left[N_{x_0}(X^n)\right] - \mathbb{E}\left[N_{x_0}(X'^n)\right] > \frac{2n}{2^\ell + 1}.$$

Moreover, since N_{x_0} are sum of independent binary random variables, Var $(N_{x_0}) \le n/4$. Now suppose $\ell < \frac{1}{4}\log n - 1$, then $n/(2^\ell + 1) > n^{3/4}$. Therefore, by Chebychev's inequality, for large n,

$$\Pr\left(N_{x_0}(X^n) - N_{x_0}(X'^n) > \frac{n}{2^\ell + 1}\right) > 0.9.$$

Since the two output distributions are indistinguishable, we have the error is at least $\frac{n}{2^{\ell+1}+2}$ for one of the cases if this event happens. Hence the expected loss would be at least $0.9 \times \frac{n}{2^{\ell+1}+2} > \frac{n}{2^{\ell+2}+4}$. □

Hence we can see when $\ell = O(1)$. We cannot learn the frequency reliably up to accuracy better than a constant. Moreover, when $\ell = o(\log n + \log(1/\varepsilon))$, we get

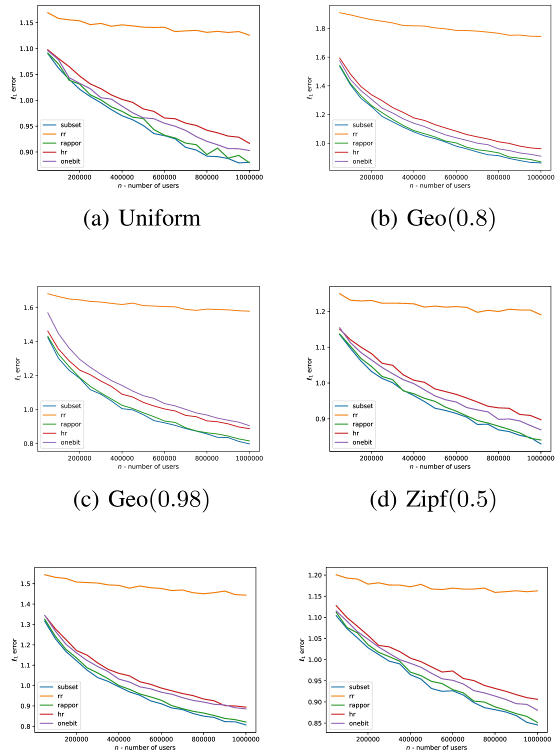
$$\frac{1}{2^{\ell+1} + 1} > \sqrt{\frac{\log k}{n\varepsilon^2}},$$

implying that optimal frequency estimation algorithms must require $\Omega(\log n + \log(1/\varepsilon))$ bits of communication when there is no public randomness. Similar to Section 3, the result also extends to non-private settings.

6. Experiments

We conduct empirical evaluations for the one bit distribution learning algorithm without public randomness proposed in Section 2. We compare the proposed algorithm (onebit) with other algorithms including Randomized Response (RR) (Warner, 1965), RAPPOR (Erlingsson et al., 2014), Hadamard Response (HR) (Acharya et al., 2019) and subset selection (subset) (Ye & Barg, 2018). To obtain samples, we generate synthetic data from various classes of distributions including uniform distribution, geometric distributions with parameter 0.8 and 0.98, Zipf distributions with parameter 1.0 and 0.5 and Two-step distribution. We conduct the experiments for $k = 1000$ and $\varepsilon = 1$. The results are shown in Figure 1. Each point is the average of 30 independent experiments.

From the figures, we can see the performance of our proposed scheme is comparable to the best among all schemes for various kinds of distributions. And the communication complexity is only one bit while the least among others is $\Omega(\log k)$ bits (Acharya et al., 2019).



(a) Uniform (b) Geo(0.8)

(c) Geo(0.98) (d) Zipf(0.5)

(e) Zipf(1.0) (f) Two Steps

Figure 1. ℓ_1-error for $k = 1000$, p from Uniform, Geo(0.8), Geo(0.98), Zipf(0.5), Zipf(1.0) and Two-step distributions

.

Acknowledgements

We thank the anonymous reviewers for pointing out the one-bit public-coin schemes in (Bassily & Smith, 2015), and other valuable suggestions, and Sourbh Bhadane for feedback on the manuscript. This research is supported by NSF-CCF-1657471, and NSF-CCF-1846300 (CAREER).

References

Acharya, J., Canonne, C. L., and Tyagi, H. Distributed simulation and distributed inference. *arXiv preprint arXiv:1804.06952*, 2018a.

Acharya, J., Canonne, C. L., and Tyagi, H. Inference under information constraints I: lower bounds from chi-square contraction. abs/1812.11476, 2018b.

Acharya, J., Sun, Z., and Zhang, H. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1120–1129. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/acharya19a.html.

Bassily, R. Linear queries estimation with local differential privacy. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 721–729. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/bassily19a.html.

Bassily, R. and Smith, A. Local, private, efficient protocols for succinct histograms. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pp. 127–135. ACM, 2015.

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

Beimel, A., Nissim, K., and Omri, E. Distributed private data analysis: Simultaneously solving how and what. In *Proceedings of the 28th Annual International Cryptology Conference*, CRYPTO '08, pp. 451–468, Berlin, Heidelberg, 2008. Springer.

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

Braverman, M., Garg, A., Ma, T., Nguyen, H. L., and Woodruff, D. P. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pp. 1011–1020. ACM, 2016.

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

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

Dagan, Y. and Shamir, O. Detecting correlations with little memory and communication. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1145–1198. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/dagan18a.html.

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

Devroye, L. and Lugosi, G. *Combinatorial Methods in Density Estimation*. Springer, 2001.

Diakonikolas, I., Grigorescu, E., Li, J., Natarajan, A., Onak, K., and Schmidt, L. Communication-efficient distributed learning of discrete distributions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6394–6404. Curran Associates, Inc., 2017.

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

Ding, B., Kulkarni, J., and Yekhanin, S. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pp. 3571–3580, 2017.

Dinur, I. and Nissim, K. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pp. 202–210, New York, NY, USA, 2003. ACM.

Duchi, J., Wainwright, M. J., and Jordan, M. I. Local privacy and minimax bounds: Sharp rates for probability estimation. In *Advances in Neural Information Processing Systems*, pp. 1529–1537, 2013a.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *Proceedings of the 54st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '13, pp. 429–438. IEEE, 2013b.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 2017.

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

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC ’06, pp. 265–284, Berlin, Heidelberg, 2006. Springer.

Erlingsson, Ú., Pihur, V., and Korolova, A. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, CCS ’14, pp. 1054–1067, New York, NY, USA, 2014. ACM.

Han, Y., Özgür, A., and Weissman, T. Geometric lower bounds for distributed parameter estimation under communication constraints. volume 75 of *Proceedings of Machine Learning Research*, pp. 3163–3188. PMLR, 2018.

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

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

Kairouz, P., Bonawitz, K., and Ramage, D. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML’16, pp. 2436–2444, 2016.

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

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pp. 94–103. IEEE, 2007.

Qin, Z., Yang, Y., Yu, T., Khalil, I., Xiao, X., and Ren, K. Heavy hitter estimation over set-valued data with local differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 192–203. ACM, 2016.

Rigollet, P. Lecture notes. 18.s997: High dimensional statistics. *MIT Courses/Mathematics, 2015. https://ocw.mit.edu/courses/mathematics/18-s997-high-dimensional-statistics-spring-2015*, 2015.

Smith, A., Thakurta, A., and Upadhyay, J. Is interaction necessary for distributed private learning? In *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 58–77. IEEE, 2017.

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

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

Wang, S., Huang, L., Wang, P., Nie, Y., Xu, H., Yang, W., Li, X.-Y., and Qiao, C. Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025*, 2016.

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

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

Ye, M. and Barg, A. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64:5662–5676, 2018.