Estimating Sparse Distributions Under Joint Communication and Privacy Constraints

Surin Ahn, Wei-Ning Chen, and Ayfer Özgür
Department of Electrical Engineering, Stanford University
Email: {surinahn, wnchen, aozgur}@stanford.edu

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834610

***Abstract*—We consider the problem of estimating a d-dimensional, s-sparse discrete distribution from independent samples subject to a joint b-bit communication constraint and ε-local differential privacy constraint. As an intermediate step, we introduce the *Privatized Random Hashing* (PRH) scheme, which concatenates a hashing-based quantization strategy with the randomized response privacy mechanism. Despite its simplicity, PRH turns out to achieve the order-optimal minimax estimation error and sample complexity in the standard (non-sparse) estimation setting, for *all* communication and privacy regimes. We then address the sparse case by developing a two-stage, non-interactive estimation scheme based on PRH in which the first half of samples are used to *localize* the unknown support of the distribution, and the remaining samples are used to obtain precise estimates of the individual probabilities. Using this scheme, we characterize the minimax sample complexity of the sparse case up to logarithmic factors, unifying existing results in the literature that considered communication and privacy constraints separately.**

I. Introduction

Recent years have witnessed an exponential growth in the volume of data generated from distributed sources at the network edge, including smartphones, wireless sensors, and wearable devices. For many machine learning tasks – particularly in federated learning and analytics [22] – the devices are required to transmit information over bandwidth-constrained wireless links to a central server or data aggregator. The resulting communication cost is often a major bottleneck to achieving the desired level of utility or accuracy for the task at hand. Moreover, when sensitive data is involved, it is critical to protect the privacy of individual users. A widely adopted notion of privacy is *local differential privacy* (LDP) [17], [25], which ensures that the central server cannot learn too much (in a statistical sense) about any individual's data. A large body of work studies the effects of imposing communication [1]–[3], [8]–[11], [15], [19], [20] and privacy [1], [2], [5]–[7], [13], [17], [18], [21], [26], [27] constraints on the statistical problem of estimating a discrete distribution from its samples.

In many applications such as language modeling and genomics, the distribution of interest is supported on a small but unknown subset (of size s) of the ambient domain (of size d, where $d \gg s$). Results from high-dimensional statistics [24], compressed sensing [12], and group testing [16] show that by exploiting such sparsity in the problem structure, the "effective dimension" of the problem can be made much smaller than d (e.g., in compressed sensing, d is replaced by $s\log(d/s)$). Inspired by these works, recent papers have demonstrated similar utility gains for sparse distribution estimation under communication and privacy constraints [4], [14], [26]. However, thus far, these constraints have largely been studied separately.

This paper makes progress toward unifying the aforementioned works by studying the effect of imposing *joint* communication and privacy constraints on one's ability to estimate a sparse distribution from its samples. First, we introduce the *Privatized Random Hashing* (PRH) scheme, which concatenates a hashing-based quantization strategy [4], [9] with the randomized response privacy mechanism [25]. We show that, surprisingly, PRH achieves the order-optimal minimax estimation error and sample complexity in *all communication and privacy regimes*, despite its apparent simplicity. For the sparse case, we develop a two-stage, non-interactive estimation scheme based on PRH in which the first half of samples are used to *localize* the unknown support of the distribution, and the remaining samples are used to obtain more precise probability estimates. Our resulting characterization of the sample complexity is tight up to logarithmic factors and recovers existing results in the literature depending on which constraint is more stringent. Furthermore, our upper bound applies to all privacy levels, whereas prior results hold only for $\varepsilon = O(\log d)$.

A. Notation and Setup

There are n clients, each of whom observes a sample $X_i \in \mathcal{X}$ drawn from an unknown discrete distribution $p \in \Delta_d$, where

$$\Delta_d \triangleq \left\{ p = (p_1, \dots, p_d) \in [0,1]^d \;\middle|\; \sum_{j\in[d]} p_j = 1 \right\}$$

is the set of all d-dimensional discrete distributions. Given X_i, the i^{th} client generates a message $Y_i \in \mathcal{Y}$ that it transmits to the central server. The message is generated using an encoding channel[1] denoted by the conditional probability $Q_i(\cdot \,|\, X_i)$, which must satisfy two constraints:

1) **Local differential privacy (LDP).** A channel Q is said to satisfy ε-LDP if

$$\frac{Q(y \,|\, x)}{Q(y \,|\, x')} \le e^{\varepsilon}, \quad \forall x, x' \in \mathcal{X}, \forall y \in \mathcal{Y}.$$

[1] An encoding channel is a randomized mapping that can potentially depend on *shared randomness*, i.e., a random variable U accessible to the clients and server. Schemes which utilize shared randomness are also known as *public-coin* schemes. In *private-coin* schemes, all channels are independent. For simplicity, we suppress the dependence on U in our notation.

2) b**-bit communication constraint.** The set of possible messages, $\mathcal{Y}$, satisfies the b-bit communication constraint if $|\mathcal{Y}| \le 2^b$, i.e., every message $Y_i \in \mathcal{Y}$ can be expressed with b bits.

Given the n messages $Y^n \triangleq (Y_1, Y_2, \ldots, Y_n)$ produced by the encoding channels $Q^n \triangleq (Q_1, Q_2, \ldots, Q_n)$, the server generates an estimate $\hat{p}(Y^n)$ of the underlying distribution p. A common objective is to design a scheme $(Q^n, \hat{p}(Y^n))$ to achieve the minimax estimation error

$$r(\ell, n, b, \varepsilon) \triangleq \min_{(Q^n, \hat{p})} \max_{p \in \Delta_d} \mathbb{E}\Big[\ell(p, \hat{p}(Y^n))\Big],$$

where $\ell = \|\cdot\|_1$ or $\|\cdot\|_2^2$.

A slightly different notion is the *minimax sample complexity* $n^*(\alpha, \ell, \Delta_d, b, \varepsilon)$, which is the smallest n for which we can achieve

$$\Pr\Big(\ell(p, \hat{p}(Y^n)) \le \alpha\Big) \ge 0.9, \quad \forall p \in \Delta_d,$$

where $\alpha \in (0, 1)$ is an accuracy parameter. It is common for ℓ to be the *total variation distance*, $\ell_{\text{TV}}(p, \hat{p}) \triangleq \sup_{\mathcal{A} \subseteq [d]} |p(\mathcal{A}) - \hat{p}(\mathcal{A})| = \frac{1}{2}\|p - \hat{p}\|_1$.

In this work, we also consider the task of estimating s-sparse distributions:

$$\Delta_{d,s} \triangleq \left\{ p = (p_1, \ldots, p_d) \in [0,1]^d \;\middle|\; \sum_{j \in [d]} p_j = 1,\ \|p\|_0 \le s \right\}.$$

The minimax sample complexity in this case is denoted by $n^*(\alpha, \ell, \Delta_{d,s}, b, \varepsilon)$.

B. Related Works

Distribution estimation under communication [1]–[3], [8], [15], [19], [20] and LDP [1], [2], [5]–[7], [17], [18], [21], [27] constraints has been studied extensively. The ℓ_1 minimax error scales as $\Theta(\sqrt{d^2/(n \min\{2^b, d\})})$ under communication constraints and $\Theta(\sqrt{d^2/(n \min\{e^\varepsilon, (e^\varepsilon - 1)^2\})})$ under LDP constraints. The ℓ_{TV} (equivalently, ℓ_1) sample complexity is $\Theta(d^2/(\alpha^2 \min\{2^b, d\}))$ and $\Theta(d^2/(\alpha^2 \min\{e^\varepsilon, (e^\varepsilon - 1)^2\}))$, respectively. Joint communication and LDP constraints were studied in [13], and it was shown that the Recursive Hadamard Response (RHR) scheme achieves the minimax error of $\Theta(\sqrt{d^2/(n \min\{e^\varepsilon,\ (e^\varepsilon - 1)^2,\ 2^b,\ d\})})$. This implies that the convergence rate is determined by the more stringent of the two constraints, allowing the other to be satisfied "for free."

For sparse distribution estimation under communication constraints, [4] obtained an upper bound of $O\Big(\frac{s^2 \max\{\log(d/s), 1\}}{\alpha^2 \min\{2^b, s\}}\Big)$ and a lower bound of $\Omega\Big(\max\Big\{\frac{s^2}{\alpha^2 \min\{2^b, s\}}, \frac{s^2 \max\{\log(d/s), 1\}}{\alpha 2^b}\Big\}\Big)$, which exhibit a logarithmic gap in various parameter regimes. Subsequently, [14] showed that the extra $\log(d/s)$ factor can be eliminated from the convergence rate if the sample size is sufficiently large. Under LDP constraints in the high-privacy regime ($\varepsilon = O(1)$), [4] established a tight sample complexity of $\Theta\Big(\frac{s^2 \max\{\log(d/s), 1\}}{\alpha^2 \varepsilon^2}\Big)$ using the 1-bit Hadamard Response with a sparse projection. Similar results are reported in [26], and their proposed scheme extends to the medium-privacy regime ($\varepsilon \in [1, \log d]$) with a resulting sample complexity of $O\Big(\frac{s^2 \log(d/s)}{\alpha^2 e^\varepsilon}\Big)$. However, their scheme requires $\Omega(\log s)$ bits of communication, which is strictly sub-optimal when $\varepsilon = O(\log s)$, according to our results.

C. Overview of Results

Our first result establishes the somewhat surprising fact that the Privatized Random Hashing (PRH) scheme achieves the order-optimal minimax estimation error for all privacy and communication regimes. The proof is provided in Section II-B.

Theorem 1 (Non-sparse estimation error)**.** *For all $p \in \Delta_d$, the estimation error of Privatized Random Hashing (PRH) satisfies*

$$\mathbb{E}\Big[\ \|\hat{p} - p\|_2^2\ \Big] \preceq \frac{d}{n \cdot \min\Big\{e^\varepsilon,\ (e^\varepsilon - 1)^2,\ 2^b,\ d\Big\}}$$

$$\mathbb{E}\Big[\ \|\hat{p} - p\|_1\ \Big] \preceq \frac{d}{\sqrt{n \cdot \min\Big\{e^\varepsilon,\ (e^\varepsilon - 1)^2,\ 2^b,\ d\Big\}}}.$$

Moreover, if $n \cdot \min\Big\{e^\varepsilon,\ (e^\varepsilon - 1)^2,\ 2^b,\ d\Big\} \ge d^2$, then PRH is order-optimal.

To the best of our knowledge, the only other scheme with this performance guarantee is the Recursive Hadamard Response (RHR) [13]. Our result demonstrates that a conceptually simpler scheme achieves the same rate-optimal performance[2], and disproves a prior belief in the literature that performing separate quantization and privatization is always strictly sub-optimal. For instance, [13] shows that the concatenation of subset selection [27] (which is optimal under LDP constraints) with grouping-based quantization [19] (which is optimal under communication constraints) yields an ℓ_2 error rate that grows *quadratically* with d, in contrast to the linear dependence exhibited by PRH and RHR. Another benefit of our concatenated scheme is that any system which already implements one of the components (either random hashing or randomized response) can easily satisfy the "missing" constraint with minimal modifications to the existing system. Finally, it can be shown that the estimation error of PRH has a strictly better leading constant than that of RHR, though we focus on order-wise bounds in the present paper.

From Theorem 1, one can derive an upper bound on the sample complexity $n^*(\alpha, \ell_{\text{TV}}, \Delta_d, b, \varepsilon)$. Using Markov's inequality, it follows that $\forall p \in \Delta_d$,

$$\Pr\Big(\ell(p, \hat{p}(Y^n)) > \alpha\Big) \le \frac{1}{\alpha}\mathbb{E}\Big[\ell_{\text{TV}}(\hat{p}, p)\Big] \preceq \frac{1}{\alpha}\frac{d}{\sqrt{n \cdot \min\Big\{e^\varepsilon,\ (e^\varepsilon - 1)^2,\ 2^b,\ d\Big\}}}.$$

[2]However, we acknowledge that this simplicity comes at the cost of requiring shared randomness (whereas RHR is strictly a private-coin scheme).

Therefore, taking

$$n = C \cdot \frac{d^2}{\alpha^2 \cdot \min\{e^\varepsilon,\, (e^\varepsilon - 1)^2,\, 2^b,\, d\}}$$

for a sufficiently large constant $C > 0$ ensures that $\Pr(\ell(p, \hat{p}(Y^n)) > \alpha)$ is bounded above by 0.1. A matching lower bound is obtained by combining existing sample complexity lower bounds in the communication- [3] and LDP-constrained settings (see [6] and references therein).

Corollary 1 (Non-sparse sample complexity)**.** *Privatized Random Hashing (PRH) achieves the minimax sample complexity for estimating distributions in Δ_d under joint b-bit communication and ε-LDP constraints, given by*

$$n^*(\alpha, \ell_{TV}, \Delta_d, b, \varepsilon) = \Theta\left(\frac{d^2}{\alpha^2 \cdot \min\left\{e^\varepsilon,\, (e^\varepsilon - 1)^2,\, 2^b,\, d\right\}}\right).$$

For the sparse setting, we derive an upper bound on the sample complexity using a two-stage, non-interactive estimation scheme based on PRH (described in Section III). Our upper bound holds in all communication and privacy regimes, extending prior results which hold only when $\varepsilon = O(\log d)$. A lower bound for $\varepsilon = O(1)$ (the high-privacy regime) is obtained by combining Theorem 1 and Theorem 2 from [4].

Theorem 2 (Sparse sample complexity)**.** *The minimax sample complexity for estimating distributions in $\Delta_{d,s}$ under joint b-bit communication and ε-LDP constraints satisfies*

$$n^*(\alpha, \ell_{TV}, \Delta_{d,s}, b, \varepsilon) = O\left(\frac{s^2 \max\{\log(d/s),\, 1\}}{\alpha^2 \cdot \min\left\{e^\varepsilon,\, (e^\varepsilon - 1)^2,\, 2^b,\, s\right\}}\right).$$

When $\varepsilon = O(1)$ (the high-privacy regime),

$$n^*(\alpha, \ell_{TV}, \Delta_{d,s}, b, \varepsilon) = \Omega\left(\max\left\{\frac{s^2 \max\{\log(d/s),\, 1\}}{\alpha^2 \varepsilon^2},\, \frac{s^2 \max\{\log(d/s),\, 1\}}{\alpha 2^b},\, \frac{s^2}{\alpha^2 \min\{2^b,\, s\}}\right\}\right).$$

As in the non-sparse case [13], our upper bound in Theorem 2 is dictated by the more stringent of the two constraints, allowing the other one to be satisfied "for free." The lower bound is tight when $\varepsilon^2 \preceq 2^b$, i.e., when the privacy level is more stringent than the communication budget. In this case, the sample complexity becomes $\Theta\left(\frac{s^2 \max\{\log(d/s),1\}}{\alpha^2 \varepsilon^2}\right)$, recovering the result of [4, Theorem 1] from the LDP-constrained setting. On the other hand, when $\varepsilon^2 \succeq 2^b$, Theorem 2 recovers [4, Theorem 2] from the communication-constrained setting, and the upper and lower bounds are separated by at most a logarithmic factor.

II. Privatized Random Hashing

In this section, we introduce and analyze the *Privatized Random Hashing* (PRH) scheme, culminating in the proof of Theorem 1.

A. Algorithm Description

PRH comprises two distinct components. Each client first quantizes its sample down to at most b bits using a random hashing approach [4], [9]. Then, the randomized response mechanism [25] is applied to the quantized sample to satisfy the LDP constraint.

More precisely, let $k \triangleq \min\{b, \lceil \varepsilon \log_2 e \rceil, \lfloor \log d \rfloor\}$, and let $\{h_i : [d] \to [2^k],\ i \in [n]\}$ be independent hash functions which are generated via public randomness and are known to both the clients and the server. Each hash function satisfies

$$\Pr(h_i(j) = y) = \frac{1}{2^k}, \quad \forall j \in [d], \forall y \in [2^k].$$

The i^{th} client first maps its sample X_i to $\tilde{Y}_i = h_i(X_i) \in [2^k]$. Next, it privatizes $\tilde{Y}_i$ using 2^k-Randomized Response (2^k-RR) and sends the resulting message Y_i to the server:

$$Y_i = \begin{cases} \tilde{Y}_i & \text{with probability } \frac{e^\varepsilon}{e^\varepsilon + 2^k - 1}, \\ \tilde{Y}_i' \in [2^k] \setminus \{\tilde{Y}_i\} & \text{with probability } \frac{1}{e^\varepsilon + 2^k - 1}. \end{cases}$$

Note that each client's message can be encoded in $k \le b$ bits. Moreover, each message satisfies the ε-LDP constraint due to the privacy guarantees of 2^k-RR.

Given all n messages, the server computes

$$N(j) \triangleq \left|\left\{i \in [n] \,:\, h_i(j) = Y_i\right\}\right|$$

for each $j \in [d]$, which is the number of messages Y_i such that symbol j lies in the pre-image of Y_i under h_i. The final estimator is given by

$$\hat{p}_j = \frac{1}{2^k - 1}\left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\right)\left(\frac{2^k}{n} N(j) - 1\right). \tag{1}$$

Lemma 1. *$N(j)$ is distributed as* $\mathsf{Binomial}(n, \gamma \cdot p_j + \beta)$*, where $\gamma = \left(1 - \frac{1}{2^k}\right)\left(\frac{e^\varepsilon - 1}{e^\varepsilon + 2^k - 1}\right)$ and $\beta = \frac{1}{2^k}$.*

Proof. First, we calculate the probability that a symbol $j \in [d]$ lies in the pre-image of a *non-privatized* message $\tilde{Y}_i$ under hash function h_i:

$$\Pr(\tilde{Y}_i = h_i(j)) = p_j + \frac{1}{2^k}(1 - p_j) = \left(1 - \frac{1}{2^k}\right) \cdot p_j + \frac{1}{2^k}.$$

Therefore, combining this with the definition of the 2^k-RR scheme, we have

$$\begin{aligned} \Pr(Y_i = h_i(j)) &= \left(\frac{e^\varepsilon}{e^\varepsilon + 2^k - 1}\right)\left[\left(1 - \frac{1}{2^k}\right) \cdot p_j + \frac{1}{2^k}\right] \\ &\quad + \frac{1}{e^\varepsilon + 2^k - 1}\left[1 - \left(1 - \frac{1}{2^k}\right) \cdot p_j - \frac{1}{2^k}\right] \\ &= \left(1 - \frac{1}{2^k}\right)\left(\frac{e^\varepsilon - 1}{e^\varepsilon + 2^k - 1}\right) \cdot p_j + \frac{1}{2^k}. \end{aligned}$$

□

Given Lemma 1, it is straightforward to verify that the estimator $\hat{p}$ is unbiased for p.

Corollary 2. *$\forall j \in [d]$, the estimator $\hat{p}_j$ given in* (1) *satisfies $\mathbb{E}[\hat{p}_j] = p_j$, i.e., it is unbiased for p_j.*

B. Proof of Theorem 1

Note that (1) can be written as $\hat{p}_j = \frac{1}{\gamma n} N(j) - \frac{\beta}{\gamma}$, where γ, β are as defined in Lemma 1. For each $j \in [d]$, we have

$$\begin{aligned}\mathbb{E}\Big[(\hat{p}_j - p_j)^2\Big] = \mathrm{Var}(\hat{p}_j) &= \frac{1}{\gamma^2 n^2} \cdot \mathrm{Var}(N(j)) \\ &\le \frac{1}{\gamma^2 n^2} \cdot \mathbb{E}[N(j)] \\ &= \frac{1}{\gamma n} \cdot \Big(p_j + \frac{\beta}{\gamma}\Big), \end{aligned} \tag{2}$$

where the first equality follows from Corollary 2, the inequality follows from the fact that if $U \sim \mathsf{Binomial}(n, q)$ then $\mathrm{Var}(U) \le nq = \mathbb{E}[U]$, and the final equality uses Lemma 1.

Therefore, the overall estimation error can be bounded as

$$\begin{aligned}\mathbb{E}\Big[\,\|\hat{p} - p\|_2^2\,\Big] = \sum_{j\in[d]} \mathbb{E}\Big[(\hat{p}_j - p_j)^2\Big] &\le \frac{1}{\gamma n}\left(\sum_{j\in[d]} \Big(p_j + \frac{\beta}{\gamma}\Big)\right) \\ &= \frac{1}{\gamma n}\Big(1 + \frac{\beta}{\gamma} \cdot d\Big) \\ &\preceq \frac{d}{n} \cdot \frac{\beta}{\gamma^2} \\ &\preceq \frac{d}{n \cdot 2^k}\Big(\frac{e^\varepsilon + 2^k}{e^\varepsilon - 1}\Big)^2,\end{aligned}$$

where the second inequality uses the fact that $\frac{\beta}{\gamma} = \frac{1}{2^k - 1}\Big(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\Big) \ge \frac{1}{d}$ since $1 \le k \le \log d$. Taking $k = \min\{b,\, \lceil \varepsilon \log_2 e \rceil,\, \lfloor \log d \rfloor\}$ results in

$$\begin{aligned}\mathbb{E}\Big[\,\|\hat{p} - p\|_2^2\,\Big] &\preceq \frac{d}{n \cdot \min\{2^b,\, e^\varepsilon,\, d\}}\Big(\frac{e^\varepsilon}{e^\varepsilon - 1}\Big)^2 \\ &\preceq \max\left\{\frac{d}{n \cdot \min\{2^b,\, d\}},\quad \frac{d}{n} \cdot \frac{e^\varepsilon}{(e^\varepsilon - 1)^2}\right\} \\ &\asymp \frac{d}{n \cdot \min\Big\{e^\varepsilon,\, (e^\varepsilon - 1)^2,\, 2^b,\, d\Big\}},\end{aligned}$$

where the second line follows from the observation that if $e^\varepsilon = O(2^b)$ then $\mathbb{E}\,\|\hat{p} - p\|_2^2 \preceq \frac{d}{n} \cdot \frac{e^\varepsilon}{(e^\varepsilon - 1)^2}$, and if $e^\varepsilon = \Omega(2^b)$ then $\mathbb{E}\,\|\hat{p} - p\|_2^2 \preceq \frac{d}{n \cdot \min\{2^b, d\}}$.

The ℓ_1 estimation error can be established using Jensen's inequality and the Cauchy-Schwarz inequality as follows:

$$\begin{aligned}\mathbb{E}\Big[\,\|\hat{p} - p\|_1\,\Big] &\le \sqrt{\mathbb{E}\Big[\,\|\hat{p} - p\|_1^2\,\Big]} \le \sqrt{d \cdot \mathbb{E}\Big[\,\|\hat{p} - p\|_2^2\,\Big]} \\ &\preceq \frac{d}{\sqrt{n \cdot \min\Big\{e^\varepsilon,\, (e^\varepsilon - 1)^2,\, 2^b,\, d\Big\}}}.\end{aligned}$$

We obtain matching lower bounds by combining the results of [27], [8], [19].

Remark 1 (Amount of shared randomness)**.** *Note that our analysis only requires the hash functions to be* pairwise independent*; that is, for any hash function h, symbols $j_1 \ne j_2 \in [d]$, and bins $y_1, y_2 \in [2^k]$, we have $\Pr\left(h(j_1) = y_1 \wedge h(j_2) = y_2\right) = 1/(2^k)^2$. It is well known that $O(\log d + k) = O(\log d)$ bits of random seeds suffice to generate such pairwise independent hash functions, and hence the amount of shared randomness used in our scheme can be reduced to $O(\log d)$.*

III. Estimating Sparse Distributions

We are now interested in characterizing the minimax sample complexity, $n^*(\alpha, \ell_{TV}, \Delta_{d,s}, b, \varepsilon)$, for learning s-sparse distributions. We consider an estimation scheme consisting of the following two stages:

1) **Localization:** The messages Y^{n_1} from the first group of $n_1 \triangleq n/2$ clients are employed to produce an estimate, $\hat{\mathcal{S}}$, of the support of the distribution, $\mathcal{S} \triangleq \{j \in [d] : p_j > 0\}$.
2) **Estimation:** Given the estimated support $\hat{\mathcal{S}}$ from the localization stage, the server estimates p using the second group of messages Y^{n_2} from $n_2 \triangleq n/2$ clients.

A similar approach was taken in [4], [14] to estimate sparse distributions under only communication constraints. We now describe each stage in more detail.

a) Localization Stage: The first group of n_1 clients encode their samples using the PRH scheme with $k \triangleq \min\{b,\, \lceil \varepsilon \log_2 e \rceil,\, \lceil \log s \rceil\}$ bits. The server collects the resulting messages and computes, for each $j \in [d]$,

$$M(j) \triangleq \Big|\Big\{i \in [n_1] \,:\, h_i(j) = Y_i\Big\}\Big|.$$

Finally, the estimated support $\hat{\mathcal{S}}$ is taken to be the set of $2s$ symbols with the largest values of the $M(j)$'s. The following lemma says that for sufficiently large n, the estimated support captures most of the probability mass.

Lemma 2. *There exists a constant $C_1 > 0$ such that for $n = C_1 \cdot s^2 \log(d/s)/(\alpha^2 \cdot \min\{e^\varepsilon,\, (e^\varepsilon - 1)^2,\, 2^b,\, s\})$, with probability at least 0.95, we have $p(\hat{\mathcal{S}}) \triangleq \sum_{j\in\hat{\mathcal{S}}} p_j \ge 1 - \alpha/2$.*

Proof Sketch. The proof is nearly identical to that of [4, Lemma 7]. We outline the key steps here, and refer the reader to that paper for further details.

We would like to show that $\Pr(p(\hat{\mathcal{S}}^{\mathsf{c}}) > \frac{\alpha}{2}) \le \frac{1}{20}$. By Lemma 1, $M(j) \sim \mathsf{Binomial}(n/2,\, \gamma \cdot p_j + \beta)$, where $\gamma = \frac{2^k - 1}{2^k}\Big(\frac{e^\varepsilon - 1}{e^\varepsilon + 2^k - 1}\Big)$ and $\beta = \frac{1}{2^k}$. Let E be the event that at most s symbols in $[d] \setminus \mathcal{S}$ appear at least M^* times, where M^* is a threshold to be determined later.

By the law of total probability and Markov's inequality,

$$\Pr\Big(p(\hat{\mathcal{S}}^{\mathsf{c}}) > \frac{\alpha}{2}\Big) \le \frac{\mathbb{E}[p(\hat{\mathcal{S}}^{\mathsf{c}}) \,|\, E]}{\alpha/2} + \Pr(E^{\mathsf{c}}).$$

Applying Markov's inequality and the multiplicative Chernoff bound [23] with $M^* = \frac{n}{2}\beta + \sqrt{3n\beta \log(d/s)}$ yields

$$\Pr(E^{\mathsf{c}}) \le \frac{1}{s} \sum_{j\in[d]\setminus\mathcal{S}} \Pr(M(j) \ge M^*) \le \frac{s}{d} \le \frac{1}{100}$$

where we assume $\frac{s}{d} \le \frac{1}{100}$ as in [4]. To prove the lemma, it now suffices to show that $\mathbb{E}[p(\hat{\mathcal{S}}^{\mathsf{c}}) \,|\, E] \le \frac{\alpha}{50}$. We have

$$\begin{aligned}\mathbb{E}[p(\hat{\mathcal{S}}^{\mathsf{c}}) \,|\, E] &= \sum_{j\in\mathcal{S}} p_j \cdot \Pr(j \notin \hat{\mathcal{S}} \,|\, E) \\ &\le \sum_{j\in\mathcal{S}} p_j \cdot \Pr(M(j) \le M^*) \end{aligned} \tag{3}$$

where the inequality follows from the following two facts: 1) conditioned on event E, $j \notin \hat{\mathcal{S}}$ only if $M(j) \le M^*$; 2) $M(j)$ is independent of event E for $j \in \mathcal{S}$.

We then consider three different sets of symbols: $\mathcal{A} \triangleq \{j \in [d] : p_j \le \frac{\alpha}{60s}\}$, $\mathcal{B} \triangleq \{j \in [d] : \frac{\alpha}{60s} < p_j \le \frac{\beta}{\gamma}\}$, and $\mathcal{C} \triangleq \{j \in [d] : p_j > \frac{\beta}{\gamma}\}$. For set $\mathcal{A}$, it holds that

$$\sum_{j\in\mathcal{A}} p_j \cdot \Pr(M(j) \le M^*) \le \sum_{j\in\mathcal{A}} p_j \le \frac{\alpha}{60}. \tag{4}$$

In what follows, we note that C_1 can be set to a sufficiently large constant such that the statements hold. For $j \in \mathcal{B} \cup \mathcal{C}$, one can show via the multiplicative Chernoff bound that

$$\Pr(M(j) \le M^*) \le \exp\Big(-\frac{\gamma^2 p_j^2 n}{8(\gamma p_j + \beta)}\Big).$$

For $j \in \mathcal{B}$, we have $\Pr(M(j) \le M^*) \le \exp\Big(-\frac{\gamma^2 p_j^2 n}{16\beta}\Big)$, and one can further show that

$$\sum_{j\in\mathcal{B}} p_j \cdot \Pr(M(j) \le M^*) \le \frac{\alpha}{500}. \tag{5}$$

For $j \in \mathcal{C}$, it holds that $\Pr(M(j) \le M^*) \le \frac{\alpha}{1000}$, so

$$\sum_{j\in\mathcal{C}} p_j \cdot \Pr(M(j) \le M^*) \le \frac{\alpha}{1000}. \tag{6}$$

Combining (3), (4), (5), and (6) proves that $\mathbb{E}[p(\hat{\mathcal{S}}^{\mathsf{c}}) \,|\, E] \le \frac{\alpha}{50}$. □

b) Estimation Stage: The second group of n_2 clients again encode their samples using PRH. Given $\hat{\mathcal{S}}$, the server computes

$$M'(j) \triangleq \Big|\Big\{i \in [n_1+1 : n] \,:\, h_i(j) = Y_i\Big\}\Big| \quad \text{for } j \in \hat{\mathcal{S}}.$$

The final estimator is given by

$$\hat{p}_j = \begin{cases} \frac{1}{2^k-1}\Big(\frac{e^\varepsilon+2^k-1}{e^\varepsilon-1}\Big)\Big(\frac{2^k}{n/2}M'(j) - 1\Big) & \text{if } j \in \hat{\mathcal{S}}, \\ 0 & \text{otherwise.} \end{cases}$$

Although our scheme comprises two stages, note that all n clients can encode and transmit their information *simultaneously*, without requiring knowledge of the estimated support, $\hat{\mathcal{S}}$, beforehand. Once the server receives all n messages, it produces its final estimate by performing the two stages of decoding described previously. In the next lemma, we bound the estimation error over $\hat{\mathcal{S}}$.

Lemma 3. *There exists a constant $C_2 > 0$ such that for $n = C_2 \cdot s^2/(\alpha^2 \cdot \min\{e^\varepsilon, (e^\varepsilon-1)^2, 2^b, s\})$, with probability at least 0.95, we have $\sum_{j\in\hat{\mathcal{S}}} |p_j - \hat{p}_j| \le \alpha/2$.*

Proof. $M(j)$ and $M'(j)$ are identically distributed, so $M'(j) \sim \mathsf{Binomial}(n/2, \gamma \cdot p_j + \beta)$. For each $j \in \hat{\mathcal{S}}$, we have $\hat{p}_j = \frac{2}{\gamma n}M'(j) - \frac{\beta}{\gamma}$. Therefore, for $j \in \hat{\mathcal{S}}$,

$$\mathbb{E}\Big[(\hat{p}_j - p_j)^2\Big] = \mathrm{Var}(\hat{p}_j) \le \frac{4}{\gamma^2 n^2}\mathbb{E}[M'(j)] = \frac{2}{\gamma n}\Big(p_j + \frac{\beta}{\gamma}\Big).$$

Note that

$$\sum_{j\in\hat{\mathcal{S}}}\Big(p_j + \frac{\beta}{\gamma}\Big) \le 1 + \frac{\beta}{\gamma}|\hat{\mathcal{S}}| = 1 + \frac{1}{2^k-1}\Big(\frac{e^\varepsilon+2^k-1}{e^\varepsilon-1}\Big)|\hat{\mathcal{S}}|.$$

Hence,

$$\begin{aligned}\mathbb{E}\Big[\sum_{j\in\hat{\mathcal{S}}}(\hat{p}_j - p_j)^2\Big] &\le \frac{2}{\gamma n}\sum_{j\in\hat{\mathcal{S}}}\Big(p_j + \frac{\beta}{\gamma}\Big) \\ &\le \frac{2}{n}\cdot\Big(\frac{2^k}{2^k-1}\Big)^2\cdot\Big(\frac{e^\varepsilon+2^k}{e^\varepsilon-1}\Big)\Big(1 + \frac{1}{2^k}\Big(\frac{e^\varepsilon+2^k}{e^\varepsilon-1}\Big)|\hat{\mathcal{S}}|\Big).\end{aligned}$$

By Jensen's inequality and the Cauchy-Schwarz inequality, and the fact that $|\hat{\mathcal{S}}| \le 2s$, we have

$$\begin{aligned}\mathbb{E}\Big[\sum_{j\in\hat{\mathcal{S}}}|\hat{p}_j - p_j|\Big] &\le \sqrt{|\hat{S}|\cdot\mathbb{E}\Big[\sum_{j\in\hat{\mathcal{S}}}(\hat{p}_j - p_j)^2\Big]} \\ &\le \sqrt{\frac{4s}{n}\Big(\frac{2^k}{2^k-1}\Big)^2\Big(\frac{e^\varepsilon+2^k}{e^\varepsilon-1}\Big)\Big(1 + \frac{1}{2^k}\Big(\frac{e^\varepsilon+2^k}{e^\varepsilon-1}\Big)2s\Big)} \\ &\preceq \sqrt{\frac{s^2}{n}\frac{1}{2^k}\Big(\frac{e^\varepsilon+2^k}{e^\varepsilon-1}\Big)^2} \\ &\preceq \sqrt{\frac{s^2}{n\min\{e^\varepsilon, (e^\varepsilon-1)^2, 2^b, s\}}}.\end{aligned}$$

Finally, by setting $n = C_2 \cdot \frac{s^2}{\alpha^2\cdot\min\{e^\varepsilon,(e^\varepsilon-1)^2,2^b,s\}}$ for a sufficiently large constant $C_2 > 0$ and invoking Markov's inequality, we establish the lemma. □

The upper bound in Theorem 2 is obtained by taking a union bound and combining Lemmas 2 and 3.

IV. Conclusion and Open Problems

In this work, we studied sparse distribution estimation under simultaneous b-bit communication and ε-LDP constraints. Our proposed Privatized Random Hashing (PRH) scheme achieves the order-optimal minimax convergence rate and sample complexity in the non-sparse setting, despite its simplicity. In the sparse case, an extended version of PRH achieves the order-optimal sample complexity up to logarithmic factors. This result unifies existing bounds in the literature, and the upper bound extends prior results to all privacy regimes.

This work naturally leads to a number of open problems for the sparse setting, including closing the gap between our upper and lower bounds in Theorem 2 and extending the lower bound beyond the high-privacy regime. It would also be interesting to characterize the sample complexity and convergence rate when restricted to private-coin schemes, or when permitting sequential interaction between the clients and server.

Acknowledgments

This work was supported in part by NSF Award # NeTS-1817205, a Cisco Systems Stanford Graduate Fellowship, and a National Semiconductor Corporation Stanford Graduate Fellowship.

REFERENCES

[1] Jayadev Acharya, Clément L Canonne, Yuhan Liu, Ziteng Sun, and Himanshu Tyagi. Interactive inference under information constraints. *arXiv preprint arXiv:2007.10976*, 2020.

[2] Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints: Lower bounds from chi-square contraction. In *Conference on Learning Theory*, pages 3–17. PMLR, 2019.

[3] Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints II: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66(12):7856–7877, 2020.

[4] Jayadev Acharya, Peter Kairouz, Yuhan Liu, and Ziteng Sun. Estimating sparse discrete distributions under privacy and communication constraints. In *Algorithmic Learning Theory*, pages 79–98. PMLR, 2021.

[5] Jayadev Acharya and Ziteng Sun. Communication complexity in locally private distribution estimation and heavy hitters. In *International Conference on Machine Learning*, pages 51–60. PMLR, 2019.

[6] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1120–1129. PMLR, 2019.

[7] Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Özgür. Fisher information under local differential privacy. *IEEE Journal on Selected Areas in Information Theory*, 1(3):645–659, 2020.

[8] Leighton Pate Barnes, Yanjun Han, and Ayfer Ozgur. Lower bounds for learning distributions under communication constraints via Fisher information. *Journal of Machine Learning Research*, 21(236):1–30, 2020.

[9] Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Thakurta. Practical locally private heavy hitters. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 2285–2293, Red Hook, NY, USA, 2017. Curran Associates Inc.

[10] Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 127–135, New York, NY, USA, 2015. Association for Computing Machinery.

[11] Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, SIGMOD/PODS '18, page 435–447, New York, NY, USA, 2018. Association for Computing Machinery.

[12] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[13] Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür. Breaking the communication-privacy-accuracy trilemma. *Advances in Neural Information Processing Systems*, 33:3312–3324, 2020.

[14] Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür. Breaking the dimension dependence in sparse distribution estimation under communication constraints. *Conference on Learning Theory*, 2021.

[15] Ilias Diakonikolas, Elena Grigorescu, Jerry Li, Abhiram Natarajan, Krzysztof Onak, and Ludwig Schmidt. Communication-efficient distributed learning of discrete distributions. In *Neural Information Processing Systems*, pages 6394–6404, 2017.

[16] Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[17] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[18] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, page 1054–1067, New York, NY, USA, 2014. Association for Computing Machinery.

[19] Yanjun Han, Pritam Mukherjee, Ayfer Ozgur, and Tsachy Weissman. Distributed statistical estimation of high-dimensional and nonparametric distributions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 506–510. IEEE, 2018.

[20] Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference on Learning Theory*, pages 3163–3188. PMLR, 2018.

[21] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, pages 2436–2444. PMLR, 2016.

[22] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[23] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge university press, 2017.

[24] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

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

[26] Zhongzheng Xiong, Zengfeng Huang, Xiaojun Mao, Jian Wang, and Shan Ying. Compressive privatization: Sparse distribution estimation under locally differentially privacy. *arXiv preprint arXiv:2012.02081*, 2020.

[27] Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676, 2018.