Proceedings of Machine Learning Research vol 132:1–20, 2021 32nd International Conference on Algorithmic Learning Theory

Estimating Sparse Discrete Distributions Under Privacy and Communication Constraints

Jayadev Acharya * ACHARYA@CORNELL.EDU
Cornell University

Peter Kairouz KAIROUZ@GOOGLE.COM
Google

Yuhan Liu * YL2976@CORNELL.EDU
Cornell University

Ziteng Sun * ZS335@CORNELL.EDU
Cornell University

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

We consider the problem of estimating sparse discrete distributions under local differential privacy (LDP) and communication constraints. We characterize the sample complexity for sparse estimation under LDP constraints up to a constant factor, and the sample complexity under communication constraints up to a logarithmic factor. Our upper bounds under LDP are based on the Hadamard Response, a private coin scheme that requires only one bit of communication per user. Under communication constraints we propose public coin schemes based on random hashing functions. Our tight lower bounds are based on recently proposed method of chi squared contractions.

Keywords: sparse distribution estimation, local differential privacy, communication constraint

1. Introduction

Estimating distributions from data samples is a central task in statistical inference. In modern learning systems such as federated learning (Kairouz et al., 2019), data is generated from distributed sources including cell phones, wireless sensors, and smart healthcare devices. Access to such data is subject to severe "local information constraints", such as communication and energy constraints, privacy concerns. For several statistical inference tasks, including distribution estimation, privacy and communication constraints lead to significant degradation in utility (see Section 1.3 for a detailed discussion). Moreover, in some applications, such as web-browsing, genomics, and language modeling, the distribution is often supported over a small unknown subset of the domain. Motivated by the utility gain in high dimensional statistics under sparsity assumptions (Wainwright, 2019), we study the problem of estimating sparse discrete distributions under privacy and communication constraints.

* Research supported by NSF 1815893, NSF 1657471, NSF 1846300 (CAREER), and a Google Faculty Research Fellowship

1.1. Notations and problem set-up

Let $[k] := \{1, 2, ..., k\}$ and $\triangle_k := \{p \in [0,1]^k \colon \sum_{x\in[k]} p(x) = 1\}$ be the set of all distributions over $[k]$. For $p \in \Delta_k$ and $S \subset [k]$, let p^S be the vector restricted on indices in S. Independent samples $X_1, \ldots, X_n$ from an unknown $p \in \triangle_k$ are observed by n users, where user i observes X_i. User i sends a message $Y_i = W_i(X_i)$ to a central server, where $W_i : [k] \to \mathcal{Y}$ is a randomized mapping (channel) with

$$W_i(y \mid x) = \Pr\left(Y_i = y \mid X_i = x\right).$$

We consider privacy and communication constrained messages in this paper, which can be enforced by restricting W_is to belong to a class $\mathcal{W}$ of *allowed* channels.

Local Differential Privacy (LDP). A channel $W : [k] \to \mathcal{Y} = \{0,1\}^*$ is ε-LDP if

$$\sup_{y\in\mathcal{Y}} \sup_{x,x'\in\mathcal{X}} \frac{W(y \mid x)}{W(y \mid x')} \leq e^{\varepsilon}. \tag{1}$$

$\mathcal{W}_\varepsilon = \{W : W \text{ is } \varepsilon\text{-LDP}\}$ is the set of all ε-LDP channels.

Communication constraints. Let $\ell < \log k$, and $\mathcal{W}_\ell = \{W : [k] \to \mathcal{Y} = \{0,1\}^\ell\}$ be the set of channels that output ℓ-bit messages, and thus characterize communication constraints.

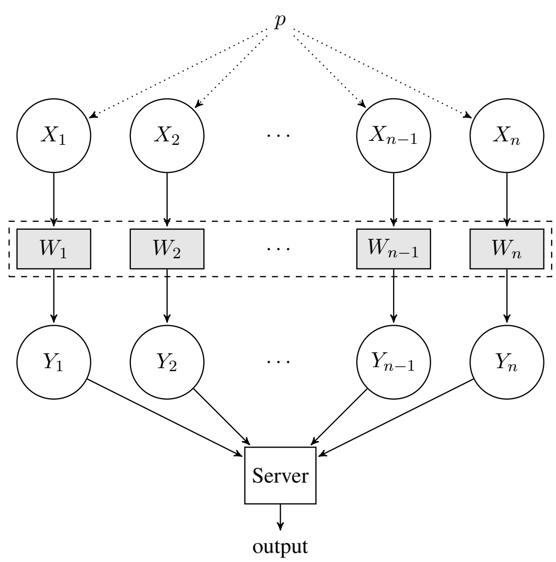

Figure 1: Distributed inference with simultaneous message passing (SMP) protocol.

Distribution estimation. Let $d : \triangle_k \times \triangle_k \to \mathbb{R}_+$ be a distance measure. A distribution estimation protocol under constraints $\mathcal{W}$ is a set of channels $W^n := (W_1, W_2, \ldots, W_n) \in \mathcal{W}^n$ and an estimator $\hat{p} : \mathcal{Y}^n \to \triangle_k$. Upon observing the n output messges $Y^n := (Y_1, Y_2, \ldots, Y_n)$, the central server outputs an estimate $\hat{p}(Y^n)$ of the underlying unknown distribution p. Let $\alpha > 0$ be an accuracy parameter. The minimax sample complexity for estimation is

$$\mathrm{SC}(\alpha, \triangle_k, \mathcal{W}, d) := \arg\min_n \left\{ \min_{\hat{p}} \min_{W^n\in\mathcal{W}^n} \max_{p\in\triangle_k} \Pr\left(d(p, \hat{p}(Y^n)) \leq \alpha\right) \geq 0.9 \right\},$$

the fewest number of samples for which we can estimate every $p \in \triangle_k$ up to α accuracy with probability at least 0.9. We will use the total variation distance, $d_{\mathrm{TV}}(\hat{p}, p) := \sup_{S \subseteq [k]} |\hat{p}(S) - p(S)| = \frac{1}{2} \|p - \hat{p}\|_1$ as the distance measure.

Sparse distribution estimation. Let $s \le k/100$[1] and

$$\triangle_{k,s} := \{p \in \triangle_k \colon \ \|p\|_0 \le s\}$$

be the distributions in $\triangle_k$ with support size at most s. Let

$$\mathrm{SC}(\alpha, \triangle_{k,s}, \mathcal{W}, d_{\mathrm{TV}}) := \arg\min_n \left\{ \min_{\hat{p}} \min_{W^n \in \mathcal{W}^n} \max_{p \in \triangle_{k,s}} \Pr\left(d_{\mathrm{TV}}(p, \hat{p}(Y^n))\right) \le \alpha) \ge 0.9 \right\}$$

denote the sample complexity of estimating s-sparse distributions to total variation distance α.

In this paper, we consider simultaneous message passing (SMP) protocols (non-interactive schemes) where all the messages from users are sent simultaneously (see Figure 1). SMP protocols are broadly classified as *private-coin* and *public-coin* protocols. In private-coin schemes the channels W_i are independent. In the more general public-coin schemes, the channels are chosen based on a function of a public randomness U observed by all the users and the server. Private-coin protocols are a strict subset of public-coin protocols. We refer the readers to Acharya et al. (2020b) for detailed definitions of these protocols.

1.2. Previous results and our contribution

Discrete distribution estimation under communication (Han et al., 2018b,a; Acharya et al., 2019b, 2020b) and LDP (Duchi et al., 2013; Erlingsson et al., 2014; Kairouz et al., 2016; Ye and Barg, 2018; Acharya et al., 2018; Acharya and Sun, 2019) constraints is well studied, and it is now known that for ℓ-bit channels $\mathcal{W}_\ell$, and ε-LDP channels $\mathcal{W}_\varepsilon$ (for $\varepsilon = O(1)$),

$$\mathrm{SC}(\alpha, \triangle_k, \mathcal{W}_\ell, d_{\mathrm{TV}}) = \Theta\left(\frac{k^2}{\alpha^2 \min\{k, 2^\ell\}}\right), \quad \mathrm{SC}(\alpha, \triangle_k, \mathcal{W}_\varepsilon, d_{\mathrm{TV}}) = \Theta\left(\frac{k^2}{\alpha^2 \varepsilon^2}\right). \tag{2}$$

Plugging $\ell = \log k$ in the first equation gives the centralized sample complexity of $\Theta\big(k/\alpha^2\big)$. Note that for $\varepsilon = O(1)$ and $\ell = O(1)$, the sample complexity increases by a factor k from the centralized setting.

We now present our results on estimating distributions in $\triangle_{k,s}$, the set of s sparse distributions in Δ_k. Our first result is a complete characterization of the sample complexity under ε-LDP up to constant factors.

Theorem 1 (Bounds for ε-LDP constraint) *For $\varepsilon = O(1)$ and $\alpha \in (0, 1)$,*

$$\mathrm{SC}(\alpha, \triangle_{k,s}, \mathcal{W}_\varepsilon, d_{\mathrm{TV}}) = \Theta\left(\frac{s^2 \max\{\log(k/s), 1\}}{\alpha^2 \varepsilon^2}\right).$$

Moreover, there exists a private-coin protocol with one-bit privatized messages that achieves the upper bound. The algorithm runs in nearly linear time in n and k.

1. Sparsity larger than $k/100$ gives same answers as the non-sparse case.

A few remarks are in order. This result shows that sparsity s is the effective domain size up to logarithmic factors. While an additional $\log k$ factor is slightly simpler to obtain, a more involved technique in sparse estimation based on a covering set argument is used to establish the upper bound with the optimal overhead factor of $\log(k/s)$. The lower bound is obtained by applying the recently developed chi-squared contraction techniques (Acharya et al., 2019b) to a new construction of distributions.

We now present our sample complexity bounds under communication constraints. Unlike LDP, our bounds are off by logarithmic factors in various parameter regimes. Resolving this gap and obtaining the tight bounds is an open question.

Theorem 2 (Bounds for ℓ-bit constraint) *For $\alpha \in (0,1)$, and channel family $\mathcal{W}_\ell$, the sample complexity for learning distributions in $\triangle_{k,s}$ satisfies,*

$$\mathrm{SC}(\alpha, \triangle_{k,s}, \mathcal{W}_\ell, d_{\mathrm{TV}}) = O\left(\frac{s^2 \max\{\log(k/s), 1\}}{\alpha^2 \min\{2^\ell, s\}}\right).$$

$$\mathrm{SC}(\alpha, \triangle_{k,s}, \mathcal{W}_\ell, d_{\mathrm{TV}}) = \Omega\left(\max\left\{\frac{s^2}{\alpha^2 \min\{2^\ell, s\}}, \frac{s^2 \max\{\log(k/s), 1\}}{\alpha 2^\ell}\right\}\right).$$

We briefly discuss the lower bound. Despite the gap, it is optimal for constant α. The first term is exactly the lower bound when the support is known. The second term dominates when $\log(k/s) > 1/\alpha$, i.e., the support is very sparse.

Organization. The remainder of the paper is organized as follows. We discuss related works in Section 1.3. We present algorithms and proofs for sparse estimation under LDP and communication constraints in Section 2 and Section 3 respectively.

1.3. Related work

Distribution estimation has a rich literature (see e.g., Barlow et al. (1972); Silverman (1986); Devroye and Györfi (1985); Devroye and Lugosi (2001), and references therein). There has been recent interest in distributed distribution estimation under communication and privacy constraints. For estimating discrete distributions under communication constraints, the optimal sample complexity is established in Han et al. (2018b); Acharya et al. (2019a); Barnes et al. (2020), and for local privacy constraints in Duchi et al. (2013); Erlingsson et al. (2014); Kairouz et al. (2016); Ye and Barg (2018); Bassily (2019); Acharya et al. (2018); Acharya and Sun (2019). Acharya et al. (2019b, 2020b) unify both constraints under the framework of distributed inference under local information constraints, where optimal bounds are obtained under both non-interactive and interactive protocols. Chen et al. (2020) considers the trade-off between privacy and communication constraints and provides optimal bounds in all parameter regimes. Murakami and Kawamoto (2019); Acharya et al. (2020a) study discrete distribution estimation under different privacy constraints on the symbols.

Kairouz et al. (2016) proposed an LDP distribution estimation algorithm for "open alphabets" which applies to the sparse setting. However, it requires public randomness and to the best of our knowledge, no theoretical guarantee of this method is provided for sparse distribution estimation. In contrast, the our algorithm is a private-coin algorithm that only requires one bit per user, and its sample complexity is proven to be optimal up to constant factors.

A closely related problem is heavy hitter detection under LDP constraints (Bassily and Smith, 2015; Bassily et al., 2020), where no distributional assumption on the data is made. A modification of their heavy hitters algorithms provides a sub-optimal $O(s^2 \log k/\alpha^2\varepsilon^2)$ sample complexity in terms of ℓ_1 error for ε-LDP distribution estimation .

Statistical inference with sparsity assumption has been studied extensively for decades. The closest to our works are the Gaussian sequence model and high dimensional linear regression (Donoho and Johnstone, 1994; Raskutti et al., 2011; Duchi and Wainwright, 2013). In these applications, it is assumed that the observations are linear transforms of the underlying parameter plus independent Gaussian noises on each dimension. In Section 2.1, it can be seen that using Algorithm 1, the histogram of observations can also be seen as a linear transform of the parameter of interest, however, with dependent noises on each dimension. We borrow ideas from these works in proving the upper bound. However, the lower bound part requires new proofs due to the dependency structure.

A few recent works study sparse estimation under information constraints. Duchi et al. (2013) and Wang and Xu (2019) consider the 1-sparse case and study mean estimation and linear regression under LDP constraints respectively. Duchi and Rogers (2019) provides lower bounds for sparse Gaussian mean estimation under LDP constraints via communication complexity. Barnes et al. (2020) considers estimating the mean of product Bernoulli distribution when the mean vector is sparse, which is different from the k-ary setting considered in this paper. Shamir (2014) considers the problem of detecting the biased coordinate of product Bernoulli distributions under communication constraints, which can be viewed as a 1-sparse detection problem. Zhang et al. (2013); Garg et al. (2014); Braverman et al. (2016); Han et al. (2018b) consider sparse Gaussian mean estimation under communication constraints (Garg et al. (2014); Braverman et al. (2016) consider interactive protocols with the goal of bounding the total amount of communication from all users). It was shown that under a fixed communication budget, the rate still scales linearly with the ambient dimension of the problem instead of the logarithmic dependence in the discrete case considered in this paper.

2. Sparse estimation under LDP constraints

In this section we will establish the sample complexity of sparse distribution estimation under LDP constraints. In Section 2.1, we analyze a private-coin algorithm where each user sends only one-bit messages detailed in Algorithm 1. The algorithm has two steps, listed below.

1. Using the private-coin Hadamard Response algorithm in Acharya and Sun (2019), players send one-bit messages.
2. The server projects a vector obtained from these messages onto $\triangle_{k,s}$ to obtain the final estimate.

We note that this algorithm is similar to that in Acharya and Sun (2019); Bassily (2019) where in the projection step they project onto $\triangle_k$ to estimate distributions without sparsity assumptions. While the algorithm is simple, to obtain the tight upper bounds, our analysis relies on a standard but involved covering-based techniques in sparse estimation. We also remark that a sample-optimal scheme can also be obtained using the popular RAPPOR mechanism (Erlingsson et al., 2014; Kairouz et al., 2016), which has higher communication overhead. We present this algorithm in the appendix for completeness.

In Section 2.2, we present a matching lower bound to prove the optimality of the aforementioned algorithm. The proof relies on applying the recently developed chi-squared contraction method (Acharya et al., 2019b) and a variant of Fano's inequality in Duchi and Wainwright (2013).

Algorithm 1 1-bit Hadamard Response with Projection

Input: $X_1, \ldots, X_n$ i.i.d. from $p \in \triangle_{k,s}$, the sparsity parameter s.
Output: $\widehat{p} \in \triangle_k$: an estimate of p.

1 Let $K = 2^{\lceil \log_2(k+1) \rceil}$ be the smallest power of 2 more than k.
2 For $y \in [K]$, let $B_y := \{x \in [K] : H_K(x, y) = 1\}$ be the rows where the yth column has 1.
3 Divide the n users into K sets $S_1, \ldots, S_K$ deterministically by assigning all $i \equiv j \mod K$ to S_j for $i \in [n]$.
4 $\forall j \in [K]$ and $\forall i \in S_j$, the distribution of the one-bit message Y_i is

$$\Pr(Y_i = 1) = \begin{cases} \frac{e^\varepsilon}{e^\varepsilon+1}, & X_i \in B_j, \\ \frac{1}{e^\varepsilon+1}, & \text{otherwise}, \end{cases} \tag{3}$$

namely if $H_K(X_i, j) = 1$, we send 1 with higher probability than 0.
5 Let $\hat{\mathbf{t}} := (\hat{t}_1, \ldots, \hat{t}_K)$ where $\forall j \in [K]$, $\hat{t}_j := \frac{1}{|S_j|} \sum_{i \in S_j} Y_i$ is the fraction of messages from S_j that are 1.
6 Compute intermediate estimates for

$$\tilde{p}_K := \frac{e^\varepsilon + 1}{K(e^\varepsilon - 1)} H_K(2\hat{\mathbf{t}} - \mathbf{1}_K).$$

7 Keep the first k elements of $\tilde{p}_K$, i.e., $\tilde{p} := \tilde{p}_K^{[k]}$ and project it onto $\triangle_{k,s}$.

$$\widehat{p} := \min_{p \in \triangle_{k,s}} \|\tilde{p} - p\|_2^2 .$$

2.1. Upper bounds under LDP constraints

We now establish the sample and time complexity of Algorithm 1.

Steps 1–6 of Algorithm 1 are identical to Acharya and Sun (2019), who showed a time complexity of $\tilde{O}(n+k)$. For the final step, where we project the vector $\tilde{p}_K$ on to $\triangle_{k,s}$, we can use (Kyrillidis et al., 2013, Algorithm 1), which runs in $\tilde{O}(k)$ time, proving the overall time complexity.

Algorithm 1 uses Hadamard matrices. For m that is a power of two, let H_m be the $m \times m$ Hadamard matrix with entries in $\{-1, 1\}$. The privacy guarantee of the algorithm follows from (3), which obeys (1). A key property we use is the following claim from Acharya and Sun (2019), which shows a relationship between underlying distribution and the message distributions.

Claim 1 (Acharya and Sun (2019)) *In (3), let $t_j := \Pr(Y_i = 1 \mid i \in S_j)$ for $j \in [K]$. Let $\mathbf{t} := (t_1, \ldots, t_K)$. Let p_K be the distribution over $[K]$ obtained by appending $K - k$ zeros to p. Then,*

$$p_K = \frac{(e^\varepsilon + 1)}{K(e^\varepsilon - 1)} H_K(2\mathbf{t} - \mathbf{1}_K). \tag{4}$$

By definiton of $\widehat{p}$, we have $\|\tilde{p}-\widehat{p}\|_2^2 \le \|\tilde{p}-p\|_2^2$. Hence

$$\|\tilde{p}-p\|_2^2 \ge \|\tilde{p}-\widehat{p}\|_2^2 = \|\tilde{p}-p\|_2^2 + \|p-\widehat{p}\|_2^2 + 2\langle \tilde{p}-p, p-\widehat{p}\rangle.$$

Rearranging the terms, we have

$$\|\widehat{p}-p\|_2^2 \le 2\langle \tilde{p}-p, \widehat{p}-p\rangle. \tag{5}$$

We bound the right hand side by analyzing the projection step (Step 7), and using Claim 1. The proof of the lemma is from standard covering number arguments from high dimensional sparse regression (Raskutti et al., 2011), and is provided in Appendix A.

Lemma 3

$$\langle \tilde{p}-p, \widehat{p}-p\rangle \le \frac{25(e^\varepsilon+1)}{(e^\varepsilon-1)} \frac{\sqrt{s\log(2k/s)}}{\sqrt{n}} \|\widehat{p}-p\|_2.$$

We can now prove the sample complexity bound as follows.

$$d_{\mathrm{TV}}(\widehat{p},p) = \frac{1}{2}\|\widehat{p}-p\|_1 \le \frac{1}{2}\sqrt{2s}\,\|\widehat{p}-p\|_2 \tag{6}$$

$$\le \frac{40s\sqrt{\log(2k/s)}}{\sqrt{n}} \frac{e^\varepsilon+1}{e^\varepsilon-1}, \tag{7}$$

where (6) applies Cauchy-Schwarz inequality on the $2s$-sparse vector $\widehat{p}-p$, and (7) is from plugging Lemma 3 in (5). Plugging in $d_{\mathrm{TV}}(\widehat{p},p)=\alpha$, and using $e^\varepsilon - 1 = O(\varepsilon)$ for $\varepsilon = O(1)$ gives us the desired sample complexity bound of $n = O\big(s^2 \max\{\log(k/s),1\}/\alpha^2\varepsilon^2\big)$.

2.2. Lower bound under LDP constraints

We now prove the sample complexity lower bound in Theorem 1 using the chi-squared contraction method in Acharya et al. (2019b) and an extension of Fano's method from Duchi and Wainwright (2013).

For simplicity of analysis we add 0 to the underlying domain and consider distributions over $[k]\cup\{0\}$. Let $\mathcal{Z}_{k,s} \subseteq \{0,1\}^k$ be all k-ary binary strings with s one's. Then, $|\mathcal{Z}_{k,s}| = \binom{k}{s}$. We will restrict to $\mathcal{P}_{k,s} := \{p_z : z \in \mathcal{Z}_{k,s}\}$, and p_z is described below for $z \in \mathcal{Z}_{k,s}$:

$$p_z(x) = \begin{cases} 1-8\alpha, & \text{for } x = 0, \\ \frac{8\alpha z_x}{s}, & \text{for } x = 1,\dots,k, \end{cases} \tag{8}$$

where z_x is the xth coordinate of z. Since s of the z_x's are one, $\sum_{x=1}^k p_z(x) = 8\alpha$ and p_z is a valid distribution.

Let $Z := (Z_1,\dots,Z_k)$ be a uniform random variable over $\mathcal{Z}_{k,s}$. Let $Y^n := (Y_1,\dots,Y_n)$ be the output of an ε-LDP scheme whose input are $X^n = (X_1,\dots,X_n)$, drawn i.i.d. from p_Z, and $\widehat{p}$ is such that $\Pr\left(d_{\mathrm{TV}}(p,\widehat{p}(Y^n)) \le \alpha\right) \ge 0.9$. In other words, we can estimate distributions in $\mathcal{P}_{k,s}$ to within α in total variation distance with probability at least 0.9. Let $\hat{Z} \in \mathcal{Z}_{k,s}$ be such that $p_{\hat{Z}}$ is the distribution in $\mathcal{P}_{k,s}$ closest to $\widehat{p}(Y^n)$ in d_{TV}. Then, we have

$$\frac{4\alpha}{s} d_{\mathrm{Ham}}(Z,\hat{Z}) = d_{\mathrm{TV}}(p_Z, p_{\hat{Z}}) \le d_{\mathrm{TV}}(\widehat{p}, p_{\hat{Z}}) + d_{\mathrm{TV}}(p_Z,\widehat{p}) \le 2d_{\mathrm{TV}}(p_Z,\widehat{p}).$$

Since $\Pr\left(d_{\mathrm{TV}}(p_Z, \widehat{p}(Y^n)) \le \alpha\right) \ge 0.9$, we have $\Pr\left(d_{\mathrm{Ham}}(Z, \hat{Z}) \le s/2\right) \ge 0.9$, which implies using the estimator $\hat{p}$, we can estimate the underlying Z to within Hamming distance $s/2$. We now state a form of Fano's inequality from Duchi and Wainwright (2013), adapted to our setting.

Lemma 4 (Corollary 1 Duchi and Wainwright (2013)) *Let $\mathcal{Z}_{k,s} \subseteq \{0,1\}^k$ and Z be uniformly distributed over $\mathcal{Z}_{k,s}$. For $t \ge 0$, define the maximum neighborhood size at radius t*

$$N_t^{max} := \max_{z \in \mathcal{Z}} \{|z' \in \mathcal{Z} : d_{\mathrm{Ham}}(z, z')| \le t\},$$

to be the maximum number of elements of $\mathcal{Z}_{k,s}$ in a Hamming ball of radius t. If $|\mathcal{Z}_{k,s}| \ge 2N_t^{max}$, then for any Markov chain $Z - Y^n - \hat{Z}$,

$$\Pr\left(d_{\mathrm{Ham}}(\hat{Z}, Z) > t\right) \ge 1 - \frac{I(Z; Y^n) + \log 2}{\log|\mathcal{Z}_{k,s}| - \log N_t^{max}}.$$

Substituting $t = s/2$, and using $\Pr\left(d_{\mathrm{Ham}}(Z, \hat{Z}) \le s/2\right) \ge 0.9$ with this lemma gives

$$\frac{I(Z; Y^n) + \log 2}{\log|\mathcal{Z}_{k,s}| - \log N_{s/2}^{max}} > 0.9. \tag{9}$$

Using chi-squared contraction bounds from Acharya et al. (2019b), we upper bound $I(Z; Y^n)$ in the next lemma.

Lemma 5 *Let Z be uniformly drawn from $\mathcal{Z}_{k,s}$ and Y^n be the outputs of n users,*

$$I(Z; Y^n) = O\left(\frac{n\alpha^2(e^\varepsilon - 1)^2}{s}\right).$$

Next. we lower bound $\log|\mathcal{Z}_{k,s}| - \log N_{s/2}^{max}$ using standard Gilbert-Varshamov type arguments in the following lemma.

Lemma 6 *Let $1 \le s \le k/100$, then*

$$\log|\mathcal{Z}_{k,s}| - \log N_{s/2}^{max} \ge \frac{s}{8} \log\left(\frac{k}{s}\right).$$

Plugging these two bounds in (9) gives the tight sample complexity lower bound for sparse distribution estimation under ε-local differential privacy.

We now prove these lemmas.

Proof [Proof of Lemma 5] For a distribution q over $[k] \cup \{0\}$ and a channel W, let q^W denote output distribution of Y when $X \sim q$ and $Y = W(X)$. Then,

$$q^W(y) = \sum_{x \in [k] \cup \{0\}} W(y \mid x) q(x). \tag{10}$$

Let p_0 be the average of all distributions in $\mathcal{P}_{k,s}$. Then $p_0(0) = 1 - 8\alpha$, and $p_0(x) = 8\alpha/k$ for $x = 1, \ldots, k$. We will use chi square contraction bound in Acharya et al. (2019b) to bound the maximum value of $I(Z; Y^n)$ in terms of the χ^2-divergence between the output distributions induced by distributions in $\mathcal{P}_{k,s}$ and by p_0 and as follows:

$$I(Z; Y^n) \le n \cdot \max_{W \in \mathcal{W}_\varepsilon} \mathbb{E}_Z \left[\chi^2\left(p_Z^W, p_0^W\right)\right] \tag{11}$$

$$= n \cdot \max_{W \in \mathcal{W}_\varepsilon} \sum_y \frac{\mathbb{E}_Z\left[\left(\sum_{x=1}^k (p_Z(x) - p_0(x)) W(y \mid x)\right)^2\right]}{\mathbb{E}_{X \sim p_0}\left[W(y \mid X)\right]}, \tag{12}$$

where (11) is from the chi-squared contraction bound, and (12) is by using (10) in the definition of χ^2-divergence[2].

For an ε-LDP channel $W \in \mathcal{W}_\varepsilon$, let $W^y_{\min} := \min_x W(y \mid x)$. By (1), we have $W(y \mid x) = W^y_{\min} + \eta^y_x \cdot W^y_{\min}$ for some $0 \le \eta^y_x \le e^\varepsilon - 1$. Furthermore, for $z \in \mathcal{Z}_{k,s}$, by the definition of p_z, $p_z(x) - p_0(x) = 8\alpha\left(\frac{z_x}{s} - \frac{1}{k}\right)$, and $\sum_x z_x = s$, thus giving

$$\sum_x (p_z(x) - p_0(x)) W(y \mid x) = 8\alpha \sum_x \left(\frac{z_x}{s} - \frac{1}{k}\right)\left(W^y_{\min} + W^y_{\min}\eta^y_x\right) = 8\alpha W^y_{\min} \cdot \sum_x \left(\frac{z_x}{s} - \frac{1}{k}\right)\eta^y_x.$$

Since Z is uniformly distributed over $\mathcal{Z}_{k,s}$, elementary computations show that $\mathbb{E}\left[Z_x\right] = \mathbb{E}\left[Z_x^2\right] = s/k$, and for $x_1 \ne x_2 \in [k]$, $\mathbb{E}\left[Z_{x_1} Z_{x_2}\right] = \binom{k-2}{s-2} / \binom{k}{s} = \frac{s(s-1)}{k(k-1)}$.

Therefore,

$$\begin{aligned}
&\mathbb{E}_Z\left[\left(\sum_x (p_Z(x) - p_0(x)) W(y \mid x)\right)^2\right] \\
&= 64\alpha^2 (W^y_{\min})^2 \cdot \left(\sum_{x_1, x_2} \mathbb{E}_Z\left[\frac{1}{k^2} - \frac{Z_{x_1} + Z_{x_2}}{sk} + \frac{Z_{x_1} Z_{x_2}}{s^2}\right] \eta^y_{x_1}\eta^y_{x_2}\right) \\
&= 64\alpha^2 (W^y_{\min})^2 \left(\sum_x \left[\frac{1}{sk} - \frac{1}{k^2}\right](\eta^y_x)^2 + \sum_{x_1 \ne x_2} \mathbb{E}_Z\left[-\frac{1}{k^2} + \frac{s-1}{sk(k-1)}\right]\eta^y_{x_1}\eta^y_{x_2}\right) \\
&\le 64\alpha^2 (W^y_{\min})^2 \left(\frac{(\max_x \eta^y_x)^2}{s}\right),
\end{aligned}$$

and

$$\frac{\mathbb{E}_Z\left[\left(\sum_x (p_Z(x) - p_0(x)) W(y \mid x)\right)^2\right]}{\mathbb{E}_{X \sim p_0}\left[W(y|X)\right]} \le 64\alpha^2 \left(\frac{(\max_x \eta^y_x)^2}{s}\right) \cdot W^y_{\min}.$$

Using $\sum_y W^y_{\min} \le 1$, and $\eta^y_x \le e^\varepsilon - 1$, we obtain

$$\mathbb{E}_Z\left[\chi^2(p_Z^W, p_0^W)\right] = O\left(\frac{\alpha^2 (e^\varepsilon - 1)^2}{s}\right).$$

2. $\chi^2(p, q) := \sum_x (p(x) - q(x))^2 / q(x)$.

Combining with (12), this completes the proof. ■

Proof [Proof of Lemma 6] Let $z \in \mathcal{Z}_{k,s}$. A vector in $\mathcal{Z}_{k,s}$ that is at most $s/2$ away from z in Hamming distance can be obtained as follows: Fix $s/2$ coordinates in z that are 1, and from the remaining $k - s/2$ coordinates, choose $s/2$ coordinates and make them 1. All other coordinates are set to zero. Therefore,

$$N_{s/2}^{max} \le \binom{s}{s/2}\binom{k-s/2}{s/2}.$$

Recall that $|\mathcal{Z}_{k,s}| = \binom{k}{s}$. Using Stirling's approximation for binomial coefficients[3], we get

$$\log \frac{|\mathcal{Z}|}{N_{s/2}^{max}} \ge \log \frac{\binom{k}{s}}{\binom{s}{s/2}\binom{k-s/2}{s/2}} \ge \log \frac{\left(\frac{k}{s}\right)^s}{(2e)^{s/2}\left(\frac{(2k-s)e}{s}\right)^{s/2}} \ge \log \frac{\left(\frac{k}{s}\right)^s}{(2e)^{s/2}\left(\frac{(2k)e}{s}\right)^{s/2}} = \frac{s}{2}\log\left(\frac{k}{4e^2 s}\right),$$

which is at least $\frac{s}{8}\log\frac{k}{s}$ when $s \le k/100$. ■

3. Sparse estimation under communication constraints

We now prove guarantees of Theorem 2 and establish upper and lower bounds for communication constrained sparse discrete distribution estimation. In Section 3.1, we propose an algorithm that requires public randomness with sample complexity given in Theorem 2. Designing a private-coin protocol for estimation is an open question. In Section 3.2 we establish the lower bounds.

3.1. Upper bounds under communication constraints

Note that the sample complexity upper bound in Theorem 2 has $\min\left\{2^\ell, s\right\}$, which equals s when $\ell \ge \log s$. We therefore only consider $\ell \le \log s$ since if $\ell > \log s$, we can just use $\log s$ bits and get the same bound.

Our first step is to use public randomness to design hash functions at the users. A randomized mapping $h : [k] \to [2^\ell]$ is a *random hash function* if $\forall x \in [k], y \in [2^\ell]$,

$$\Pr\left(h(x) = y\right) = \frac{1}{2^\ell}.$$

The scheme. Let $h_1, \ldots, h_n$ be n independent hash functions, available at the users and at the server. User i's ℓ bit output is $Y_i = h_i(X_i) \in [2^\ell]$. The probability of $x \in [k]$ being in the preimage of user i's message Y_i is,

$$\Pr\left(Y_i = h_i(x)\right) = p(x) + (1 - p(x))\frac{1}{2^\ell} = p(x)\left(1 - \frac{1}{2^\ell}\right) + \frac{1}{2^\ell} =: b(x). \tag{13}$$

The estimator. Upon receiving messages $Y_1, \ldots, Y_n$, the estimator is as follows,

1. The first $n/2$ messages are used to obtain a set $T \subseteq [k]$ with $|T| = O(s)$ such that with high probability $p(T) > 1 - \alpha/2$.
2. With the remaining messages we estimate $p(x)$ for $x \in T$.

3. For $1 \le s \le k$, we have $\left(\frac{k}{s}\right)^s \le \binom{k}{s} \le \left(\frac{ke}{s}\right)^s$.

We now describe and analyze the two steps.

Step 1. Let $Y_1, \ldots, Y_{n/2}$ be the first $n/2$ messages, where $Y_i = h_i(X_i)$. For $x = 1, \ldots, k$, let

$$M(x) := |\{i : h_i(x) = Y_i, 1 \le i \le n/2\}|$$

be the number of these messages in the first half whose preimage x belongs to. Let $T \subseteq [k]$ be the set of symbols with the largest $|T| = 2s$ values of $M(x)$'s. If $p(x)$ is large we expect $M(x)$ to be large. In particular, we show that for sufficiently large n, the probability of symbols not in T is small.

Lemma 7 *There is a constant C_1 such that for $n = C_1 \cdot s^2 \log(k/s)/(\alpha^2 \min\{2^\ell, s\})$ with probability at least 0.95, $p(T) := \sum_{x \in T} p(x) \ge 1 - \alpha/2$.*

Step 2. For $x = 1, \ldots, k$, let $N(x) := |\{i : h_i(x) = Y_i, n/2 < i \le n\}|$ be the number of messages in the second half such that x belongs to the preimage of Y_i. Our final estimator is given by

$$\widehat{p}(x) = \begin{cases} \frac{(2^\ell N(x)/(n/2)) - 1}{2^\ell - 1}, & \text{if } x \in T \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

The following lemma shows that $\widehat{p}$ converges to p over T.

Lemma 8 *There is a constant C_2 such that for $n = C_2 \cdot s^2/(\alpha^2 \min\{2^\ell, s\})$ with probability at least 0.95,*

$$\sum_{x \in T} |\widehat{p}(x) - p(x)| \le \frac{\alpha}{2}.$$

Combining Lemma 7 and Lemma 8, by the union bound we get that with n samples, with probability at least 0.9, $\|\widehat{p} - p\|_1 \le \alpha$, establishing upper bound of Theorem 2.

3.2. Lower bounds under communication constraints

We now prove the sample complexity lower bound in Theorem 2. The first term of $\Omega(\frac{s^2}{\alpha^2 \min\{2^\ell, s\}})$ follows from (2) and holds even with the knowledge of the support S.

We prove the second term by considering the construction $\mathcal{P}_{k,s}$ as in Section 2.2, and bound (9). The lower bound on $\log(|\mathcal{Z}|/N_t^{max})$ is the same as from Lemma 6. Analogous to Lemma 5, we will now bound the mutual information $I(Z; Y^n)$ by $O(n\alpha 2^\ell/s)$ as follows. As in (12), we have the following bound

$$I(Z; Y^n) \le n \cdot \max_{W \in \mathcal{W}_\ell} \sum_y \frac{\mathbb{E}_Z \left[\left(\sum_{x=1}^k (p_Z(x) - p_0(x)) W(y \mid x) \right)^2 \right]}{\mathbb{E}_{X \sim p_0} \left[W(y \mid X) \right]}, \tag{15}$$

and we now bound it for each $y \in [2^\ell]$. Similar to the expansion in proving the LDP lower bounds, we have

$$
\begin{aligned}
&\mathbb{E}_Z\left[\left(\sum_x (p_Z(x) - p_0(x))W(y \mid x)\right)^2\right] \\
&= 64\alpha^2 \mathbb{E}_Z\left[\sum_{x_1,x_2}\left(\frac{1}{k^2} - \frac{Z_{x_1}+Z_{x_2}}{sk} + \frac{Z_{x_1}Z_{x_2}}{s^2}\right)W(y \mid x_1)W(y \mid x_2)\right] \\
&= 64\alpha^2\sum_{x=1}^k\left(\frac{1}{sk} - \frac{1}{k^2}\right)W(y \mid x)^2 + 64\alpha^2\sum_{x_1\neq x_2}\left(-\frac{1}{k^2} + \frac{s-1}{sk(k-1)}\right)W(y \mid x_1)W(y \mid x_2) \\
&\leq 64\alpha^2\sum_{x=1}^k\left(\frac{1}{sk} - \frac{1}{k^2}\right)W(y \mid x)^2.
\end{aligned}
$$

Note that $\mathbb{E}_{X\sim p_0}\left[W(y \mid X)\right] = (1-8\alpha)W(y \mid 0) + \sum_{x=1}^k \frac{8\alpha}{k}W(y \mid x)$. Hence,

$$
\begin{aligned}
\mathbb{E}_Z\left[\chi^2(p_Z^W, p_0^W)\right] &\leq 64\alpha^2\sum_y \frac{\left(\sum_{x=1}^k\left(\frac{1}{sk} - \frac{1}{k^2}\right)W(y \mid x)^2\right)}{(1-8\alpha)W(y \mid 0) + \sum_{x=1}^k\frac{8\alpha}{k}W(y \mid x)} \\
&\leq 64\alpha^2\left(\frac{1}{sk} - \frac{1}{k^2}\right)\sum_y\frac{\left(\sum_{x=1}^k W(y \mid x)^2\right)}{\sum_{x=1}^k\frac{8\alpha}{k}W(y \mid x)} \\
&\leq \frac{8\alpha}{s}\sum_y\frac{\sum_{x=1}^k W(y \mid x)^2}{\sum_{x=1}^k W(y \mid x)} \\
&\leq \frac{8\alpha}{s}2^\ell,
\end{aligned}
$$

where we used that $W(y|x)^2 \leq W(y|x)$, proving the lower bound.

4. Experiments

To verify our bounds, we evaluate our algorithms on sythetic datasets. We fix the support size $k = 5000$ and draw the data from distributions uniform over a subset with size s which takes values in $2^i, i = 1, \ldots, 12$.

For LDP, we set $\varepsilon = 0.5, 0.9$ and draw $n = 3 \times 10^6$ samples. For communication, we set $\ell = 1, \ldots, 7$ and $n = 400000$. The estimation errors in d_{TV} with respect to sparsity s are shown in Figure 2. In both experiments, we observe significant increase in accuracy when the sparsity decreases. And in both experiments, we observe that under the same sparsity, larger ε (LDP) or ℓ (communication) leads to a better utility, which is consistent with our theoretical analysis.

Moreover, for LDP we compare the Hadamard Response algorithm with sparse projection (exact procedures in Algorithm 1) and regular projection (projecting onto $\triangle_k$ instead of $\triangle_{k,s}$ in Step 7 of Algorithm 1). The results show that sparse projection, which needs the knowledge of s, leads to a significant improvement in utility. Straightforward modifications of our proof shows that projecting onto $\triangle_k$, which doesn't need to know s in advance, will lead to a risk bound of $O(s\sqrt{\log k}/\sqrt{n})$

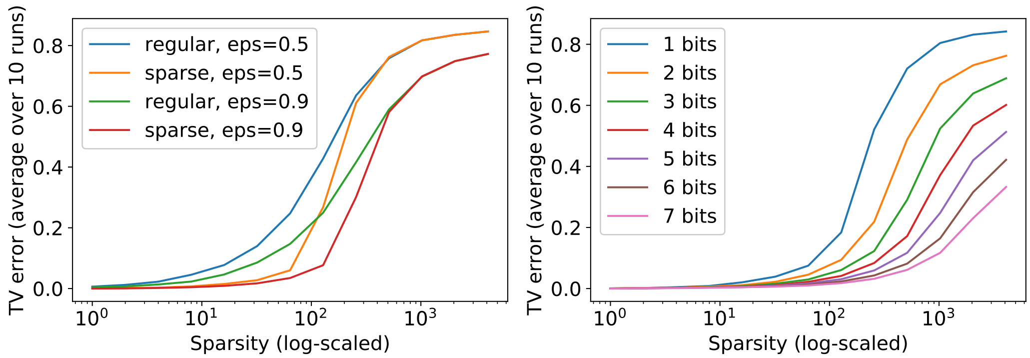

Figure 2: Estimation errors of the proposed algorithms under different support sizes. **Left**: LDP (comparing regular and sparse projections); **Right**: communication constraints.

instead of $O(s\sqrt{\log(k/s)}/\sqrt{n})$ in Theorem 1. Whether this gap is inevitable is an interesting direction to explore.

References

Jayadev Acharya and Ziteng Sun. Communication complexity in locally private distribution estimation and heavy hitters. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. In *Advances in Neural Information Processing Systems*, pages 6879–6891, 2018.

Jayadev Acharya, Clement Canonne, and Himanshu Tyagi. Communication-constrained inference and the role of shared randomness. volume 97 of *Proceedings of Machine Learning Research*, pages 30–39, Long Beach, California, USA, 09–15 Jun 2019a. PMLR. URL http://proceedings.mlr.press/v97/acharya19a.html.

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints: Lower bounds from chi-square contraction. *Proceedings of Machine Learning Research vol*, 99:1–15, 2019b.

Jayadev Acharya, Kaylee Bonawitz, Peter Kairouz, Daniel Ramage, and Ziteng Sun. Context-aware local differential privacy. In *Proceedings of the 37th International Conference on Machine Learning*, 2020a.

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

Richard E. Barlow, David J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.

L. P. Barnes, W. N. Chen, and A. Özgür. Fisher information under local differential privacy. *IEEE Journal on Selected Areas in Information Theory*, 1(3):645–659, 2020. doi: 10.1109/JSAIT.2020.3039461.

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. URL http://jmlr.org/papers/v21/19-737.html.

Raef Bassily. Linear queries estimation with local differential privacy. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 721–729, 2019.

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

Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Thakurta. Practical locally private heavy hitters. *Journal of Machine Learning Research*, 21(16):1–42, 2020. URL http://jmlr.org/papers/v21/18-786.html.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. 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*, pages 1011–1020. ACM, 2016.

Wei-Ning Chen, Peter Kairouz, and Ayfer Ozgur. Breaking the communication-privacy-accuracy trilemma. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3312–3324. Curran Associates, Inc., 2020.

Luc Devroye and László Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, 1985.

Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

David L Donoho and Iain M Johnstone. Minimax risk overl p-balls forl p-error. *Probability Theory and Related Fields*, 99(2):277–303, 1994.

John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. volume 99 of *Proceedings of Machine Learning Research*, pages 1161–1191, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/duchi19a.html.

John C Duchi and Martin J Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Proceedings of the 54st Annual IEEE Symposium on Foundations of Computer Science*, FOCS ’13, pages 429–438. IEEE, 2013.

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

Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2014.

Yanjun Han, Pritam Mukherjee, Ayfer Özgür, 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, 2018a.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. volume 75 of *Proceedings of Machine Learning Research*, pages 3163–3188. PMLR, 2018b.

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2436–2444, 2016.

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

Anastasios Kyrillidis, Stephen Becker, Volkan Cevher, and Christoph Koch. Sparse projections onto the simplex. In *International Conference on Machine Learning*, pages 235–243, 2013.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

Takao Murakami and Yusuke Kawamoto. Utility-optimized local differential privacy mechanisms for distribution estimation. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1877–1894, 2019.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10): 6976–6994, 2011.

Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems*, pages 163–171, 2014.

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

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. ISBN 9781108498029. URL https://books.google.com/books?id=8C8nuQEACAAJ.

Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *International Conference on Machine Learning*, pages 6628–6637, 2019.

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

Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013.

Appendix A. Proof of Lemma 3

We will use the following bound on the covering number of s-sparse vectors.

Claim 2 (Raskutti et al. (2011)) *Let $S(r,2s)=\{\xi\in\mathbb{R}^k : \|\xi\|_2\le r, \|\xi\|_0\le 2s\}$. There exists a ρr-covering (in ℓ_2) $\mathcal{C}_r\subset S(r,2s)$ of $S(r,2s)$ with size*

$$|\mathcal{C}_r| = N(S(r,2s),\rho r)\le \binom{k}{2s}\left(\frac{1}{\rho}\right)^{2s}.$$

The bound follows from the fact that $\binom{k}{2s}$ $2s$-dimensional subspaces are sufficient to cover $S(r,2s)$, and each subspace can be covered with $(1/\rho)^{2s}$ ℓ_2 balls of radius ρr.

Let $\mathcal{C}_1$ be a ρ-covering of $S(1,2s)$ with size $N(S(1,2r),\rho)$. Let $r=\|\widehat{p}-p\|_2$, then $\widehat{p}-p\in S(r,2s)$, and one can obtain a covering $\mathcal{C}_r$ of $S(r,2s)$ by multiplying each vector in $\mathcal{C}_1$ by r. Since $\widehat{p}-p$ is $2s$-sparse, let ξ^* be the closest point in $\mathcal{C}_r$ to $\widehat{p}-p$. Then we can bound the right hand side of (5).

$$\begin{aligned}\langle\tilde{p}-p,\widehat{p}-p\rangle &\le |\langle\tilde{p}-p,\xi^*\rangle| + |\langle\tilde{p}-p,\widehat{p}-p-\xi^*\rangle|\\ &\le \max_{\xi\in\mathcal{C}_r}|\langle\tilde{p}-p,\xi\rangle| + \rho r\,\|\tilde{p}-p\|_2\\ &= r\max_{\xi\in\mathcal{C}_r}|\langle\tilde{p}-p,\xi/r\rangle| + \rho r\,\|\tilde{p}-p\|_2\\ &= r\max_{\xi\in\mathcal{C}_1}|\langle\tilde{p}-p,\xi\rangle| + \rho r\,\|\tilde{p}-p\|_2\,.\end{aligned}$$

$\tilde{p}-p$ is the first k entries of $\tilde{p}_K-p_K$, and by Claim 1, we have

$$\tilde{p}_K-p_K=\frac{2(e^\varepsilon+1)}{K(e^\varepsilon-1)}H_K(\hat{\mathbf{t}}-\mathbf{t}).$$

Note that in Algorithm 1, $\forall j\in[K]$, $\hat{t}_j$ is the average of $|S_j|\ge K/(2n)$ i.i.d. Bernoulli random variables with the mean satisfying $\mathbb{E}\left[\hat{\mathbf{t}}\right]=\mathbf{t}$. Hence $\forall j\in[K]$, $\hat{t}_j-t_j$ is a zero-mean sub-Gaussian random variable with variance proxy at most $\frac{K}{2n}$. Moreover, $\hat{t}_j$'s are independent since S_j's are disjoint. Note for all $\forall\xi\in\mathcal{C}_1$,

$$\langle\tilde{p}_K-p_K,\xi\rangle=\frac{2(e^\varepsilon+1)}{K(e^\varepsilon-1)}\xi^T H_K(\hat{\mathbf{t}}-\mathbf{t}),$$

which are linear combinations of $(\hat{t}_j - t_j)$'s. Since $\tilde{p} - p$ is the first k entries of $\tilde{p}_K - p_K$, we have $\forall \xi \in \mathcal{C}_1$, $\langle \tilde{p} - p, \xi \rangle$ is also sub-Gaussian (see Corollary 1.7 in Rigollet and Hütter (2015)) with variance proxy at most

$$\begin{aligned} \frac{K}{2n} \left\| \frac{2(e^\varepsilon + 1)}{K(e^\varepsilon - 1)} \xi^T H_K \right\|_2^2 &= \frac{2}{nK} \left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1} \right)^2 \left\| \xi^T H_K \right\|_2^2 \\ &= \frac{2}{n} \left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1} \right)^2 \|\xi\|_2^2 \qquad (16) \\ &\le \frac{2}{n} \left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1} \right)^2 =: \sigma^2, \end{aligned}$$

where (16) follows since $H_m^T H_m = m\mathbb{I}_k$ by the orthogonality of Hadamard matrices.

Therefore using maximal inequalties of sub-Gaussian random variables (Rigollet and Hütter, 2015, Theorem 1.14), with probability at least $19/20$, we have

$$\max_{e \in \mathcal{C}_1} |\langle \tilde{p} - p, \xi \rangle| < \sigma \sqrt{14 \log |\mathcal{C}_1|}. \qquad (17)$$

By the utility guarantee of Hadamard Response (Acharya and Sun, 2019), with probability at least $19/20$,

$$\|\tilde{p} - p\|_2^2 \le 20 \mathbb{E} \left[\|\tilde{p} - p\|_2^2 \right] \le \frac{40k(e^\varepsilon + 1)^2}{n(e^\varepsilon - 1)^2} \implies \|\tilde{p} - p\|_2 \le 7 \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \sqrt{\frac{k}{n}}. \qquad (18)$$

By union bound, conditioned on (17) and (18), which happens with probability at least $9/10$,

$$\begin{aligned} \langle \tilde{p} - p, \widehat{p} - p \rangle &\le \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \frac{r}{\sqrt{n}} \sqrt{56 \log |\mathcal{C}_1|} + 7\rho r \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \sqrt{\frac{k}{n}} \\ &\le \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \frac{r}{\sqrt{n}} \sqrt{112 s \log \left(\frac{ek}{2s} \right) + 112 s \log \frac{1}{\rho}} + 7\rho r \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \sqrt{\frac{k}{n}}. \end{aligned}$$

Taking $\rho = \sqrt{\frac{s}{k}}$,

$$\langle \tilde{p} - p, \widehat{p} - p \rangle \le 25 \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \frac{r\sqrt{s \log(2k/s)}}{\sqrt{n}} = 25 \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \frac{\sqrt{s \log(2k/s)}}{\sqrt{n}} \|\widehat{p} - p\|_2 ,$$

concluding the proof.

Appendix B. Proof of Lemma 7 and Lemma 8

Proof [Proof of Lemma 7] $M(x)$ is distributed as a Binomial $Bin(n/2, b(x))$. Hence, $\mathbb{E}[M(x)] = n \cdot b(x)/2$. By (13) and (14) we have $\mathbb{E}[\widehat{p}(x)] = p(x)$. Using the variance formula of Binomials, we know $\text{Var}(M(x)) = n \cdot b(x) \cdot (1 - b(x))/2 \le n \cdot b(x)/2$.

Set $\gamma = 1 - 1/2^\ell$ and $\beta = 1/2^\ell$, then $b(x) = \gamma \cdot p(x) + \beta$. We will use the following multiplicative Chernoff bound.

Lemma 9 (Multiplicative Chernoff bound (Mitzenmacher and Upfal, 2017)) *Let $Y_1, \ldots, Y_n$ be independent random variables with $Y_i \in 0, 1$, and $Y = Y_1 + \cdots + Y_n$, and $\mu = \mathbb{E}[Y]$. Then for $\tau > 0$,*

$$\Pr(Y \geq (1+\tau)\mu) \leq e^{-\frac{\tau^2 \mu}{2+\tau}}, \quad \Pr(Y \geq (1-\tau)\mu) \leq e^{-\frac{\tau^2\mu}{2}}.$$

Let $S := \{x : p(x) > 0\}$. Therefore, for $x \in [k] \setminus S$, $p(x) = 0$. By Lemma 9,

$$\Pr\left(M(x) \geq \frac{n}{2}\beta + \sqrt{3n\beta \log \frac{k}{s}}\right) \leq \left(\frac{s}{k}\right)^2.$$

Let E be the event that at most s symbols in $[k] \setminus S$ appear at least $M^* := \frac{n}{2}\beta + \sqrt{3n\beta \log \frac{k}{s}}$ times. By Markov's inequality,

$$\Pr(E^c) = \Pr\left(\left|x \in [k] \setminus S : M(x) \geq \frac{n}{2}\beta + \sqrt{3n\beta \log \frac{k}{s}}\right| > s\right) \leq \frac{s}{k} \leq \frac{1}{100}. \tag{19}$$

We condition on E in the remainder of the proof. Note that it suffices to show

$$\mathbb{E}[p(T^c) \mid E] := \sum_x p(x) \Pr(x \text{ not selected} \mid E) \leq \frac{\alpha}{50}, \tag{20}$$

since if (20) is true, by Markov inequality,

$$\Pr\left(p(T^c) > \frac{\alpha}{2} \;\middle|\; E\right) \leq \frac{1}{25}.$$

which, combined with (19), implies Lemma 7.

Next we prove (20). Conditioned on E, a symbol x is not selected after the first stage only if it appears at most M^* times, which implies $\Pr(x \text{ not selected} \mid E) \leq \Pr(M(x) \leq M^* \mid E)$. Moreover, since $\forall x_1 \in S, x_2 \notin S$, $M(x_1)$ and $M(x_2)$ are independent, we have $M(x_1)$ is independent of event E. Thus:

$$\mathbb{E}[p(T^c) \mid E] = \sum_x p(x) \Pr(x \text{ not selected} \mid E) \leq \sum_x p(x) \Pr(M(x) \leq M^*).$$

Next we divide symbols into three sets based on their probability mass: $A = \{x \in [k] : p(x) \leq \frac{\alpha}{60s}\}$, $B = \{x \in [k] : \frac{\alpha}{60s} < p(x) \leq \beta/\gamma\}$ and $C = \{x \in [k] : p(x) > \beta/\gamma\}$. For set A, we have:

$$\sum_{x \in A} p(x) \Pr(M(x) \leq M^*) \leq \sum_{x \in A} p(x) \leq \frac{\alpha}{60}. \tag{21}$$

Next we bound the sum over set B and C. $\forall x \in B \cup C, p(x) > \alpha/60s$. In the rest of the proof, we set the constant $C_1 = 700000$. For $n = C_1 \cdot s^2 \log(k/s)/(\alpha^2 \min\{2^\ell, s\})$,

$$\frac{n}{4}\gamma p(x) - \sqrt{3n\beta \log \frac{k}{s}} \geq 0.$$

Hence,

$$\mathbb{E}\left[M(x) - M^*\right] = \frac{n}{2}\gamma p(x) - \sqrt{3n\beta \log\frac{k}{s}} \geq \frac{n\gamma p(x)}{4},$$

Using Lemma 9,

$$\begin{aligned}\Pr\left(M(x) \leq M^*\right) &= \Pr\left(M(x) \leq \mathbb{E}\left[M(x)\right] - \left(\mathbb{E}\left[M(x)\right] - M^*\right)\right) \\ &\leq \exp\left(-\left(\frac{\gamma p(x)/2}{\gamma p(x) + \beta}\right)^2 \frac{n}{2}(\gamma p(x) + \beta)\right) \\ &= \exp\left(-\frac{\gamma^2 p(x)^2 n}{8(\gamma p(x) + \beta)}\right)\end{aligned}$$

If $x \in C$, i.e., $p(x) > \beta/\gamma$, we have

$$\Pr\left(M(x) \leq M^*\right) \leq \exp\left(-\frac{\gamma^2 p(x)^2 n}{16\gamma p(x)}\right) = \exp\left(-\frac{\gamma p(x) n}{16}\right) \leq \frac{16}{\gamma p(x) n} \leq \frac{16}{n\beta} \leq \frac{\alpha}{1000},$$

where we use $n\beta > C_1 s^2 \log(k/s)/(\alpha^2 2^{2\ell}) \geq C_1/\alpha^2$ when $s \geq 2^\ell$. This implies

$$\sum_{x \in C} p(x) \Pr\left(M(x) < M^*\right) \leq \sum_{x \in A} p(x)\frac{\alpha}{1000} \leq \frac{\alpha}{1000}. \tag{22}$$

When $x \in B$, i.e, $p(x) \leq \beta/\gamma$,

$$\Pr\left(M(x) \leq M^*\right) \leq \exp\left(-\frac{\gamma^2 p(x)^2 n}{16\beta}\right).$$

Now let $p(x) = (1 + \zeta_x)\alpha/60s$ where $\zeta_x > 0$, we have

$$\frac{\gamma^2 p(x)^2 n}{16\beta} \geq 2(1 + \zeta_x)^2 \log\frac{k}{s}.$$

$$\sum_{x \in C} p(x) \Pr\left(M(x) \leq M^*\right) \leq \frac{\alpha}{60s}\sum_{x \in C}(1 + \zeta_x)\exp\left(-2(1 + \zeta_x)^2 \log\frac{k}{s}\right) \leq \frac{\alpha}{500}. \tag{23}$$

Combining (21), (22) and (23), we get (20), and thus proving the lemma. ■

Proof [Proof of Lemma 8] Note that $N(x)$ and $M(x)$ are identically distributed, and therefore, $M(x)$ is distributed $Bin(n/2, b(x))$, and $\mathbb{E}\left[N(x)\right] = n \cdot b(x)/2$, and $\mathrm{Var}\left(N(x)\right) \leq n \cdot b(x)/2$.

$$\mathbb{E}\left[(\widehat{p}(x) - p(x))^2\right] = \left(\frac{2 \cdot 2^\ell}{n(2^\ell - 1)}\right)^2 \cdot \mathrm{Var}\left(N(x)\right) \leq \frac{2}{n}\left(\frac{2^\ell}{2^\ell - 1}\right)^2 b(x). \tag{24}$$

Now, note that $\sum_{x \in T} b(x) = p(T)(1 - 1/2^\ell) + |T|/2^\ell$, and therefore,

$$\mathbb{E}\left[\left\|\widehat{p}^T - p^T\right\|_2^2\right] = \sum_{x \in T}\mathbb{E}\left[(\widehat{p}(x) - p(x))^2\right] \leq \frac{2}{n}\left(\frac{2^\ell}{2^\ell - 1}\right)^2 \sum_{x \in T} b(x) \leq \frac{2(|T| + 2^\ell)2^\ell}{n(2^\ell - 1)^2}.$$

Using Jensen's inequality and Cauchy-Schwarz,

$$\mathbb{E}\left[\left\|\widehat{p}^T - p^T\right\|_1\right] \leq \sqrt{\mathbb{E}\left[\|\widehat{p}^T - p^T\|_1^2\right]} \leq \sqrt{|T| \cdot \mathbb{E}\left[\|\widehat{p}^T - p^T\|_2^2\right]} \leq \sqrt{\frac{4s2^\ell(2^\ell + 2s)}{n(2^\ell - 1)^2}}.$$

Setting $C_2 = 6400$, the lemma follows by Markov's inequality. ■

Appendix C. An LDP estimation scheme using RAPPOR

We first describe the high level idea of the algorithm for LDP estimation. All users send their privatized data using RAPPOR (Erlingsson et al., 2014; Kairouz et al., 2016). As in Section 3.1, we use the first half of privatized samples to estimate a subset $T \subseteq [k]$ with size $O(s)$ which contains most of the probability densities; we then use the remaining samples to estimate the distribution only on this set T. Details are described in Algorithm 2.

Algorithm 2 Sparse estimation using RAPPOR

Input: n i.i.d. samples from unknown s-sparse p.

Each user randomizes its sample using RAPPOR: Each sample X_i is first mapped to a one-hot vector $Z_i \in \{0,1\}^k$ which has a 1 at the X_i'th coordinate and 0's elsewhere. Then each bit is flipped independently with probability $1/(e^{\varepsilon/2}+1)$ to obtain $Y_i \in \{0,1\}^k$

Compute $M := [M(1), \ldots, M(k)] = \sum_{i=1}^{n/2} Y_i$ using the first $\frac{n}{2}$ samples.

Construct the set $T \subseteq [n]$ by keeping the $2s$ symbols with highest $M(x)$'s.

Obtain $\widehat{p}$: estimate the distribution over T using the remaining $\frac{n}{2}$ samples.

We note that

$$\mathbb{E}\left[M(x)\right] = \frac{n}{2}\left(p(x)\frac{e^\varepsilon - 1}{e^\varepsilon + 1} + \frac{1}{e^\varepsilon + 1}\right),$$

which has a similar form as (13). Hence setting $\beta = 1/(e^\varepsilon + 1)$ and $\gamma = (e^\varepsilon - 1)/(e^\varepsilon + 1)$, we can follow the steps in the communication constrained setting and obtain with probability at least $9/10$, $d_{TV}(\widehat{p}, p) \leq \alpha$ using

$$n = O\left(\beta\frac{s^2 \log\frac{k}{s}}{\alpha^2\gamma^2}\right) = O\left(\frac{s^2 \max\{\log(k/s), 1\}}{\alpha^2\varepsilon^2}\right),$$

when $\varepsilon = O(1)$.