Pointwise Bounds for Distribution Estimation under Communication Constraints

Wei-Ning Chen
Department of Electrical Engineering
Stanford University
wnchen@stanford.edu

Peter Kairouz
Google Research
kairouz@google.com

Ayfer Özgür
Department of Electrical Engineering
Stanford University
aozgur@stanford.edu

Abstract

We consider the problem of estimating a d-dimensional discrete distribution from its samples observed under a b-bit communication constraint. In contrast to most previous results that largely focus on the global minimax error, we study the local behavior of the estimation error and provide *pointwise* bounds that depend on the target distribution p. In particular, we show that the ℓ_2 error decays with $O\left(\max\left(\frac{\|p\|_{1/2}}{n2^b}, \frac{1}{n}\right)\right)$ when n is sufficiently large, hence it is governed by the *half-norm* of p instead of the ambient dimension d. For the achievability result, we propose a two-round sequentially interactive estimation scheme that achieves this error rate uniformly over all p. This two-round scheme extends to ℓ_q loss with $q \geq 1$, and hence gives pointwise upper bounds on ℓ_q error. We also develop a new local minimax lower bound with (almost) matching ℓ_2 error, showing that any interactive scheme must admit a $\Omega\left(\frac{\|p\|_{(1+\delta)/2}}{n2^b}\right)$ ℓ_2 error for any $\delta > 0$.

Our upper and lower bounds together indicate that the $\mathsf{H}_{1/2}(p) \triangleq \log(\|p\|_{1/2})$ bits of communication is both sufficient and necessary to achieve the optimal (centralized) performance, where $\mathsf{H}_{1/2}(p)$ is the Rényi entropy of order 2. Therefore, under the ℓ_2 loss, the correct measure of the local communication complexity at p is its Rényi entropy.

1 Introduction

Learning a distribution from its samples has been a fundamental task in unsupervised learning dating back to the late nineteenth century [1]. This task, especially under distributed settings, has gained growing popularity in the recent years as data is increasingly generated "at the edge" by countless sensors, smartphones, and other devices. When data is distributed across multiple devices, communication cost and bandwidth often become a bottleneck hampering the training of high-accuracy machine learning models [2]. This is even more so for federated learning and analytics type settings [3] which rely on wireless mobile links for communication.

To resolve this issue, several communication-efficient distribution learning schemes have been recently proposed and studied in the literature (see Section 1.2 for a thorough discussion). On the positive side, the state-of-the-art schemes are known to be worst-case (minimax) optimal as they have been shown to achieve the information-theoretic lower bounds on the global minimax error [4–8]. On the

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

negative side, however, the ℓ_2 estimation error achieved by these schemes scales as $O(\frac{d}{n2^b})$ under a b-bit communication constraint on each sample, where d is the alphabet size of the unknown discrete distribution p. This suggests that without additional assumptions on p the error scales *linearly* in d, i.e. the introduction of communication constraints introduces a penalty d on the estimation accuracy. This is true even if we allow for interaction between clients [8,9].

A recent work [10] has moved a step forward from the "global minimax regime" by restricting the target distribution p to be s-sparse and showing that the ℓ_2 error can be reduced to $O(\frac{s\log(d/s)}{n2^b})$ in this case, i.e. the error depends on the sparsity s rather than the ambient dimension d. More recently, [11] has improved on this result by developing a group-testing based scheme that reduces the ℓ_2 error to $O(\frac{s}{n2^b})$, therefore completely removing the dependence on the ambient dimension d and matching the information-theoretic lower bound.

However, both of these schemes and their performance guarantees are limited to the s-sparse setting. Little is known when the target distribution deviates slightly from being exactly s-sparse. Moreover, these schemes are essentially still worst-case schemes as they are designed to be minimax optimal, albeit over a smaller class of distributions, i.e. s-sparse.

In this paper, we argue that all these results can be overly pessimistic, as worst-case notions of complexity and schemes designed to optimize these worst-case notions can be too conservative. Instead, we seek a measure of local complexity that captures the hardness of estimating a specific instance p. Ideally, we want a scheme that adapts to the hardness of the problem instead of being tuned to the worst-case scenario; that is, a scheme achieving smaller error when p is "simpler."

Our contributions Motivated by these observations, in this work we consider the *local* minimax complexity of distribution estimation and quantify the hardness of estimating a specific p under communication constraints. In particular, under the ℓ_2 loss, we show that the local complexity of estimating p is captured by its half-norm[1] $\|p\|_{\frac{1}{2}}$: we propose a two-round interactive scheme that uniformly achieves the $O\left(\max\left(\frac{\|p\|_{1/2}}{n2^b}, \frac{1}{n}\right)\right)$ error under ℓ_2 loss[2] which requires no prior information on p. On the impossibility side, we also show that for any (arbitrarily) interactive scheme, the local minimax error (which is formally defined in Theorem 2.4) must be at least $\Omega\left(\max\left(\frac{\|p\|_{1+\delta/2}}{n2^b}, \frac{1}{n}\right)\right)$ for any $\delta > 0$ when n is sufficiently large.

These upper and the lower bounds together indicate that $\|p\|_{1/2}$ plays a fundamental role in distributed estimation and that the $\lceil \log(\|p\|_{1/2})\rceil$ bits of communication is both sufficient and necessary to achieve the optimal (centralized) performance when n is large enough. Indeed, this quantity is exactly the Rényi entropy of order 2, i.e. $\mathsf{H}_{1/2}(p) \triangleq \log(\|p\|_{1/2})$, showing that under the ℓ_2 loss, the correct measure of the local communication complexity at p is the Rényi entropy of p.

Compared to the global minimax results where the error scales as $O(\frac{d}{n2^b})$, we see that when we move toward the local regime, the linear dependency on d in the convergence rate is replaced by $\|p\|_{1/2}$. This dimension independent convergence is also empirically verified by our experiments (see Section 3 for more details). Note that $\|p\|_{1/2} < d$, so our proposed scheme is also globally minimax optimal. Moreover, since $\|p\|_{1/2} < \|p\|_0$, our scheme also achieves the $O(\frac{s}{n2^b})$ convergence rate as in [11] under the s-sparse model (though admittedly, the scheme in [11] is designed under the more stringent non-interactive setting). As another immediate corollary, our pointwise upper bounds indicate that 1 bit suffices to attain the performance of the centralized model when the target distribution is highly skewed, such as the (truncated) Geometric distributions and Zipf distributions with degree greater than two.

Our techniques Our proposed two-round interactive scheme is based on a local refinement approach, where in the first round, a standard (global) minimax optimal estimation scheme is applied to localize p. In the second round, we use additional $\Theta(n)$ samples (clients), together with the information obtained from the previous round, to locally refine the estimate. The localization-refinement procedure enables us to tune the encoders (in the second round) to the target distribution p and hence attain the optimal pointwise convergence rate uniformly.

[1] Here we generalize the notion of q-norm $\|p\|_q \triangleq \left(\sum_{i=1}^d p_i^q\right)^{1/q}$ to $q \in [0,1]$.

[2] Our scheme also guarantees pointwise upper bounds on ℓ_1 or general ℓ_q errors. See Theorem 2.2.

On the other hand, our lower bound is based on the quantized Fisher information framework introduced in [12]. However, in order to obtain a local bound around p, we develop a new approach that first finds the best parametric sub-model containing p, and then upper bounds its Fisher information in a neighborhood around p. To the best of our knowledge, this is the first impossibility result that allows to capture the local complexity of high dimensional estimation under information constraints and can be of independent interest, e.g. to derive pointwise lower bounds for other estimation models under general information constraints.

1.1 Notation and Setup

The general distributed statistical task we consider in this paper can be formulated as follows. Each one of the n clients has local data $X_i \sim p$, where $p \in \mathcal{P}_d$ and $\mathcal{P}_d \triangleq \left\{ p \in \mathbb{R}_+^d \,\middle|\, \sum_j p_j = 1 \right\}$ is the collection of all d-dimensional discrete distributions. The i-th client then sends a message $Y_i \in \mathcal{Y}$ to the server, who upon receiving Y^n aims to estimate the unknown distribution p.

At client i, the message Y_i is generated via a *sequentially interactive* protocol; that is, samples are communicated sequentially by broadcasting the communication to all nodes in the system including the server. Therefore, the encoding function W_i of the i-th client can depend on all previous messages $Y_1, ..., Y_{i-1}$. Formally, it can be written as a randomized mapping (possibly using shared randomness across participating clients and the server) of the form $W_i(\cdot|X_i, Y^{i-1})$, and the b-bit communication constraint restricts $|\mathcal{Y}| \leq 2^b$. As a special case, when W_i depends only on X_i and is independent of the other messages Y^{i-1} for all i (i.e. $W_i(\cdot|X_i, Y^{i-1}) = W_i(\cdot|X_i)$), we say the corresponding protocol is *non-interactive*. Finally, we call the tuple $(W^n, \hat{p}(Y^n))$ an estimation scheme, where $\hat{p}\,(Y^n)$ is an estimator of p. We use Π_{seq} and Π_{ind} to denote the collections of all sequentially interactive and non-interactive schemes respectively.

Our goal here is to design a scheme $(W^n, \hat{p}\,(Y^n))$ to minimize the ℓ_2 (or ℓ_1) estimation error: $r\left(\ell_2, p, (W^n, \hat{p})\right) \triangleq \mathbb{E}[\|p - \hat{p}\,(Y^n)\|_2^2]$, for all $p \in \mathcal{P}_d$, as well as characterizing the best error achievable by any scheme in Π_{seq}. We note that while our impossibility bounds hold for any scheme in Π_{seq}, the particular scheme we propose uses only one round of interaction. For ℓ_1 error, we replace $\|\cdot\|_2^2$ in the above expectations with $\|\cdot\|_1$. In this work, we mainly focus on the regime $1 \ll d \ll n$, aiming to characterize the statistical convergence rates when n is sufficiently large.

1.2 Related works

Estimating discrete distributions is a fundamental task in statistical inference and has a rich literature [13–16]. Under communication constraints, the optimal convergence rate for discrete distribution estimation was established in [4–8, 17] for the non-interactive setting, and [8, 9] for the general (blackboard) interactive model. The recent work [10, 11] considers the same task under the s sparse assumption for the distribution under communication or privacy constraints. However, all these works study the global minimax error and focus on minimizing the worst-case estimation error. Hence the resultant schemes are tuned to minimize the error in the worst-case which may be too pessimistic for most real-world applications.

A slightly different but closely related problem is distributed estimation of distributions [17–21] and heavy hitter detection under local differential privacy (LDP) constraints [22–25]. Although originally designed for preserving user privacy, some of these schemes can be made communication efficient. For instance, the scheme in [21] suggests that one can use 1 bit communication to achieve $O(d/n)$ ℓ_2 error (which is global minimax optimal); the non-interactive tree-based schemes proposed in [23, 24] can be cast into a 1-bit frequency oracle, but since the scheme is optimized with respect to ℓ_∞ error, directly applying their frequency oracle leads to a sub-optimal convergence $O\left(\frac{s(\log d + \log n)}{n}\right)$ in ℓ_2.

In the line of distributed estimation under LDP constraint, the recent work [26] studies *local* minimax lower bounds and shows that the local modulus of continuity with respect to the variation distance governs the rate of convergence under LDP. However, the notion of the local minimax error in [26] is based on the two-point method, so their characterized lower bounds may not be uniformly attainable in general high-dimensional settings.

Organization The rest of the paper is organized as follows. In Section 2, we present our main results, including pointwise upper bounds and an (almost) matching local minimax lower bound. We provide examples and experiments in Section 3 to demonstrate how our pointwise bounds improve upon previous global minimax results. In Section 4, we introduce our two-round interactive scheme that achieves the pointwise bound. The analysis of this scheme, as well as the proof of the upper bounds, is given in Section 5. Finally, in Section 6 we provide the proof of the local minimax lower bound.

2 Main results

Our first contribution is the design of a two-round interactive scheme (see Section 4 for details) for the problem described in the earlier section. The analysis of the convergence rate of this scheme leads to the following pointwise upper bound on the ℓ_2 error:

Theorem 2.1 (Local ℓ_2 upper bound) *For any $b \leq \lfloor \log_2 d \rfloor$, there exist a sequentially interactive scheme $(W^n, \check{p}) \in \Pi_{\mathsf{seq}}$, such that for all $p \in \mathcal{P}_d$,*

$$r\left(\ell_2, p, (W^n, \check{p})\right) \leq C_1 \frac{1}{n} + C_2 \frac{\|p\|_{\frac{1}{2}}}{n2^b} + C_3 \frac{d^3 \log(nd)}{(n2^b)^2} \asymp \frac{\|p\|_{\frac{1}{2}} + o_n(1)}{n2^b}, \tag{1}$$

where $C_1, C_2, C_3 > 0$ are some universal constants (which are explicitly specified in Section 5). This implies that as long as $n = \Omega\left(\frac{d^3 \log d}{2^b \|p\|_{1/2}}\right)$, $r\left(\ell_2, p, (W^n, \check{p})\right) = O\left(\max\left(\frac{\|p\|_{1/2}}{n2^b}, \frac{1}{n}\right)\right)$.

In addition to the ℓ_2 error, by slightly tweaking the parameters in our proposed scheme, we can obtain the following pointwise upper bound on the ℓ_1 (i.e. the total variation) error:

Theorem 2.2 (Local ℓ_1 upper bound) *For any $b \leq \lfloor \log_2 d \rfloor$, there exists a sequentially interactive scheme $\left(\tilde{W}^n, \check{p}\right) \in \Pi_{\mathsf{seq}}$, such that for all $p \in \mathcal{P}_d$,*

$$r\left(\ell_1, p, \left(\tilde{W}^n, \check{p}\right)\right) \leq C_1 \sqrt{\frac{\|p\|_{\frac{1}{2}}}{n}} + C_2 \sqrt{\frac{\|p\|_{\frac{1}{3}}}{n2^b}} + C_3 \sqrt{\frac{d^4 \log(nd)}{(n2^b)^2}} \asymp \sqrt{\frac{\|p\|_{\frac{1}{3}} + o_n(1)}{n2^b}},$$

where $C_1, C_2, C_3 > 0$ are some universal constants. Hence as long as $n = \Omega\left(\frac{d^4 \log d}{2^b \|p\|_{1/3}}\right)$, $r\left(\ell_1, p, (W^n, \check{p})\right) = O\left(\max\left(\sqrt{\frac{\|p\|_{1/3}}{n2^b}}, \sqrt{\frac{\|p\|_{1/2}}{n}}\right)\right)$.

Theorem 2.1 implies that the convergence rate of the ℓ_2 error is dictated by the half-norm of the target distribution p while Theorem 2.2 implies that the ℓ_1 error is dictated by the one-third norm of the distribution. In general, we can optimize the encoding function with respect to any ℓ_q loss for $q \in [1, 2]$ and the quantity that determines the convergence rate becomes the $\frac{q}{q+2}$-norm (see Remark C.1 for details). We also remark that the large second order terms in Theorem 2.1 and Theorem 2.2 are generally inevitable. See Appendix A.1 for a discussion. The proofs of Theorem 2.1 and Theorem 2.2 are given in Section 5 and Appendix C respectively.

Next, we complement our achievability results with the following minimax lower bounds.

Theorem 2.3 (Global minimax lower bound) *For any (possibly interactive) scheme, it holds that*

$$\inf_{(W^n, \hat{p}) \in \Pi_{\mathsf{seq}}} \sup_{p: \|p\|_{1/2} \leq h} \mathbb{E}_{X \sim p}\left[\|\hat{p}(W^n(X^n)) - p\|_2^2\right] = \Omega\left(\frac{h}{n2^b}\right).$$

Note that the lower bound in Theorem 2.3 can be uniformly attained by our scheme even if no information on the target set of distributions, i.e. the parameter h is available, indicating that our proposed scheme is global minimax optimal. The proof can be found in Section D.

Theorem 2.4 (Local minimax lower bound) *Let* $p \in \mathcal{P}'_d \triangleq \left\{p \in \mathcal{P}_d \middle| \frac{1}{2} < p_1 < \frac{2}{3}\right\}$. *Then for any* $\delta > 0, B \geq \sqrt{\frac{\|p\|_{1/2}}{d2^b}}$, *as long as* $n = \Omega\left(\frac{d^3 \log d}{\|p\|_{1/2}}\right)$, *it holds that*[3]

$$\inf_{(W^n,\hat{p})\in\Pi_{\mathsf{seq}}} \sup_{p':\|p'-p\|_\infty \leq \frac{B}{\sqrt{n}}} \mathbb{E}_{p'}\left[\|\hat{p}\left(W^n(X^n)\right) - p'\|_2^2\right] \geq \frac{1}{n2^b} \max\left(c_\delta \|p\|_{\frac{1+\delta}{2}}, \frac{c_1 \|p\|_{\frac{1}{2}}}{\log d}\right) + \frac{c_2}{n},$$

for some $c_\delta, c_1, c_2 > 0$.

Remark 2.1 *Note that in Theorem 2.4, the* ℓ_∞ *neighborhood* $\left\{p' : \|p' - p\|_\infty \leq \frac{B}{\sqrt{n}}\right\}$ *is strictly smaller than the following* ℓ_2 *neighborhood:* $\left\{p' : \|p' - p\|_2 \leq \sqrt{\|p\|_{1/2}/(2^b n)}\right\}$ *as* $\mathcal{B}_{d,\ell_\infty}(\frac{1}{\sqrt{d}}) \subset \mathcal{B}_{d,\ell_2}(1)$.

Theorem 2.4 indicates that our proposed scheme is (almost) optimal when n is sufficiently large and that $\|p\|_{1/2}$ is the fundamental limit for the estimation error. This conclusion implies that under ℓ_2 loss, the right measure of the hardness of an instance p is its half-norm, or equivalently its Rényi entropy of order $1/2$. The proof of Theorem 2.4 is given in Section 6.

Corollary 2.1 *When* n *is sufficiently large,* $\lceil \log(\|p\|_{1/2}) \rceil$ *bits of communication is both sufficient and necessary to achieve the convergence rate of the centralized setting under* ℓ_2 *loss. Similarly,* $\lceil \log(\|p\|_{1/3}) \rceil$ *bits are sufficient for* ℓ_1 *loss.*

In addition, observe that the quantities $\log(\|p\|_{1/2})$ and $\frac{1}{2}\log\left((p\|_{1/3}\right)$ are the Rényi entropies of order $1/2$ and $1/3$, denoted as $\mathsf{H}_{1/2}(p)$ and $\mathsf{H}_{1/3}(p)$, respectively. In other words, the communication required to achieve the optimal (i.e. the centralized) rate is determined by the Rényi entropy of the underlying distribution p.

Finally, we remark that when the goal is to achieve the centralized rate with minimal communication (instead of achieving the best convergence rate for a fixed communication budget as we have assumed so far), the performance promised in the above corollary can be achieved without knowing $\mathsf{H}_{1/2}(p)$ beforehand. See Appendix A.2 for a discussion.

3 Examples and experiments

Next, we demonstrate by several examples and experiments that our results can recover earlier global minimax results and can significantly improve them when the target distribution is highly skewed.

Corollary 3.1 (s-sparse distributions) *Let* $P_{s,d} \triangleq \left\{p \in [0,1]^d \middle| \sum p_i = 1, \|p\|_0 \leq s\right\}$. *Then as long as* n *large enough, there exists interactive schemes* $(W^n, \check{p})$ *and* $(\tilde{W}^n, \check{p})$ *such that* $r\left(\ell_2, p, (W^n, \check{p})\right) = O(\frac{s}{n2^b})$ *and* $r\left(\ell_1, p, \left(\tilde{W}^n, \check{p}\right)\right) = O(\sqrt{\frac{s}{n2^b}})$.

The above result shows that our scheme is minimax optimal over $\mathcal{P}_{d,s}$. This recovers and improves (by a factor of $\log d/s$) the results from [10] on the ℓ_1 and ℓ_2 convergence rates for s-sparse distributions[4].

Corollary 3.2 (Truncated geometric distributions) *Let* $p \triangleq \mathsf{Geo}_{\beta,d}$ *be the truncated geometric distribution. That is, for all* $\beta \in (0,1]$ *and* $k \in [d]$, $\mathsf{Geo}_{\beta,d}(k) \triangleq \frac{1-\beta}{\beta(1-\beta^d)}\beta^k$. *Then if* $n = \Omega\left(d^3 \log d/2^b\right)$,

$$r\left(\ell_2, p, (W^n, \hat{p})\right) = O\left(\frac{1}{n2^b}\frac{(1+\sqrt{\beta})(1-\sqrt{\beta}^d)}{(1-\sqrt{\beta})(1+\sqrt{\beta}^d)}\right) = O\left(\max\left(\frac{1}{n(1-\sqrt{\beta})2^b}, \frac{1}{n}\right)\right).$$

[3] Indeed, the lower bound holds for blackboard interactive schemes [27], a more general class of interactive schemes than Π_{seq}. See [8] for a discussion of blackboard schemes.

[4] We remark, however, that their scheme is *non-interactive* and has smaller minimum sample size requirement.

This result shows that if β is constant for the truncated geometric distribution, 1 bit suffices to achieve the centralized convergence rate $O\left(\frac{1}{n}\right)$ in this case. Note that this is a significant improvement over previous minimax bounds on the the ℓ_2 error which are $O(\frac{d}{n2^b})$ [4]. This suggests that the corresponding minimax optimal scheme is suboptimal by a factor of d when the target distribution is truncated geometric. Our results suggest that the ℓ_2 error should not depend on d at all, and Figure 1 provides empirical evidence to justify this observation.

Corollary 3.3 (Truncated Zipf distributions with $\lambda > 2$) *Let $p \triangleq \mathsf{Zipf}_{\lambda,d}$ be a truncated Zipf distribution with $\lambda \geq 2$. That is, for $k \in [d]$, $\mathsf{Zipf}_{\lambda,d}(k) \triangleq \frac{k^{-\lambda}}{\sum_{k'=1}^{d}(k')^{-\lambda}}$. Then the local complexity of p is characterized by $\|p\|_{1/2} = \Theta\left(\left(\frac{1-d^{-\lambda/2+1}}{1-d^{-\lambda+1}}\right)^2\right) = \Theta(1)$. So in this case, $r\left(\ell_2, p, (W^n, \hat{p})\right) = O\left(\frac{1}{n}\right)$, as long as $n = \Omega\left(d^3 \log d/2^b\right)$.*

We leave the complete characterization for all $\lambda > 0$ to Section E in appendix. Note that since both $\mathsf{Geo}_{\beta,d}$ and $\mathsf{Zipf}_{\lambda,d}$ are not sparse distributions, [10] cannot be applied here. Therefore, the best previously known scheme for these two cases is the global minimax optimal scheme achieving an $O\left(\frac{d}{n2^b}\right)$ ℓ_2 error. Again, our results suggest that this is suboptimal by a factor of d.

In Figure 1, we empirically compare our scheme with [4] (which is globally minimax optimal). In the left figure, we see that the error of our scheme is an order of magnitude smaller than the minimax scheme and remains almost the same under different values of d. We illustrate that more clearly in the right figure, where we fix n and increase d. It can be observed that the error of our scheme remains bounded when d increases, while the error of the minimax scheme scales linearly in d. This phenomenon is justified by Corollary 3.2.

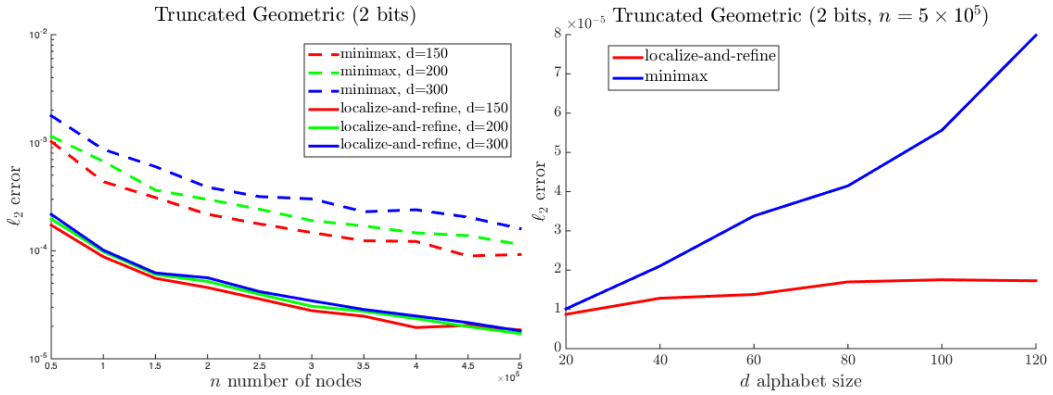

Figure 1: Comparisons between our scheme (labeled as 'localize-and-refine') and the globally minimax optimal scheme (labeled as 'minimax') [4]. The underlying distribution is set to be truncated geometric distribution with $\beta = 0.8$.

4 The localization-refinement scheme

In this section, we introduce our two-round localization-refinement scheme and show how it locally refines a global minimax estimator $\hat{p}$ to obtain a pointwise upper bound.

The first round of our two-round estimation scheme (the localization phase) is built upon the global minimax optimal scheme [4], in which symbols are partitioned into $\lfloor d/(2^b - 1) \rfloor$ subsets each of size $(2^b - 1)$, and each of the first $n/2$ clients is assigned one of these subsets (each subset is assigned to $(n/2)/(d/(2^b - 1))$ clients). The client reports its observation only if it is in its assigned subset by using b bits. This first round allows us to obtain a coarse estimate $\hat{p}$ with error $\mathbb{E}[\|p - \hat{p}\|_2^2] \asymp \sum_j p_j/n2^b = O(d/n2^b)$. Then in the second round (the refinement phase), we locally refine the estimate by adaptively assigning more "resources" (i.e. samples), according to $\hat{p}$ obtained from the first round, to the more difficult symbols (i.e. $j \in [d]$ with larger error

Algorithm 1: uniform grouping [4] (at client i)

Input: $X_i \in [d], b \in \mathbb{N}$
Compute $M = d/\left(2^b - 1\right)$;
$m \leftarrow i \bmod M, Y_i \leftarrow 0$;
if $X_i \in \left[(m-1)(2^b-1)+1 : m(2^b-1)\right]$ **then**
  $Y_i \leftarrow X_i \bmod (2^b - 1)$
end
return Y_i

Algorithm 2: localize-and-refine (at client i)

Input: $X_i \in [d], b \in \mathbb{N}, \hat{p}$
$(\mathcal{G}_1, ..., \mathcal{G}_d) \leftarrow \texttt{gen_group}\,(\hat{p}, d), Y_i \leftarrow 0$;
Compute $\mathcal{J}_i \triangleq \{j \in [d] | i \in \mathcal{G}_j\}$;
if $X_i \in \mathcal{J}_i$ **then**
  $Y_i \leftarrow$ the ranking of X_i in $\mathcal{J}_i$
end
return Y_i

$\mathbb{E}\left[(\hat{p}_j - p_j)^2\right]$). In particular, for the ℓ_2 error, the number of samples assigned to symbol j will be roughly in proportional to $\sqrt{p_j}$.

It turns out that the local refinement step can effectively mitigate the estimation error, enabling us to move from a worst-case bound to a pointwise bound.

Round 1 (localization): In the first round, the first $n/2$ clients collaboratively produce a coarse estimate $\hat{p} = (\hat{p}_1, ..., \hat{p}_j)$ for the target distribution p via the grouping scheme described above (this is the minimax optimal scheme proposed in [4], see Algorithm 1 for details). Note that in general, this scheme can be replaced by any non-interactive minimax optimal scheme.

Round 2 (local refinement): Upon obtaining $\hat{p}$ from the first round, the server then computes $\pi(\hat{p}) : \mathcal{P}_d \to \mathcal{P}_d$, where the exact definition of π depends on the loss function. π will be used in the design of the rest $n/2$ clients' encoders. In particular, for ℓ_2 loss, we set $\pi_j(p) \triangleq \frac{\sqrt{p_j}}{\sum_{k\in[d]}\sqrt{p_k}}$; for ℓ_1 loss, we set $\pi_j(p) \triangleq \frac{\sqrt[3]{p_j}}{\sum_{k\in[d]}\sqrt[3]{p_k}}$. For notational convenience, we denote $\pi\left(\hat{p}\right)$ as $\hat{\pi} = (\hat{\pi}_1, ..., \hat{\pi}_d)$.

To design the encoding functions of the remaining $n/2$ clients, we group them into d (possibly overlapping) sets $\mathcal{G}_1, ..., \mathcal{G}_d$ with

$$|\mathcal{G}_j| = n_j \triangleq \frac{n}{2}\left(\min\left(1, (2^b - 1)\left(\frac{\hat{\pi}_j}{4} + \frac{1}{4d}\right)\right)\right), \tag{2}$$

for all $j \in [d]$.

We require the grouping $\{\mathcal{G}_j\}$ to satisfy the following properties: 1) each $\{\mathcal{G}_j\}$ consists of n_j distinct clients, and 2) each client i is contained in at most $(2^b - 1)$ groups. Notice that these requirements can always be attained (see, for instance, Algorithm 3 in Section B). We further write $\mathcal{J}_i$ to be the set of indices of the groups that client i belongs to (i.e. $\mathcal{J}_i \triangleq \{j \in [d] | i \in \mathcal{G}_j\}$), so the second property implies $|\mathcal{J}_i| \leq 2^b - 1$ for all i.

As in the first round, client i will report its observation if it belongs to the subset $\mathcal{J}_i$. Since $|\mathcal{J}_i| \leq 2^b - 1$, client i's report can be encoded in b bits. Note that with this grouping strategy each symbol j is reported by clients in $\mathcal{G}_j$, hence the server can estimate p_j by computing $\check{p}_j(Y^n) \triangleq \frac{\sum_{i\in\mathcal{G}_j}\mathbb{1}_{\{X_i=j\}}}{n_j}$. Note that the size of $\mathcal{G}_j$ is dictated by the estimate for p_j obtained in the first round. See Algorithm 2 for the details of the second round.

Remark 4.1 *In the above two-round scheme, we see that the local refinement step is crucial for moving from the worst-case performance guarantee to a pointwise one. Therefore we conjecture that the local lower bound in Theorem 2.4 cannot be achieved by any non-interactive scheme.*

5 Analysis of the ℓ_2 error (proof of Theorem 2.1)

In this section, we analyze the estimation errors of the above two-round localization-refinement scheme and prove Theorem 2.1. Before entering the main proof, we give the following lemma that controls the estimation error of between $\hat{p}$ and π.

Lemma 5.1 *Let $(W^n, \hat{p})$ be the grouping scheme given in Algorithm 1. Then with probability at least $1 - \frac{1}{nd}$, it holds that $\left|\sqrt{p_j} - \sqrt{\hat{p}_j}\right| \leq \sqrt{\varepsilon_n}$ and $\left|\sqrt[3]{p_j} - \sqrt[3]{\hat{p}_j}\right| \leq \sqrt[3]{\varepsilon_n}$, where $\varepsilon_n \triangleq \frac{3d\log(nd)}{n2^b}$.*

Now consider the scheme described in Section 4. After the first round, we obtain a coarse estimate $\hat{p} = (\hat{p}_1, ..., \hat{p}_j)$. Set $\mathcal{E} \triangleq \bigcap_{j\in[d]} \left\{\sqrt{\hat{p}_j} \in \left[\sqrt{p_j} - \sqrt{\varepsilon_n},\ \sqrt{p_j} + \sqrt{\varepsilon_n}\right]\right\}$. Then by Lemma 5.1 and taking the union bound over $j \in [d]$, we have $\mathbb{P}\{\mathcal{E}\} \geq 1 - \frac{1}{n}$. In order to distinguish the estimate obtained from the first round to the final estimator, we use $\check{p}_j$ to denote the final estimator.

Now observe that the ℓ_2 estimation error can be decomposed into

$$\mathbb{E}\left[\|p - \check{p}\|_2^2\right] = \mathbb{P}\{\mathcal{E}^c\}\,\mathbb{E}\left[\|p - \check{p}\|_2^2 \middle| \mathcal{E}^c\right] + \mathbb{P}\{\mathcal{E}\}\,\mathbb{E}\left[\|p - \check{p}\|_2^2 \middle| \mathcal{E}\right] \overset{\text{(a)}}{\leq} \frac{2}{n} + \mathbb{E}\left[\|p - \check{p}\|_2^2 \middle| \mathcal{E}\right]$$

where (a) holds since $\|p - \check{p}\|_2^2 \leq 2$ almost surely. Hence it remains to bound $\mathbb{E}\left[\|p - \check{p}\|_2^2 \middle| \mathcal{E}\right]$. Next, as described in Section 4, we partition the second $n/2$ clients into d overlapping groups $\mathcal{G}_1, ..., \mathcal{G}_d$ according to (2). The reason we choose n_j in this way is to ensure that 1) for symbols with larger p_j (which implies larger estimation error), we allocate them more samples; 2) every symbol is assigned with at least $\Theta(n/d)$ samples.

Clients in $\mathcal{G}_j$ then collaboratively estimate p_j. In particular, client i reports her observation if $X_i \in \mathcal{J}_i \triangleq \{j | i \in \mathcal{G}_j\}$, and the server computes $\check{p}_j(Y^n|\hat{p}) \triangleq \frac{\sum_{i\in\mathcal{G}_j} \mathbb{1}_{\{X_i = j\}}}{n_j} \sim \frac{1}{n_j}\mathsf{Binom}(n_j, p_j)$. Finally, the following lemma controls $\mathbb{E}\left[\|p - \check{p}\|_2^2 \middle| \mathcal{E}\right]$, completing the proof of Theorem 2.1.

Lemma 5.2 *Let $\check{p}$ be defined as above. Then* $\mathbb{E}\left[\|p - \check{p}\|_2^2 \middle| \mathcal{E}\right] \leq \frac{6}{n2^b}\left(\sum_{j\in[d]}\sqrt{p_j}\right)^2 + \frac{10d^2\varepsilon_n}{n2^b} + \frac{1}{n}$.

For the ℓ_1 error, we set $\pi_j(p) \triangleq \frac{\sqrt[3]{p_j}}{\sum_k \sqrt[3]{p_k}}$. Following the similar but slightly more sophisticated analysis, we obtain Theorem 2.2. The detailed proof is left to Section C in appendix.

6 The local minimax lower bound (proof of Theorem 2.4)

Our proof is based on the framework introduced in [8], where a global upper bound on the quantized Fisher information is given and used to derive the minimax lower bound on the ℓ_2 error. We extend their results to the local regime and develop a *local* upper bound on the quantized Fisher information around a neighborhood of p.

To obtain a local upper bound, we construct an h-dimensional parametric sub-model Θ_p^h that contains p and is a subset of $\mathcal{P}_d$, where $h \in [d]$ is a tuning parameter and will be determined later. By considering the sub-model Θ_p^h, we can control its Fisher information around p with a function of h and p. Optimizing over $h \in [d]$ yield an upper bound that depends on $\|p\|_{\frac{1}{2}}$. Finally, the local upper bound on the quantized Fisher information can then be transformed to the local minimax lower bound on the ℓ_2 error via van Tree's inequality [28].

Before entering the main proof, we first introduce some notation that will be used through this section. Let $(p_{(1)}, p_{(2)}, ..., p_{(d)})$ be the sorted sequence of $p = (p_1, p_2, ..., p_d)$ in the non-increasing order; that is, $p_{(i)} \geq p_{(j)}$ for all $i > j$. Denote $\pi : [d] \to [d]$ as the corresponding sorting function[5], i.e. $p_{(i)} = p_{\pi(i)}$ for all $i \in [d]$.

Constructing the sub-model Θ_p^h We construct Θ_p^h by "freezing" the $(d-h)$ smallest coordinates of p and only letting the largest $(h-1)$ coordinates to be free parameters. Mathematically, let

$$\Theta_p^h \triangleq \left\{(\theta_2, \theta_3, ..., \theta_h) \middle| \pi^{-1}\left(\theta_1, \theta_2, ..., \theta_h, p_{(h+1)}, p_{(h+2)}, ..., p_{(d)}\right) \in \mathcal{P}_d\right\}, \tag{3}$$

where $\theta_1 = 1 - \sum_{i=2}^h \theta_i - \sum_{i=h+1}^d p_i$ is fixed when $(\theta_2, ..., \theta_h)$ are determined. For instance, if $p = \left(\frac{1}{16}, \frac{1}{8}, \frac{1}{2}, \frac{1}{16}, \frac{1}{4}\right)$ (so $d = 5$) and $h = 3$, then the corresponding sub-model is

$$\Theta_p^h \triangleq \left\{(\theta_2, \theta_3) \middle| \left(\frac{1}{16}, \theta_3, \theta_1, \frac{1}{16}, \theta_2\right) \in \mathcal{P}_d\right\}.$$

[5] With a slight abuse of notation, we overload π so that $\pi\left((p_1, ..., p_d)\right) \triangleq (p_{\pi(1)}, ..., p_{\pi(d)})$

Bounding the quantized Fisher information Now recall that under this model, the score function

$$S_\theta(x) \triangleq (S_{\theta_2}(x), ..., S_{\theta_h}(x)) \triangleq \left(\frac{\partial \log p(x|\theta)}{\partial \theta_2}, ..., \frac{\partial \log p(x|\theta)}{\partial \theta_h}\right)$$

can be computed as

$$S_{\theta_i}(x) = \begin{cases} \frac{1}{\theta_i}, & \text{if } x = \pi(i),\ 2 \le i \le h \\ -\frac{1}{\theta_1}, & \text{if } x = \pi(1) \\ 0, & \text{otherwise} \end{cases}$$

The next lemma shows that to bound the quantized Fisher information, it suffices to control the variance of the score function.

Lemma 6.1 (Theorem 1 in [8]) *Let W be any b-bit quantization scheme and $I_W(\theta)$ is the Fisher information of Y at θ where $Y \sim W(\cdot|X)$ and $X \sim p_\theta$. Then for any $\theta \in \Theta \subseteq \mathbb{R}^h$,*

$$\mathsf{Tr}\left(I_W(\theta)\right) \le \min\left(I_X(\theta), 2^b \max_{\|u\|_2 \le 1} \mathsf{Var}\left(\langle u, S_\theta(X)\rangle\right)\right).$$

Therefore, for any unit vector $u = (u_2, ..., u_h)$ with $\|u\|_2 = 1$, we control the variance as follows:

$$\begin{aligned} \mathsf{Var}\left(\langle u, S_\theta(X)\rangle\right) &\overset{\text{(a)}}{=} \sum_{i=1}^h \theta_i \left(\sum_{j=2}^h u_j S_{\theta_j}(\pi(i))\right)^2 \\ &= \theta_1 \left(\sum_{j=2}^h u_j\right)^2 \left(\frac{1}{\theta_1}\right)^2 + \sum_{i=2}^h \theta_i u_i^2 \frac{1}{\theta_i^2} \\ &= \frac{\left(\sum_{j=2}^h u_j\right)^2}{\theta_1} + \sum_{j=2}^h \frac{u_i^2}{\theta_i} \\ &\le \frac{h}{\theta_1} + \frac{1}{\min_{j\in\{2,...,h\}} \theta_j}, \end{aligned} \tag{4}$$

where (a) holds since the score function has zero mean. This allows us to upper bound $I_W(\theta)$ in a neighborhood around $\theta(p)$, where $\theta(p)$ is the location of p in the sub-model Θ_p^h, i.e.

$$\theta(p) = (\theta_2(p), ..., \theta_h(p)) \triangleq (p_{(2)}, ..., p_{(h)}).$$

In particular, for any $0 < B \le \frac{p_{(h)}}{3}$ and $p \in \mathcal{P}_d' \triangleq \left\{p \in \mathcal{P}_d \middle| \frac{1}{2} < p_1 < \frac{5}{6}\right\}$, the neighborhood $\mathcal{N}_{B,h}(p) \triangleq \theta(p) + [-B, B]^h$ must be contained in Θ_p^h. In addition, for any $\theta' \in \mathcal{N}_{B,h}(p)$, it holds that

$$\theta_1' \ge \theta_1(p) - \frac{hp_{(h)}}{3} \ge \frac{1}{6},$$

where the second inequality holds since 1) $\theta_1(p) = p_{(1)} \ge \frac{1}{2}$ by our definition of $\mathcal{P}_d'$, and 2) $\frac{hp_{(h)}}{3} \le \frac{\sum_{i=1}^h p_{(1)}}{3} \le \frac{1}{3}$. We also have

$$\min_{j\in\{2,...,h\}} \theta_j' \ge \min_{j\in\{2,...,h\}} \theta_j(p) - \frac{p_{(h)}}{3} \ge \frac{2p_{(h)}}{3}.$$

Therefore (4) implies for any $\theta' \in \mathcal{N}_{B,h}(p)$,

$$\mathsf{Var}\left(\langle u, S_\theta'(X)\rangle\right) \le 6h + \frac{3}{2p_{(h)}}.$$

Together with Lemma 6.1, we arrive at

$$\forall \theta' \in \mathcal{N}_{B,h}(p),\ \mathsf{Tr}\left(I_W(\theta')\right) \le 2^b \left(6h + \frac{3}{2p_{(h)}}\right).$$

Bounding the ℓ_2 error Applying [8, Theorem 3] on $\mathcal{N}_{B,h}(p)$, we obtain

$$\sup_{\theta' \in \mathcal{N}_{B,h}(p)} \mathbb{E}\left[\left\|\hat{\theta} - \theta'\right\|_2^2\right] \geq \frac{h^2}{n2^b\left(6h + \frac{3}{2p_{(h)}}\right) + \frac{h\pi^2}{B^2}} \geq \frac{h^2 p_{(h)}}{10n2^b + \frac{10hp_{(h)}}{B^2}} \geq \frac{h^2 p_{(h)}}{20n2^b}, \tag{5}$$

where the second inequality is due to $hp_{(h)} \leq 1$, and the third inequality holds if we pick $B \geq \sqrt{\frac{hp_{(h)}}{n2^b}}$. Notice that in order to satisfy the condition $\mathcal{N}_{B,h}(p) \subseteq \Theta_p^h$, B must be at most $\frac{p_{(h)}}{3}$, so we have an implicit sample size requirement: n must be at least $\frac{3h}{2^b p_{(h)}}$.

Optimizing over h Finally, we maximize $h^2 p_{(h)}$ over $h \in [d]$ to obtain the best lower bound. The following simple but crucial lemma relates $h^2 p_{(h)}$ to $\|p\|_{\frac{1}{2}}$.

Lemma 6.2 *For any $p \in \mathcal{P}_d$ and $\delta > 0$, it holds that*

$$\|p\|_{\frac{1}{2}} \geq \max_{h \in [d]} h^2 p_{(h)} \geq \max\left(C_\delta \|p\|_{\frac{1+\delta}{2}}, C\|p\|_{\frac{1}{2}}/\log d\right),$$

for $C_\delta \triangleq \left(\frac{\delta}{1+\delta}\right)^{\frac{2}{1+\delta}}$ and a universal constant C small enough.

Picking $h^* = \operatorname{argmax}_{h \in [d]} h^2 p_{(h)}$ and by Lemma 6.2 and (5), we obtain that for all $\hat{\theta}$

$$\sup_{\theta' \in \mathcal{N}_{B_n,h^*}(p)} \mathbb{E}\left[\left\|\hat{\theta} - \theta'\right\|_2^2\right] \geq \max\left(C'_\delta \frac{\|p\|_{\frac{1+\delta}{2}}}{n2^b}, C' \frac{\|p\|_{\frac{1}{2}}}{n2^b \log d}\right),$$

as long as $p \in \mathcal{P}'_d$ and $B_n = \sqrt{\frac{h^* p_{(h^*)}}{n2^b}} \overset{\text{(a)}}{\leq} \sqrt{\frac{d}{n2^b \|p\|_{\frac{1}{2}}}}$, where (a) holds due to the second result of Lemma 6.2 and $h^* \leq d$. In addition, the sample size constraint that n must be larger than $\frac{3h^*}{2^b p_{(h^*)}}$ can be satisfied if $n = \Omega\left(\frac{d^3 \log d}{2^b \|p\|_{\frac{1}{2}}}\right)$ since $\frac{h^*}{p_{(h^*)}} \leq \frac{(h^*)^3 \log d}{C\|p\|_{\frac{1}{2}}} \leq \frac{d^3 \log d}{C\|p\|_{\frac{1}{2}}}$, where the first inequality is due to Lemma 6.2 and the second one is due to $h^* \leq d$. The proof is complete by observing that

$$\inf_{(W^n,\hat{p})} \sup_{p' : \|p'-p\|_\infty \leq B_n} \mathbb{E}\left[\|\hat{p} - p'\|_2^2\right] \geq \inf_{(W^n,\hat{\theta})} \sup_{\theta' \in \mathcal{N}_{B_n,h^*}(p)} \mathbb{E}\left[\left\|\hat{\theta} - \theta'\right\|_2^2\right].$$

7 Conclusion and open problems

We have investigated distribution estimation under b-bit communication constraints and characterized the local complexity of a target distribution p. We show that under ℓ_2 loss, the half-norm of p dictates the convergence rate of the estimation error. In addition, to achieve the optimal (centralized) convergence rate, $\Theta(\mathsf{H}_{1/2}(p))$ bits of communication is both necessary and sufficient.

Many interesting questions remain to be addressed, including investigating if the same lower bound can be achieved by non-interactive schemes, deriving the local complexity under general information constraints (such as ε-local differential privacy constraint), and extending these results to the distribution-free setting (i.e. the frequency estimation problem).

Acknowledgments

This work was supported in part by a Google Faculty Research Award, a National Semiconductor Corporation Stanford Graduate Fellowship, and the National Science Foundation under grants CCF-1704624 and NeTS-1817205.

References

[1] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[2] A. Garg, T. Ma, and H. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," in *Advances in Neural Information Processing Systems*, pp. 2726–2734, 2014.

[3] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[4] Y. Han, P. Mukherjee, A. Ozgur, and T. Weissman, "Distributed statistical estimation of high-dimensional and nonparametric distributions," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 506–510, IEEE, 2018.

[5] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints: Lower bounds from chi-square contraction," in *Conference on Learning Theory*, pp. 3–17, PMLR, 2019.

[6] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints ii: Communication constraints and shared randomness," *arXiv preprint arXiv:1905.08302*, 2019.

[7] Y. Han, A. Özgür, and T. Weissman, "Geometric lower bounds for distributed parameter estimation under communication constraints," in *Conference On Learning Theory*, pp. 3163–3188, PMLR, 2018.

[8] L. P. Barnes, Y. Han, and A. Ozgur, "Lower bounds for learning distributions under communication constraints via fisher information," 2019.

[9] J. Acharya, C. L. Canonne, and H. Tyagi, "General lower bounds for interactive high-dimensional estimation under information constraints," *arXiv preprint arXiv:2010.06562*, 2020.

[10] J. Acharya, P. Kairouz, Y. Liu, and Z. Sun, "Estimating sparse discrete distributions under local privacy and communication constraints," in *Algorithmic Learning Theory*, PMLR, 2021.

[11] W.-N. Chen, P. Kairouz, and A. Ozgur, "Breaking the dimension dependence in sparse distribution estimation under communication constraints," in *Proceedings of Thirty Fourth Conference on Learning Theory*, vol. 134 of *Proceedings of Machine Learning Research*, pp. 1028–1059, PMLR, 15–19 Aug 2021.

[12] L. P. Barnes, W.-N. Chen, and A. Ozgur, "Fisher information under local differential privacy," *arXiv preprint arXiv:2005.10783*, 2020.

[13] R. E. Barlow, "Statistical inference under order restrictions; the theory and application of isotonic regression," tech. rep., 1972.

[14] L. Devroye and L. Gábor, *Nonparametric Density Estimation: The L1 View*. Wiley Interscience Series in Discrete Mathematics, Wiley, 1985.

[15] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. Springer Series in Statistics, Springer New York, 2012.

[16] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall, 1986.

[17] W.-N. Chen, P. Kairouz, and A. Ozgur, "Breaking the communication-privacy-accuracy trilemma," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[18] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under local differential privacy," in *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 759–763, June 2017.

[19] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *26th {USENIX} Security Symposium ({USENIX} Security 17)*, pp. 729–745, 2017.

[20] J. Acharya, Z. Sun, and H. Zhang, "Hadamard response: Estimating distributions privately, efficiently, and with little communication," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1120–1129, 2019.

[21] J. Acharya and Z. Sun, "Communication complexity in locally private distribution estimation and heavy hitters," in *International Conference on Machine Learning*, pp. 51–60, 2019.

[22] R. Bassily and A. Smith, "Local, private, efficient protocols for succinct histograms," in *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, (New York, NY, USA), p. 127–135, Association for Computing Machinery, 2015.

[23] R. Bassily, K. Nissim, U. Stemmer, and A. Thakurta, "Practical locally private heavy hitters," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 2285–2293, Curran Associates Inc., 2017.

[24] M. Bun, J. Nelson, and U. Stemmer, "Heavy hitters and the structure of local privacy," *ACM Transactions on Algorithms (TALG)*, vol. 15, no. 4, pp. 1–40, 2019.

[25] W. Zhu, P. Kairouz, B. McMahan, H. Sun, and W. Li, "Federated heavy hitters discovery with differential privacy," in *International Conference on Artificial Intelligence and Statistics*, pp. 3837–3847, PMLR, 2020.

[26] J. C. Duchi and F. Ruan, "The right complexity measure in locally private estimation: It is not the fisher information," *arXiv preprint arXiv:1806.05756*, 2018.

[27] E. Kushilevitz, "Communication complexity," in *Advances in Computers*, vol. 44, pp. 331–360, Elsevier, 1997.

[28] R. D. Gill, B. Y. Levit, *et al.*, "Applications of the van trees inequality: a bayesian cramér-rao bound," *Bernoulli*, vol. 1, no. 1-2, pp. 59–79, 1995.