Proceedings of Machine Learning Research vol 134:1–32, 2021 34th Annual Conference on Learning Theory

Breaking The Dimension Dependence in Sparse Distribution Estimation under Communication Constraints

Wei-Ning Chen WNCHEN@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Peter Kairouz KAIROUZ@GOOGLE.COM
Google Research

Ayfer Özgür AOZGUR@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We consider the problem of estimating a d-dimensional s-sparse discrete distribution from its samples observed under a b-bit communication constraint. The best-known previous result on ℓ_2 estimation error for this problem is $O\left(\frac{s\log(d/s)}{n2^b}\right)$. Surprisingly, we show that when sample size n exceeds a minimum threshold $n^*(s,d,b)$, we can achieve an ℓ_2 estimation error of $O\left(\frac{s}{n2^b}\right)$. This implies that when $n > n^*(s,d,b)$ the convergence rate does not depend on the ambient dimension d and is the same as knowing the support of the distribution beforehand.

We next ask the question: "what is the minimum $n^*(s,d,b)$ that allows dimension-free convergence?". To upper bound $n^*(s,d,b)$, we develop novel localization schemes to accurately and efficiently localize the unknown support. For the non-interactive setting, we show that $n^*(s,d,b) = O\left(\min\left(d^2\log^2 d/2^b, s^4\log^2 d/2^b\right)\right)$. Moreover, we connect the problem with non-adaptive group testing and obtain a polynomial-time estimation scheme when $n = \tilde{\Omega}\left(s^4\log^4 d/2^b\right)$. This group testing based scheme is adaptive to the sparsity parameter s, and hence can be applied without knowing it. For the interactive setting, we propose a novel tree-based estimation scheme and show that the minimum sample-size needed to achieve dimension-free convergence can be further reduced to $n^*(s,d,b) = \tilde{O}\left(s^2\log^2 d/2^b\right)$.

1. Introduction

Estimating a distribution from its samples is a fundamental unsupervised learning problem that has been studied in statistics since the late nineteenth century (Pearson, 1894). Motivated by the fact that data is increasingly being generated "at the edge" (Kairouz et al., 2019) by countless sensors, smartphones, and other devices, away from the central servers that churn through it, there has been significant recent interest in studying this problem in a distributed setting. Assume we observe n i.i.d. samples from an unknown discrete distribution p, $X_1, X_2, \ldots, X_n \sim p$, but each sample X_i is observed at a different client i. Each client has a finite communication budget, say b bits to communicate its sample to a central server which wants to estimate the unknown distribution p under squared ℓ_2 loss. Recent works (Han et al., 2018a; Barnes et al., 2019; Han et al., 2018b) showed that the estimation error in the distributed case can be as large as $O\left(\frac{d}{n2^b}\right)$ where d is the domain size of the distribution. As compared to the classical (centralized) case where a simple empirical frequency estimator yields an ℓ_2 error of $O(\frac{1}{n})$, this implies a significant penalty when b is small and d is large. Moreover in the classical case, the empirical frequency estimator can be applied without any knowledge of the domain of the distribution while the distributed scheme in (Han et al., 2018a) requires d to be known ahead of time at all the nodes.

Fortunately, in many real-world applications such as location tracking, language modeling and web-browsing, the underlying distribution is often supported only on a sparse but unknown subset of size s, denoted $\mathsf{supp}(p)$, of the ambient domain of size d. Motivated by the utility gains due to sparsity in many high-dimensional statistical problems, we can ask whether the factor of d penalty in distributed estimation can be avoided in settings where the underlying distribution is inherently sparse. For example, compressed sensing (Donoho, 2006; Candes et al., 2006) suggests that a sparse vector of dimension d with s non-zero elements can be losslessly compressed into a $s\log(d/s)$-dimensional vector. This may lead one to suspect that in the sparse case, the dimensionality penalty d can be replaced by the "effective" dimension of the problem $s\log(d/s)$, yielding an ℓ_2 error of $O\left(\frac{s\log(d/s)}{n2^b}\right)$ for distributed estimation (assuming that $b \leq \log s$). This result has been recently shown in (Acharya et al., 2021). Note that when the support of the sparse distribution is given beforehand, the earlier results of (Han et al., 2018a; Barnes et al., 2019; Han et al., 2018b) imply an ℓ_2 error of $O\left(\frac{s}{n2^b}\right)$. However, the additional logarithmic dependence on d in (Acharya et al., 2021) appears natural and inevitable in light of classical results on sparsity.

Our contributions In this paper, we prove that one can surprisingly eliminate the $\log d$ term and achieve a *dimension-free*[1] convergence rate $O\left(\frac{s}{n2^b}\right)$, as long as n is larger than a threshold $n^*(s,d,b)$. To achieve $O\left(\frac{s}{n2^b}\right)$ convergence, we propose a two-stage scheme where in the first stage we use a subset of the samples to *localize* the support of p, and then use the remaining samples to refine the estimation. Our key contribution is to carefully design the localization stage, which allows us to improve the best-known result $O\left(\frac{s\log(d/s)}{n2^b}\right)$ to $O\left(\frac{s}{n2^b}\right)$. Note that $O\left(\frac{s}{n2^b}\right)$ is optimal since it is also the minimax convergence rate when $\mathsf{supp}(p)$ is given beforehand. To our knowledge, this is the first work that observes such dimension independent convergence in a sparse setting.

A natural next step for the sparse distribution estimation task is to ask: *"what is the minimum sample size $n^*(s,d,b)$ needed to achieve the dimension-free convergence rate $O\left(\frac{s}{n2^b}\right)$?"* We investigate this question in the non-interactive (each client encodes its observation independently) and interactive settings (the clients can interact in a sequential fashion). For the non-interactive setting, a simple grouping scheme for localization leads to an upper bound of $n^*(s,d,b) = O\left(d^2\log^2 d/2^b\right)$. On the other hand, by using carefully constructed random hash functions in the localization step, we show that $n^*(s,d,b) = O\left(s^4\log^2 d/2^b\right)$, which depends only logarithmically on d. However, the decoding algorithm for this scheme involves searching over all possible s-sparse supports and is computationally inefficient. To resolve the computational issue, we make a non-trivial connection to non-adaptive group testing, showing that as long as $n = \tilde{\Omega}\left(s^4\log^4 d/2^b\right)$[2], we can achieve the optimal estimation error in $\mathsf{poly}(s,\log d)$ time. The resultant group testing based scheme adapts to the sparsity parameter s, and can be applied without knowing it. For the sequentially interactive case, we propose a tree-based scheme that achieves the optimal convergence rate when $n = \tilde{\Omega}\left(s^2\log^2 d/2^b\right)$, showing that interaction between nodes can lead to smaller sample size. Lower bounds on ℓ_1 sample complexity developed in (Acharya et al., 2021, 2019a) imply that 1) $n^*(s,d,b) = \Omega\left(\frac{s^2\log(d/s)}{2^b}\right)$ in the non-interactive case; and $n^*(s,d,b) = \Omega\left(\frac{s^2}{2^b}\right)$ for the interactive setting. This implies that the requirement on n in our sequentially interactive scheme is tight

[1] This means that the convergence rate does not depend on the ambient dimension d.

[2] For ease of presentation, we use the notation $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to hide all dependence on $\log s$ and $\log\log d$.

up to logarithmic terms, while there is a gap between upper and lower bounds on $n^*(s,d,b)$ (in terms of their dependence on s) in the non-interactive case. See Table 1 for a comparison between different localization schemes.

We also study several natural extensions of sparse distribution estimation. When the data is *distribution-free* and the goal is to estimate the empirical frequency of the symbols held by different nodes, we show that our schemes extend naturally and we obtain the same convergence guarantees as in the distributional settings. Second, we show that our schemes extend to the case when p is *approximately-sparse*, and we provide an upper bound on the estimation error.

Notation and setup The general distributed statistical tasks we consider in this paper can be formulated as follows: each one of the n clients has local data $X_i \sim p$ and sends a message $Y_i \in \mathcal{Y}$ to the server, who upon receiving Y^n aims to estimate the underlying distribution $p \in \mathcal{P}_{s,d}$, where

$$\mathcal{P}_{s,d} \triangleq \left\{ p = (p_1, ..., p_d) \,\middle|\, p \in [0,1]^d, \sum_j p_j = 1, \|p\|_0 \leq s \right\}$$

is the collection of all d-dimensional s-sparse discrete distributions.

At client i, the message Y_i is generated via some encoding channel (a randomized mapping that possibly uses shared randomness across participating clients and the server) denoted by a conditional probability $W_i(\cdot|X_i)$ (for the non-interactive setting) or $W_i(\cdot|X_i, Y^{i-1})$ (for the interactive setting). The b-bit communication constraint restricts $|\mathcal{Y}| \leq 2^b$, so without loss of generality we assume $\mathcal{Y} = [2^b]$. When the context is clear, we sometimes view W_i (in the non-interactive setting) as a $2^b \times d$ stochastic matrix, with $[W_i]_{y,x} \triangleq W_i(y|x)$. Finally, we call the tuple $(W^n, \hat{p}(Y^n))$ an estimation scheme, where $\hat{p}\,(Y^n)$ is an estimator of p.

Let $\Pi_{\mathsf{non-int}}$ be the collection of all non-interactive estimation schemes (i.e. W_i is non-interactive for all $i \in [n]$) and Π_{seq} be the collection of sequentially interactive schemes. Our goal is to design a scheme $(W^n, \hat{p}\,(Y^n))$ to minimize the ℓ_2 (or ℓ_1) estimation error:

$$r_{\mathsf{non-int}}\,(\ell_2, n, b) \triangleq \min_{(W^n,\hat{p})\in\Pi_{\mathsf{non-int}}} \max_{p\in\mathcal{P}_{s,d}} \mathbb{E}\left[\|p - \hat{p}\,(Y^n)\|_2^2\right],$$

and r_{seq} defined similarly with the minimum taken in Π_{seq}. For ℓ_1 error, we replace $\|\cdot\|_2^2$ in the above expectations with $\|\cdot\|_1$.

Related works Estimating discrete distributions is a fundamental task in statistical inference and has a rich literature (Barlow, 1972; Devroye and Gábor, 1985; Devroye and Lugosi, 2012; Silverman, 1986). In the case of communication constraints, the optimal convergence rate for discrete

	$n^*\,(d,s,b)$	Decoding time	Interactivity	Randomness
A. uniform grouping	$\Omega\left(\frac{d^2\log^2 d}{2^b}\right)$	$O\left(n2^b\right)$	non-interactive	public-coin
B. random hashing	$\Omega\left(\frac{s^4\log^2\left(\frac{d}{s}\right)}{2^b}\right)$	$O\left(n2^b d^s\right)$	non-interactive	public-coin
C. group testing	$\tilde{\Omega}\left(\frac{s^4\log^4 d}{2^b}\right)$	$O\left(n2^b + \mathsf{poly}(s,\log d)\right)$	non-interactive	public-coin
D. tree-based	$\tilde{\Omega}\left(\frac{s^2\log^2 d}{2^b}\right)$	$\tilde{O}\left(n2^b + s\log d\right)$	interactive	private-coin

Table 1: Performance of using different localization schemes.

distribution estimation was established in (Han et al., 2018b,a; Barnes et al., 2019; Acharya et al., 2019a,b; Chen et al., 2020) for the non-interactive setting, and (Barnes et al., 2019; Acharya et al., 2020) for general interactive models. The recent work of (Acharya et al., 2021) considers the same problem under an s-sparsity assumption for the distribution. Our result improves on their result by removing the dimension-dependent $\log\left(\frac{d}{s}\right)$ term in their upper bound and hence matching a natural lower bound for the error (i.e. when the sparse support is known beforehand).

A slightly different but closely related problem is distributed heavy hitter detection and distribution estimation under local differential privacy constraints (Bassily and Smith, 2015; Bassily et al., 2017; Bun et al., 2019; Zhu et al., 2020) where no distributional assumption on the data is made. Although originally designed for preserving user privacy, the non-interactive tree-based schemes proposed in (Bassily et al., 2017; Bun et al., 2019) can be modified to communication efficient (indeed 1-bit) distribution estimation schemes. However, in the heavy-hitter problem, most of the results optimize with respect to ℓ_∞ error, and directly applying their frequency oracles leads to a sub-optimal convergence $O\left(\frac{s(\log d+\log n)}{n}\right)$ in ℓ_2 (see Section 3 for a brief discussion). On the other hand, the interactive scheme in (Zhu et al., 2020), which identifies (instead of estimating the frequencies of) heavy-hitter symbols, is similar to our proposed interactive scheme in nature. Nevertheless, we extend their result to a communication efficient scheme and explicitly characterize the convergence rate.

In our construction of a non-interactive polynomial-time decodable scheme (Section 4.3), we map the support localization task into the non-adaptive combinatorial group testing problem (Dorfman, 1943; Du et al., 2000; Ngo and Du, 2000). In particular, we show that the Kautz and Singleton's construction of test measurement matrices (Kautz and Singleton, 1964) can be used to design the local encoding channels and obtain a polynomial time decoding algorithm at the cost of a slightly larger sample size requirement. This novel connection opens the possibility to further harness the rich literature on group testing for building structured schemes for high-dimensional statistical problems with sparsity of the type we study here.

Organization The rest of the paper is organized as follows: in Section 2, we present our main results, including the convergence rate and bounds on the minimum sample size requirement. In Section 3, we introduce the main idea of the generic two-stage scheme and propose a non-interactive construction for the second stage (i.e. the estimation phase). In Section 4, we give three non-interactive localization schemes with different sample size requirements and decoding complexity. In Section 5, we introduce a tree-based interactive localization scheme and show how interactivity can be beneficial. Finally, we conclude our paper with a few non-trivial extensions and interesting open problems in Section 6.

2. Main Results

Our main contribution is the design of both non-interactive and interactive schemes that achieve a dimension-free convergence rate $O\left(\frac{s}{n2^b}\right)$ for the problem described in the earlier section. These schemes can generally be described by a two-stage protocol (see Algorithm 1): the server uses the first n_1 clients to localize the support of p, denoted $\mathsf{supp}(p)$, and the remaining $n-n_1$ clients, in addition to the output of the first stage, to estimate p. We will later see that to estimate $\mathsf{supp}(p)$ accurately, there is a minimum requirement on n. We then propose different localization schemes that aim to minimize this requirement on n in different parameter regimes. See Table 1 for a detailed comparison. Our first theorem upper bounds the requirement on n under the non-interactive setting.

Theorem 1

1. *As long as* $n = \Omega\left(\min\left(\frac{d^2\log^2 d}{\min(d,2^b)}, \frac{s^4\log^2\left(\frac{d}{s}\right)}{\min(2^b,s)}\right)\right)$*, the* ℓ_2 *and* ℓ_1 *error for the non-interactive scheme is*

$$\begin{cases} r_{\mathsf{non-int}}\left(\ell_2, n, b\right) = \Theta\left(\max\left(\frac{s}{n2^b}, \frac{1}{n}\right)\right), \\ r_{\mathsf{non-int}}\left(\ell_1, n, b\right) = \Theta\left(\max\left(\frac{s}{\sqrt{n2^b}}, \sqrt{\frac{s}{n}}\right)\right). \end{cases} \tag{1}$$

2. *Moreover, if* $n = \Omega\left(\frac{s^4\log^4 d(\log s + \log\log d)^2}{\min(2^b,s)}\right)$*, then there exists a non-interactive scheme based on group testing that achieves the convergence rate* (1) *with polynomial time decoding complexity* $O\left(n2^b + \mathsf{poly}(s, \log d)\right)$*. Further, the scheme is adaptive to* s *(i.e. requires no knowledge of* s*).*

Note that (1) is the convergence rate *with the knowledge* of $\mathsf{supp}(p)$ given beforehand, so the lower bound follows directly from standard distribution estimation results under communication constraints (by assuming $\mathsf{supp}(p)$ is known). See, for instance, (Barnes et al., 2019). The achievability schemes (i.e. the upper bounds) are given in Section 3 and Section 4. Our results improve the best-known result (Acharya et al., 2021), which has convergence rate $O\left(\frac{s\log(d/s)}{n2^b}\right)$, by a factor of $\log\left(d/s\right)$.

We next show that with interactive schemes, we can further reduce the minimum sample-size requirement to $n = \tilde{\Omega}\left(s^2\log d/2^b\right)$.

Theorem 2 *As long as* $n = \Omega\left(\frac{s^2\log^2 d(\log s + \log\log d)^2}{\min(2^b,s)}\right)$*, the errors for sequentially interactive schemes* $r_{\mathsf{seq}}\left(\ell_2, n, b\right)$ *and* $r_{\mathsf{seq}}\left(\ell_1, n, b\right)$ *are the same as* (1).

The lower bound on the convergence rate also follows from (Barnes et al., 2019), and the upper bound is proved in Section 5.

To translate Theorem 1 and Theorem 2 into the language of sample complexity, let $\mathsf{SC}_{\mathsf{non-int}}(\beta, d, s, b)$ be the ℓ_1[3] sample complexity of non-interactive setting, which is defined as

$$\mathsf{SC}_{\mathsf{non-int}}(\beta, d, s, b) \triangleq \min\left\{n \in \mathbb{N} \,\middle|\, \min_{(W^n,\hat{p})\in\Pi_{\mathsf{non-int}}} \max_{p\in\mathcal{P}_{s,d}} \mathbb{P}\left\{\left\|p - \hat{p}\left(Y^n\right)\right\|_{\mathsf{TV}} \leq \beta\right\} \geq 0.9\right\},$$

and let $\mathsf{SC}_{\mathsf{seq}}$ be defined in the similar way.

Corollary 3 (Sample complexity) *For* $\beta \in (0,1)$*,* $\mathsf{SC}_{\mathsf{non-int}}$ *and* $\mathsf{SC}_{\mathsf{seq}}$ *satisfy*

$$\mathsf{SC}_{\mathsf{non-int}}(\beta, d, s, b) = O\left(\max\left(\frac{s^2}{\beta^2\min\left(2^b, s\right)}, \min\left(\frac{d^2\log^2 d}{\min(2^b, d)}, \frac{s^4\log^2\left(\frac{d}{s}\right)}{\min(2^b, s)}\right)\right)\right), \text{ and}$$

$$\mathsf{SC}_{\mathsf{seq}}(\beta, d, s, b) = O\left(\max\left(\frac{s^2}{\beta^2\min\left(2^b, s\right)}, \frac{s^2\log^2 d\left(\log s + \log\log d\right)^2}{\min(2^b, s)}\right)\right).$$

[3] Note that for discrete distributions, ℓ_1 distance is the same as total variation distance.

Moreover, it holds that

$$\mathsf{SC}_{\mathsf{non-int}}(\beta, d, s, b) = \Omega\left(\max\left(\frac{s^2}{\beta^2 \min\left(2^b, s\right)}, \frac{s^2 \log\left(d/s\right)}{\beta 2^b}\right)\right), \textit{ and}$$
$$\mathsf{SC}_{\mathsf{seq}}(\beta, d, s, b) = \Omega\left(\frac{s^2}{\beta^2 \min\left(2^b, s\right)}\right).$$

The lower bounds in Corollary 3 are from (Acharya et al., 2021, 2019a). Corollary 3 directly implies the following two facts. First, comparing the lower bounds with the upper bounds, we see that our achievability results are tight when 1) $0 < \beta \leq \frac{1}{s\log(d/s)}$ for the non-interactive setting, and 2) $0 < \beta \leq \frac{1}{\log d(\log s + \log\log d)}$ for the sequentially interactive setting. This suggests that for small enough β, increasing the ambient dimension d does not increase the sample complexity at all. Second, we observe from Corollary 3 that the ℓ_1 estimation error must be at least $\Theta(1)$ when 1) $n = O\left(\frac{s^2\log(d/s)}{\min(2^b,s)}\right)$ for the non-interactive model and 2) $n = O\left(\frac{s^2}{\min(2^b,s)}\right)$ for the sequentially interactive model. This results in the following lower bounds for the minimum sample size requirement: 1) $n^*(s,d,b) = \Omega\left(\frac{s^2\log(d/s)}{\min(2^b,s)}\right)$ in the non-interactive case and 2) $n^*(s,d,b) = \Omega\left(\frac{s^2}{\min(2^b,s)}\right)$ for the interactive setting. This implies that the requirement on n in our sequentially interactive scheme is tight up to logarithmic terms, while there is a gap between upper and lower bounds on $n^*(s,d,b)$ (in terms of their dependence on s) in the non-interactive case. We believe that the lower bound on $n^*(s,d,b)$ is also tight in the non-interactive case and the non-interactive schemes can be further improved.

Next, we show that our results extend naturally to the distribution-free setting where we do not assume any underlying distribution on X^n, and the goal is to estimate the empirical distribution $\pi \triangleq (\pi_1, ..., \pi_d)$ of n local observations $X_1, ..., X_n \in (j_1, ..., j_s) \triangleq \mathcal{J} \subset [d]$. Formally, let $\pi_j \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i = j\}}$, and define

$$\tilde{r}_{\mathsf{non-int}}\left(\ell_2, n, b\right) \triangleq \min_{(W^n, \hat{\pi}) \in \Pi_{\mathsf{non-int}}} \max_{X^n \in \mathcal{J}^n} \mathbb{E}\left[\left\|\pi - \hat{\pi}\left(Y^n\right)\right\|_2^2\right]$$

and $\tilde{r}_{\mathsf{seq}}$ similarly. The expectation is over the (possibly shared) randomness used in the algorithm.

Theorem 4 *The convergence rates as well as the sample-size requirements given in Theorem 1 and Theorem 2 hold for* $\tilde{r}_{\mathsf{non-int}}$ *and* $\tilde{r}_{\mathsf{seq}}$.

Finally, when the target distribution p is no longer s-sparse, we have the following bound on the ℓ_2 estimation error:

Theorem 5 *There exists a non-interactive scheme such that as long as* $n = \Omega\left(s^4 \log^2\left(\frac{d}{s}\right)/\min(2^b, s)\right)$,

$$\mathbb{E}\left[\sum_{j\in[d]} \left(\hat{p}_j - p\right)^2\right] \leq C_2 \cdot \left(\max\left(\frac{s}{n2^b}, \frac{1}{n}\right)\right) + C_3 \cdot \sqrt{n\log n} \cdot 2^b \cdot \left(1 - P_{\mathcal{S}}\right),$$

where $P_{\mathcal{S}} \triangleq \sum_{j=1}^s p_{(j)}$ *and* $p_{(j)}$ *is the* j*-th largest value of* p.

Remark 6 *Note that if* p *is exactly* s*-sparse, then* $P_{\mathcal{S}} = 1$*, and Corollary 5 recovers the convergence rate and the second bound on the sample-size requirement in Corollary 1. Moreover, if* $1 - P_{\mathcal{S}} < \frac{\min\left(1, s/2^b\right)}{n^{\frac{3}{2}} \log^{\frac{1}{2}} n2^b}$*, then we recover the convergence rate with exact* s*-sparsity.*

The proofs of Theorem 4 and Theorem 5 can be found in Section E.

```
Algorithm 1: Two-stage decoding alg
Input: $Y^n = (Y^{n_1}, Y^{n_2})$, $\alpha > 0$
Output: $\hat{p} = (\hat{p}_1, ..., \hat{p}_d)$
$\hat{J}_\alpha \leftarrow$ support_localization $(Y^{n_1}, \alpha)$;
$\hat{p} \leftarrow$ estimation $\left(Y^{n_2}, \hat{J}_\alpha\right)$;
return $\hat{p}$
```

```
Algorithm 2: Encoder in estimation stage
Input: $X_i \in [d]$, $b \in \mathbb{N}$
Output: $Y_i$
Generate random hash function
  $h_i(\cdot) : [d] \to [2^b]$ by shared randomness;
return $Y_i = h_i(X_i)$
```

3. A Two-stage Decoding Algorithm

All of our schemes, both interactive and non-interactive, can be generally described by a two-step protocol (Algorithm 1): we partition all participating clients into two groups, such that

- the first group of $n_1 \leq n/2$ clients are used to *localize* the support of p, i.e. estimate $\mathsf{supp}(p)$. In particular, let $\mathcal{J}_\alpha \triangleq \{j \in [d] \,|\, p_j \geq \alpha\}$ be the collection of high probability symbols. Then from the reports Y^{n_1} of the first group of clients, the server generates an estimate on $\mathcal{J}_\alpha$, such that 1) $\mathcal{J}_\alpha \subseteq \hat{\mathcal{J}}_\alpha$ with high probability; and 2) $|\hat{\mathcal{J}}_\alpha| \leq s$ almost surely.
- the reports Y^{n_2} of the second group of $n_2 \triangleq n - n_1$ clients are used to estimate p_j with the knowledge of $\mathcal{J}_\alpha$ from the first stage.

Notice that although the decoding algorithm is two-stage, encoding at each client can be done simultaneously and does not require the knowledge of $\hat{\mathcal{J}}_\alpha$. Also note that n_1, n_2 and α are design parameters that will be specified later. For ease of presentation, we call the first phase *localization* and the second phase *estimation*.

For the estimation phase, we adopt a scheme similar to (Acharya et al., 2021), where each client encodes their local observation via an independent b-bit random hash function[4] $h_i : [d] \to [2^b]$ as described in Algorithm 2. Upon receiving the hashed values from n_2 clients in the second stage, the server estimates the empirical frequencies of all symbols $j \in \hat{\mathcal{J}}_\alpha$ by counting the number of clients $i \in [n_1 + 1 : n]$ such that $Y_i = h_i(j)$, and sets $\hat{p}_j = 0$ for all $j \notin \hat{\mathcal{J}}_\alpha$:

$$\hat{p}_j\left(Y^{n_2}\right) = \begin{cases} \frac{(2^b-1)\cdot\left(\sum_{i=n_1+1}^{n} \mathbb{1}_{\{h_i(j)=Y_i\}}\right)}{n_2\cdot 2^b} - \frac{1}{2^b}, & \text{if } j \in \hat{\mathcal{J}}_\alpha \\ 0, & \text{else.} \end{cases}$$

We provide a more formal description of the scheme in Section A of the appendix. The ℓ_2 estimation error of this scheme can be controlled by the following lemma:

Lemma 7 *The estimation error of the two-stage scheme is upper bounded by*

$$\mathbb{E}\left[\sum_{j\in[d]} \left(\hat{p}_j - p_j\right)^2\right] \leq 2\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} + s\alpha^2 + \frac{s}{n_2 2^b} + \frac{1}{n_2}. \tag{2}$$

The first term in the left side of (2) corresponds to the probability of failure in the first stage, and the remaining terms correspond to the the ℓ_2 error resulting from the second stage provided that the localization was done correctly in the first stage.

[4] A randomized mapping $[d] \to 2^b$ is called random hash function if $\forall x \in [d],\ y \in [2^b],\ \mathbb{P}\left\{h(x) = y\right\} = \frac{1}{2^b}$.

We will later see that we can make the failure probability decay exponentially fast with $n_1\alpha$ (recall that n_1 is the number of clients participating in the localization phase), that is, $\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} = \exp\left(-\frac{n_1\alpha}{f(s,d)}\right)$, where $f(s,d)$ depends only on s and d. Therefore, if we pick $\alpha = \frac{1}{\sqrt{n2^b}}$ and $n_1 = \Omega\left(f(s,d)\sqrt{n2^b}\log n\right)$, we arrive at the dimension-free convergence in Theorem 1.

However, the condition that $n_1 \leq n$ imposes a minimum requirement on the amount of participating clients in the first stage, i.e. $n = \Omega\left(f^2(s,d)\log^2 f(s,d)\right)$. Hence it is critical to carefully design the localization scheme to get the best scaling for $f(s,d)$ with s and d. For instance, if we use count-sketch based methods as in the heavy-hitter literature (Bassily et al., 2017; Bun et al., 2019), we can only localize to the resolution level $\alpha = O\left(\sqrt{(\log n + \log d)/n}\right)$[5], which is not enough to get the $O\left(\frac{s}{n2^b}\right)$ rate. In the next two sections, we introduce three non-interactive and one interactive localization schemes, which have different minimum sample-size requirements and decoding complexities, as summarized in Table 1.

4. Non-interactive (SMP) Localization Schemes

In this section, we present three non-interactive schemes for the localization stage, each offering a different trade-off between the minimum number of samples to achieve the dimension-free convergence rate and decoding time.

4.1. Localization scheme A: uniform grouping

Under the b-bit communication constraint, our first encoding scheme is based on the grouping idea (Han et al., 2018a), where each client only encodes symbols in a pre-specified subset of $[d]$ and ignores others. In particular, we partition the d symbols and n_1 clients into M equal-sized groups (disjoint subsets) denoted by $\mathcal{B}_1, ..., \mathcal{B}_M$ and $\mathcal{G}_1, ..., \mathcal{G}_M$, respectively. Clients $i \in \mathcal{G}_m$ are assigned to the subset of symbols $\mathcal{B}_m$. This means that they only encode symbols in $\mathcal{B}_m$ (and ignore their sample if it is not in $\mathcal{B}_m$ and set $Y_i = 0$). We set $M \triangleq d/(2^b-1)$ so that each $\mathcal{B}_m$ contains exactly $2^b - 1$ symbols, and thus the encoded message can be described in b bits. Upon observing all messages from the clients, the server computes $\hat{\mathcal{J}}_\alpha$ that contains all symbols successfully signaled to it. Note that a symbol $j \in \mathcal{J}_\alpha \cap \mathcal{B}_m$ will be in $\hat{\mathcal{J}}_\alpha$ if a client $i \in \mathcal{G}_m$ observes j. We derive the following bound on the failure probability.

Lemma 8 *Under the above encoding and decoding schemes (see Algorithm 4, 5 in Section D for the formal descriptions), we have* $\mathbb{P}\left\{\hat{\mathcal{J}}_\alpha \not\subset \mathcal{J}_\alpha\right\} \leq s\exp\left(-n_1(2^b-1)\alpha/d\right)$.

We describe the details of the encoding and decoding schemes in Section B, and leave the proof of Lemma 8 to Section E.

Finally, taking $\alpha = \frac{1}{\sqrt{n2^b}}$ and combining Lemma 7 and Lemma 8, we arrive at the following bound for $r_{\mathsf{non-int}}(\ell_2, n, b)$:

$$\mathbb{E}\left[\sum_{j\in[d]}(\hat{p}_j - p_j)^2\right] \leq 2se^{-\frac{n_1\cdot(2^b-1)}{d\cdot\sqrt{n2^b}}} + \frac{2s}{(n-n_1)\cdot 2^b} + \frac{1}{n-n_1} \leq C\cdot\left(\frac{s}{n2^b} \vee \frac{1}{n}\right), \tag{3}$$

[5] See, for instance Table 1 in (Bun et al., 2019), where we pick $\varepsilon = \Theta(1)$, $\beta = \Theta(1/n)$ and $|X| = d$ to translate to our settings.

where in the last inequality we choose $n_1 = n/2$ and assume $n \succeq \frac{d^2 \log^2 d}{2^b}$. This gives the first sample-size requirement in Theorem 1. To bound $r_{\mathsf{non-int}}(\ell_1, n, b)$, we apply Jensen's and Cauchy-Schwarz inequalities to obtain

$$r_{\mathsf{non-int}}(\ell_1, n, b) = \mathbb{E}\left[\|\hat{p} - p\|_1\right] \leq \sqrt{\mathbb{E}\left[\|\hat{p} - p\|_1^2\right]} \leq \sqrt{2s \cdot \mathbb{E}\left[\|\hat{p} - p\|_2^2\right]}. \tag{4}$$

The last inequality holds since by our construction, $|\mathsf{supp}\,(\hat{p}) \cup \mathsf{supp}\,(p)| < 2s$.

4.2. Localization Scheme B: non-uniform random hashing

Though Scheme A achieves the minimax estimation error when $p \in \mathcal{P}_d$ (see (Han et al., 2018a,b)), it is indeed inefficient under the s-sparse assumption. This is because only a small fraction of symbols can be observed with non-zero probability, so for clients assigned to blocks that did not contain these symbols, they always encode their observations to $Y_i = 0$. In our second localization scheme, we aim to improve the encoding efficiency by using random hash functions. However, unlike in the estimation stage described in Section A, the random hash functions we use for localization are generated non-uniformly.

Encoding For $i \in [n_1]$, client i generates their local random hash function $W_i(y|x)$ as follows: each column of W_i, denoted $W_i(\cdot|x) \in \{0,1\}^{2^b}$, is defined as the one-hot representation of $L_{i,x}$, where $L_{i,x} \in [2^b]$ follows a multinomial distribution

$$L_{i,x} \overset{\text{i.i.d.}}{\sim} \mathsf{Mult}\left(1, \left(\frac{1}{s}, \frac{1}{s}, ..., \frac{1}{s}, 1 - \frac{2^b - 1}{s}\right)\right). \tag{5}$$

Formally, we can be express

$$W_i(\cdot|x) \triangleq e_{L_{i,x}} \in \{0,1\}^{2^b \times 1} \tag{6}$$

for all $x \in [d]$ (where e_L is the L-th standard basis vector). Since the W_i corresponds to a deterministic mapping (but randomly generated as described above), sometimes we write $Y_i = h_i(X_i)$ for simplicity. The encoding algorithm resembles the one in the estimation stage (Algorithm 2), except that now the random hash functions are generated according to (6).

Decoding The decoding rule is based on exhaustive search. Due to the s-sparse assumption, there are at most $N \triangleq \binom{d}{s}$ possibilities for $\mathsf{supp}(p)$, which we index by $\mathcal{C}_1, ..., \mathcal{C}_N$. Hence the localization step can be cast into a multiple hypothesis testing problem: let H_ℓ be the hypothesis such that $\mathcal{J}_\alpha \subseteq \mathcal{C}_\ell$, for $\ell \in [N]$. To proceed, we first define the notion of consistency.

Definition 9 *We say H_ℓ (or $\mathcal{C}_\ell$) is* consistent *with observations (W^n, Y^n) if $\mathbb{P}\left\{Y^n | H_\ell, W^n\right\} > 0$.*

The decoding rule is as follows: upon observing local encoding functions and reports $\{(W_i, Y_i), i = 1, ..., n_1\}$ from all clients, the server searches for all candidates $\mathcal{C}_1, ..., \mathcal{C}_N$ and randomly picks one which is consistent with (W^n, Y^n) as our estimate of $\mathcal{J}_\alpha$. By using non-uniform hash functions (i.e. generating $W_i(\cdot|x)$ according to (6)), we can improve the distinguishability of W_i (which is formally defined in Definition 18), which reduces the probability of accepting false hypothesis H_ℓ for some $\mathcal{J}_\alpha \not\subseteq \mathcal{C}_\ell$. Indeed, we can obtain the following bound on $\left\{\mathcal{J}_\alpha \not\subseteq \hat{\mathcal{J}}_\alpha\right\}$:

Lemma 10 *Under the above encoding and decoding rules,*

$$\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} \leq \exp\left(-n_1\alpha(2^b-1)/4s + C_0 s\log\left(d/s\right)\right),$$

for some universal constant $C_0 > 0$. Moreover, with probability 1, $|\hat{\mathcal{J}}_\alpha| = s$.

Again, we pick $n_1 = \frac{n}{2}$ and $\alpha = \frac{1}{\sqrt{n2^b}}$. By combining Lemma 7 and Lemma 10, we arrive at

$$\mathbb{E}\left[\sum_{j\in[d]}\left(\hat{p}_j - p\right)^2\right] \leq 2\exp\left(-\sqrt{\frac{n\left(2^b-1\right)^2}{2^b}}\cdot\frac{1}{8s} + C_0 s\log\left(\frac{d}{s}\right)\right) + \frac{3s}{n2^b} + \frac{2}{n}. \tag{7}$$

To ensure the first term less than $O(1/n)$, we introduce the following simple but useful lemma.

Lemma 11 *Let $f_1(s,d,b) \geq 300$, $f_2(s,d,b) \geq 0$. Then as long as $n \geq 4\cdot f_1^2\cdot\max\left(f_2^2, 16\log^2\left(f_1\right)\right)$,*

$$\exp\left(-\frac{\sqrt{n}}{f_1(s,d,b)} + f_2(s,d,b)\right) \leq \frac{1}{n}.$$

Taking $f_1 = 8s/\sqrt{2^b-1}$, $f_2 = C_0 s\log\left(\frac{d}{s}\right)$ and applying Lemma 11, we see that as long as $n \succeq s^4\log^2\left(\frac{d}{s}\right)/2^b$, the ℓ_2 error (7) is $O\left(\frac{s}{n2^b}\vee\frac{1}{n}\right)$. By (4), we obtain the upper bound on ℓ_1 error. This gives the second bound on the sample-size requirement in Theorem 1.

4.3. Localization Scheme C: non-adaptive combinatorial group testing

The non-uniform random hashing scheme presented in Section 4.2 provides a substantial reduction in the minimum number of samples needed to break the dimension dependence in sample complexity. However, this comes at the expense of increased decoding complexity as it relies on exhaustive search. We now present a group testing based scheme that combines the best of both worlds.

Group testing preliminaries Group testing is the problem of identifying s defective items in a large set of cardinality d by making tests on groups of items. A *group test* is applied to a subset of items $\mathcal{S} \subseteq [d]$. The test outcome Z is *positive* (i.e. $Z = 1$) if at least one item in $\mathcal{S}$ is defective. A group testing algorithm describes how to design $\mathcal{S}_1, ..., \mathcal{S}_T$ and select T such that the defective items can be identified from the test outcomes $Z_1, ..., Z_T$. In the non-adaptive setting, all T tests must be designed and fixed before they are conducted. Therefore, each single test $\mathcal{S}_t$ can be characterized by a row vector $m_t \in \{0,1\}^{1\times d}$, where $m_t(j) = 1$ if the j-th item is included in the t-th test. Therefore, the collection of T tests can be represented by a $T\times d$ binary measurement matrix $M = \left[m_1^\intercal, ..., m_T^\intercal\right]^\intercal$.

The goal of non-adaptive combinatorial group testing (NCGT) is to design the measurement matrix M such that 1) the number of tests T is minimized, and 2) the defective items can be identified correctly (i.e. with *zero-error*) and efficiently (i.e. in $O\left(\mathsf{poly}\left(\log d\right)\right)$ time). In particular, if a matrix M satisfies the *s-disjunct* property described below, then a cover decoder (summarized in Algorithm 8 of Section D) can identify all s defective items in $O(Td)$ time.

Definition 12 (s-disjunct) *Let M be a $T\times d$ binary matrix, M_j be the j-th column of M, and $\mathsf{supp}(M_j) \triangleq \{t|t\in[T],\, M_{t,j} = 1\}$. Then M is said to be s-disjunct if $\mathsf{supp}(M_j) \not\subset \bigcup_{j'\in\mathcal{K}}\mathsf{supp}(M_{j'})$, for all $\mathcal{K}\subset[d]$ such that 1) $|\mathcal{K}| = s$ and 2) $\mathcal{K}$ does not contain j.*

Let $T_{\mathsf{disjunct}}\left(s,d\right)$ denote the minimum number of rows of an d-column s-disjunct matrix. It has been known for about 40 years (D'yachkov and Rykov, 1982) that when $s = O(\sqrt{d})$,

$$\Omega\left(s^2 \log d / \log s\right) \leq T_{\mathsf{disjunct}}\left(s,d\right) \leq O(s^2 \log d).$$

4.3.1. SUPPORT LOCALIZATION VIA GROUP TESTING: 1-BIT CASE

Next, we map the support localization problem into NCGT for the case $b = 1$ by viewing $\mathcal{J}_\alpha \subseteq \mathsf{supp}(p)$ as the defective items that need to be identified (recall that our goal here is to only specify $\mathcal{J}_\alpha$ and $|\mathcal{J}_\alpha| \leq s$). The main difference between the localization task and group testing is as follows: in group testing, a group test provides information about *all* the defective items participating in the test, while in the localization task each client can observe only a single symbol, and hence makes an observation regarding only one "defective item".

To match the two problems, we use multiple clients to simulate a single group test. We partition clients into T bins, and the clients in the τ-th bin encode their observations according to the same group test $\mathcal{S}_\tau$. If the clients in the τ-th bin observe all symbols in $\mathcal{J}_\alpha$, then by taking the Boolean OR of their reported bits, the server can recover the outcome corresponding to the test $\mathcal{S}_\tau$.

Encoding Let $M = \left[m_1^\intercal, ..., m_T^\intercal\right]^\intercal \in \{0,1\}^{T\times d}$ be any zero-error NCGT measurement matrix satisfying the s-disjunct property. Each client encodes their local observation according to a row (i.e. an individual group test) of M. Define $t(i) \triangleq i \,\mathsf{mod}\, T$. We then uniformly partition n_1 clients into T bins by assigning client i into the $t(i)$-th bin. Client i then generates their 1-bit report by setting $Y_i = M_{t(i),X_i}$. Equivalently, client i's 1-bit encoding channel matrix W_i is $W_i(y=1|x) = m_{t(i)}$.

Decoding Let $\mathcal{G}_\tau \triangleq \{i \in [n_1] | t(i) = \tau\}$ denote the τ-th bin of clients. For each $\tau \in [T]$, the server computes $\hat{Z}_\tau \triangleq \bigvee_{i\in\mathcal{G}_\tau} Y_i$. If all symbols in $\mathcal{J}_\alpha$ appear at least once in $\mathcal{G}_\tau$'s observations (i.e. if $\mathcal{J}_\alpha \subseteq \{X_i | i \in \mathcal{G}_\tau\}$), then $\hat{Z}_\tau$ is the same as the result of the τ-th group test of M measuring on $\mathcal{J}_\alpha$, which we denote by Z_τ. Therefore, as long as n_1 is large enough, $\bigcup_{\tau\in T}\left\{\hat{Z}_\tau = Z_\tau\right\}$ holds with high probability, and the server can then identify $\mathcal{J}_\alpha$ by running a standard cover decoder (which is summarized in Algorithm 8 of Section D).

Lemma 13 *Under the above encoding and decoding schemes (see Algorithm 7, 8 for the formal descriptions) with measurement matrix $M \in \{0,1\}^{T\times d}$, we have*

$$\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} \leq \exp\left(-n_1\alpha/T + (\log s + \log T)\right).$$

In addition, with probability 1, $|\hat{\mathcal{J}}_\alpha| \leq s$, and the decoding complexity is $O(n + dT)$.

4.3.2. GENERAL b-BIT CASE

For the general b-bit case, one may attempt to repeatedly apply the 1-bit encoding scheme for b times. That is, each bin of clients $\mathcal{G}_\tau$ simulates b group tests at a time. This can reduce the total number of bins required from T to T/b (and thus increases $|\mathcal{G}_\tau|$ by a factor of b), equivalently yielding a boost on the sample size from n_1 to $n_1 b$. However, according to Lemma 10, we see that by carefully designing the encoding channels $W_i(y|x)$, it is possible to achieve an *exponential* gain on the sample size, i.e. from n_1 to $n_1 2^b$. Therefore, our goal is to design the measurement matrix M in a way that each bin of clients $\mathcal{G}_\tau$ can simulate $\Theta(2^b)$ group tests at a time. Towards this goal, we want M to have the following properties:

1. M is s-disjunct (so that the cover decoder applies).
2. $T = O\left(s^2 \cdot \mathsf{polylog}(d)\right)$.
3. M is (2^b-1)-*blockwise sparse*, that is, for each column M_j, every block
$$M_j\left((\tau-1)(2^b-1)+1 : \tau(2^b-1)\right),\ \forall \tau \in \left[\frac{T}{2^b-1}\right]$$
contains at most one 1.

With the (2^b-1) block-wise sparse structure, the τ-th bin of clients select their channel matrices as

$$W_i\left(1:2^b-1\middle|j\right) = M_j\left((\tau-1)\cdot(2^b-1)+1 : \tau\cdot(2^b-1)\right),\ \forall j \in [d], \tau \in \left[\frac{T}{2^b-1}\right]. \tag{8}$$

Notice that 1) $W_i(1:2^b-1|j)$ determines $W_i(2^b|j)$, and 2) the (2^b-1) block-wise sparse structure ensures that W_i's are valid channels.

To find a measurement matrix M that satisfies the above three criterion, we use the celebrated Kautz and Singleton's construction (Kautz and Singleton, 1964). This construction uses a $[m,k]_q$ Reed-Solomon code as an outer code C_{out} and the identity code C_{in} (i.e. one-hot encoding $I_q : [q] \to \{0,1\}^{q\times 1}$) as the inner code, and the measurement matrix is the concatenation of C_{out} and C_{in}: $M_{\mathsf{KS}} \triangleq C_{\mathsf{out}} \circ C_{\mathsf{in}} \in \{0,1\}^{mq\times q^k}$.

For NCGT, we pick $m = q$ (so $T = q^2$ and $d = q^k$) and set the rate $\frac{k}{m} = \frac{1}{s+1}$ to ensure that M_{KS} is s-disjunct. Thus by selecting $q = \Theta\left(s\log d\right)$, M_{KS} satisfies Property 1 and is $\Theta\left(s\log d\right)$-blockwise sparse (so Property 3 holds for all $b \leq \log\left(s\log d\right)$), with $T_{\mathsf{KS}} = O\left(s^2\log^2 d\right)$ rows. For more details on Kautz and Singleton's construction, we refer the reader to (Inan et al., 2019; Indyk et al., 2010). By adopting M_{KS} as the encoding channel matrix (as described in (8)), we extend the previous scheme to the b-bit setting (see Algorithm 9 and Algorithm 10 in Section D for the details).

Lemma 14 *Under the above encoding and decoding schemes (i.e. Algorithm 9, 10) with measurement matrix M_{KS}, we have*

$$\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} \leq \exp\left(-C_0 \cdot \frac{n_1 2^b \alpha}{s^2\log^2 d} + (\log s + \log\log d)\right).$$

In addition, $|\hat{\mathcal{J}}_\alpha| \leq s$ with probability 1, and the decoding complexity is $O(n2^b + s^2 d\log^2 d)$.

Picking $n_1 = \frac{n}{2}$, $\alpha = \frac{1}{\sqrt{n2^b}}$, and combining Lemma 7 and Lemma 14, we arrive at (12) as long as $n \succeq \frac{s^4\log^4(d/s)}{2^b}\left(\log s + \log\log d\right)^2$. This proves the second part of Theorem 1.

Remark 15 *In Algorithm 10, we present a naive NCGT cover decoder, which takes $O\left(d\cdot T_{\mathsf{KS}}\right) = O(s^2 d\log^2 d)$ time. However, since M_{KS} is constructed based on Reed-Solomon codes, one can leverage the efficient list recovery algorithm (i.e. Guruswami-Sudan algorithm (Guruswami and Sudan, 1998)) to decode $\hat{\mathcal{J}}_\alpha$ in $\mathsf{poly}(s, \log d)$ time, improving the dependency on d from $\mathsf{poly}(d)$ to $\mathsf{poly}(\log d)$.*

Removing the use of shared randomness In Scheme C, the local encoding functions are constructed deterministically according to M_{KS} and hence do not involve any randomization, so the only use of shared randomness is in the estimation phase. However, we can circumvent it by considering the following two-round encoding scheme: the second n_2 clients encode their observations according to $\hat{\mathcal{J}}_\alpha\left(Y^{n_1}\right)$, which can be done, for instance, by using the grouping idea in Algorithm 4, but now we only group symbols in $\hat{\mathcal{J}}_\alpha$. Unlike the interactive scheme in the next section (i.e. Scheme D) that requires $\log d$ rounds of interaction, the resulting scheme involves only two-round interaction and no longer needs shared randomness.

5. Interactive Localization Scheme: Tree-based Approach

Unlike non-interactive localization schemes introduced in the previous section, if we allow for sequential interaction between the server and clients, we can localize $\mathcal{J}_\alpha$ more efficiently (i.e. using less samples) and obtain a smaller requirement on the sample size (as described in Theorem 2). We briefly introduce a tree-based $\log d$-round interactive localization scheme below. We refer the reader to Section C for the details of the tree-based interactive scheme.

Sketch of the scheme We represent each symbol $j \in [d]$ by a $\log d$ bits binary string, and our algorithm discovers elements in $\mathcal{J}_\alpha$ by learning the prefixes of their bit representations sequentially across $\log d$ rounds. In particular, at each round t, the goal is to estimate the set of all length-t prefixes in $\mathcal{J}_\alpha$, which we denote by $\mathcal{J}_{\alpha,t} \triangleq \left\{\mathsf{prefix}_t(j) \,|\, j \in \mathcal{J}_\alpha\right\}$ (so $\mathcal{J}_\alpha = \mathcal{J}_{\alpha,\log d}$).

Towards this goal, we first partition n_1 clients into $\log d$ equal-sized groups $\mathcal{G}_1, ..., \mathcal{G}_{\log d}$ with clients in $\mathcal{G}_t$ participating in round t. At round t, clients encode their observations according to $\hat{\mathcal{J}}_{\alpha,t-1}$, where $\hat{\mathcal{J}}_{\alpha,t-1}$ is an estimate of $\mathcal{J}_{\alpha,t-1}$ obtained from the previous round. The encoding rule is based on the grouping idea described in Scheme A, but since now each client has partial knowledge $\hat{\mathcal{J}}_{\alpha,t-1}$, they only group symbols whose prefixes lie in $\hat{\mathcal{J}}_{\alpha,t-1}$ (instead of grouping the entire $[d]$). This leads to a more efficient way to use the samples and improves the sample size requirement from $O(d^2\log^2 d/2^b)$ to $\tilde{O}(s^2\log^2 d/2^b)$.

Upon observing $\mathcal{G}_t$'s reports, the system (i.e. all subsequent clients and the server) updates $\hat{\mathcal{J}}_{\alpha,t-1}$ accordingly to generate $\hat{\mathcal{J}}_{\alpha,t}$. When n_1 is large enough, this protocol successfully localizes $\mathcal{J}_\alpha$ with high probability. Indeed, we have the following bound on the probability of error:

Lemma 16 *Under the above encoding and decoding rules, the failure probability is bounded by*

$$\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} \leq \exp\left(-\frac{n_1(2^b-1)\alpha}{2s\log d} + \left(\log\log d + \log\left(\frac{2s}{2^b-1}\right)\right)\right).$$

Picking $n_1 = \frac{n}{2}$, $\alpha = \frac{1}{\sqrt{n2^b}}$ and combining Lemma 16 and Lemma 7, we conclude that as long as $n \succeq \frac{s^2\log^2 d}{2^b}\left(\log s + \log\log d\right)^2$, (1) holds. This establishes Theorem 2.

6. Concluding Remarks and Open Problems

In this work, we characterize the convergence rate of estimating s-sparse distributions, showing that by carefully designing encoding and decoding schemes, one can achieve a dimension-free estimation rate, which is the same rate as knowing the sparse support of the distribution beforehand. As a natural next step, we study the threshold on number of samples needed to achieve such a dimension-free estimation rate. We give upper bounds on this threshold by developing three non-interactive schemes and one interactive scheme. Our results establish an interesting connection between distribution estimation and group testing, suggesting that non-adaptive group testing can be useful in designing efficient decoding and encoding schemes with small sample complexity.

There are several open research directions emerging from our work. First, under the non-interactive model, there exists a gap between our upper and lower bounds on the minimum sample size required $n^*(s, b, d)$ to achieve dimension-free convergence (see the discussion after Corollary 3 for more details). We conjecture that the lower bound on $n^*(s, b, d)$ is tight, and the non-interactive schemes can be further improved. Closing this gap remains an open problem. Second, we note that in the estimation phase of our two-stage scheme, we rely on random hash functions to encode local data. It remains unclear whether or not there exists private-coin schemes that achieve the same estimation error.

Finally, we note an interesting contrast with sparse distribution estimation under local differential privacy (LDP) constraints. There are several recent works in the literature that observe a symmetry between LDP and communication constraints (Chen et al., 2020; Acharya et al., 2019a,b; Han et al., 2018b; Barnes et al., 2019, 2020). In particular, without the sparse assumption, previous works (Barnes et al., 2019, 2020; Acharya et al., 2019a; Han et al., 2018b) show that the minimax ℓ_2 estimation error under b-bit and ε-local differential privacy (LDP) constraints are $\Theta\left(\frac{d}{n2^b}\right)$ and $\Theta\left(\frac{d}{ne^\varepsilon}\right)$ (for $\varepsilon = \Omega(1)$). This implies that in the non-sparse case, compression and LDP have the same sample complexity as long as $b \approx \varepsilon$. This symmetry between communication and LDP constraints has also been observed in other statistical models such as mean estimation.

To the best of our knowledge, the result we derive in this paper is the first to break the symmetry between communication and privacy constraints in distributed estimation. Under the s-sparse assumption, (Acharya et al., 2021) shows that $\Theta\left(\frac{s\log(d/s)}{ne^\varepsilon}\right)$ error is fundamental under ε-LDP. Given the symmetry between communication and LDP constraints in previous results as mentioned earlier, one might have been tempted to expect the error under a b-bit constraint to be of the form $O\left(\frac{s\log(d/s)}{n2^b}\right)$. Our results suggest that one can achieve $\Theta\left(\frac{s}{n2^b}\right)$ error under a b-bit constraint, implying that when estimating sparse distributions, the communication constraint and the LDP constraint behave differently and the latter is strictly more stringent than the former. This loss of symmetry makes it difficult, for example, to postulate the fundamental limit (and how to achieve it) under joint communication and LDP constraints, a direction that has been settled in (Chen et al., 2020) in the non-sparse case. Understanding the convergence rate for sparse distribution estimation under joint communication and LDP constraints remains an open problem. Further, exploring interactions with secure aggregation is of practical interest, especially in the federated learning and analytics settings.

Acknowledgments

The authors would like to thank Professor Mary Wootters for pointing out the efficient decoding algorithm for Kautz and Singleton's group testing scheme, which helped us to improve the time complexity of the localization scheme.

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

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

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints ii: Communication constraints and shared randomness. *arXiv preprint arXiv:1905.08302*, 2019b.

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. General lower bounds for interactive high-dimensional estimation under information constraints. *arXiv preprint arXiv:2010.06562*, 2020.

Jayadev Acharya, Peter Kairouz, Yuhan Liu, and Ziteng Sun. Estimating sparse discrete distributions under local privacy and communication constraints. In *Algorithmic Learning Theory*. PMLR, 2021.

Richard E Barlow. Statistical inference under order restrictions; the theory and application of isotonic regression. Technical report, 1972.

Leighton Pate Barnes, Yanjun Han, and Ayfer Ozgur. Lower bounds for learning distributions under communication constraints via fisher information, 2019.

Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Ozgur. Fisher information under local differential privacy. *arXiv preprint arXiv:2005.10783*, 2020.

Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 127–135, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746632. URL https://doi.org/10.1145/2746539.2746632.

Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Thakurta. Practical locally private heavy hitters. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 2285–2293, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. *ACM Transactions on Algorithms (TALG)*, 15(4):1–40, 2019.

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

Wei-Ning Chen, Peter Kairouz, and Ayfer Ozgur. Breaking the communication-privacy-accuracy trilemma. *Advances in Neural Information Processing Systems*, 33, 2020.

L. Devroye and L. Gábor. *Nonparametric Density Estimation: The L1 View*. Wiley Interscience Series in Discrete Mathematics. Wiley, 1985. ISBN 9780471816461. URL https://books.google.com.tw/books?id=ZVALbrjGpCoC.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer New York, 2012. ISBN 9781461301257. URL https://books.google.com.tw/books?id=1QEMCAAAQBAJ.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

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

Dingzhu Du, Frank K Hwang, and Frank Hwang. *Combinatorial group testing and its applications*, volume 12. World Scientific, 2000.

Arkadii Georgievich D'yachkov and Vladimir Vasil'evich Rykov. Bounds on the length of disjunctive codes. *Problemy Peredachi Informatsii*, 18(3):7–13, 1982.

Venkatesan Guruswami and Madhu Sudan. Improved decoding of reed-solomon and algebraic-geometric codes. In *Proceedings 39th Annual Symposium on Foundations of Computer Science (Cat. No. 98CB36280)*, pages 28–37. IEEE, 1998.

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

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. *arXiv preprint arXiv:1802.08417*, 2018b.

Huseyin A Inan, Peter Kairouz, Mary Wootters, and Ayfer Özgür. On the optimality of the kautz-singleton construction in probabilistic group testing. *IEEE Transactions on Information Theory*, 65(9):5592–5603, 2019.

Piotr Indyk, Hung Q Ngo, and Atri Rudra. Efficiently decodable non-adaptive group testing. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 1126–1142. SIAM, 2010.

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.

William Kautz and Roy Singleton. Nonrandom binary superimposed codes. *IEEE Transactions on Information Theory*, 10(4):363–377, 1964.

Hung Q Ngo and Ding-Zhu Du. A survey on combinatorial group testing algorithms with applications to dna library screening. *Discrete mathematical problems with medical applications*, 55: 171–182, 2000.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

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

Wennan Zhu, Peter Kairouz, Brendan McMahan, Haicheng Sun, and Wei Li. Federated heavy hitters discovery with differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 3837–3847. PMLR, 2020.

Appendix A. Estimation stage: random hashing

Encoding For the second group of clients $i \in [n_1+1:n]$, they report their local samples through n_2 independent random hash functions $h_{n_1+1}, ..., h_n$(which are generated via shared randomness). Equivalently, client i's encoding channel $W_i(y|x)$ is constructed as follows: each column of the channel matrix W_i can be viewed as the one-hot representation of $L_{i,x}$, where

$$L_{i,x} \overset{\text{i.i.d.}}{\sim} \mathsf{uniform}\left(2^b\right), \; \forall i \in [n_1+1:n],\, x \in [d].$$

Formally,

$$W_i(\cdot|x) \triangleq \left[\mathbb{1}_{\{L_{i,x}=1\}}, ..., \mathbb{1}_{\left\{L_{i,x}=2^b\right\}}\right]^{\mathsf{T}}.$$

Note that since $W_i \in \{0,1\}^{2^b \times d}$, the local encoder is *deterministic*, so we can also write $Y_i = h_i\left(X_i\right)$ for some deterministic hash function $h_i(x)$.

Decoding Upon obtaining $\hat{\mathcal{J}}_\alpha$ from the first stage and receiving Y^{n_2}[6], the server computes count on each symbol j $N_j(Y^{n_2}) \triangleq |\{i \in [n_1+1:n] : h_i(j) = Y_i\}|$. Note that $\mathbb{P}\left\{h_i(j) = Y_i\right\} = \frac{p_j\left(2^b-1\right)+1}{2^b} \triangleq b_j$. The final estimator $\hat{p}_j\left(Y^{n_2}\right)$ is then defined as

$$\hat{p}_j\left(Y^{n_2}\right) = \begin{cases} \frac{N_j\left(2^b-1\right)}{n_2 \cdot 2^b} - \frac{1}{2^b}, & \text{if } j \in \hat{\mathcal{J}}_\alpha \\ 0, & \text{else.} \end{cases} \tag{9}$$

We summarize the encoding algorithm in Algorithm 2. The decoding algorithm (Algorithm 3) can be found in Section D.

Bounds on the estimation error Let $\mathcal{E}_g$ be the event that the localization phase succeeds, i.e. $\mathcal{E}_g \triangleq \left\{\mathcal{J}_\alpha \subseteq \hat{\mathcal{J}}_\alpha\right\}$. Then conditioned on $\mathcal{E}_g$, we can control the ℓ_2 estimation error as follows:

Lemma 17 *Let $\mathcal{E}_g$ and $\hat{p}$ be defined as above. Then conditioned on $\mathcal{E}_g$, we have*

$$\mathbb{E}\left[\sum_{j\in[d]} \left(\hat{p}_j\left(Y^{n_2}\right) - p_j\right)^2 \middle| \mathcal{E}_g\right] \leq s\alpha^2 + \frac{s}{n_2 2^b} + \frac{1}{n_2}.$$

Proof The proofs of Lemma 17 can be found in Section E. ■

Appendix B. Localization scheme A: uniform grouping

Under b-bit communication constraint, a straightforward encoding approach is based on the grouping idea (Han et al., 2018a), where each client only encodes symbols in a pre-specified subset of $[d]$ and ignores others. The subset assigned to each client contains $2^b - 1$ symbols, so the encoded message can be described by b bits.

[6] With a slight abuse of notation, we use Y^{n_2} to denote the collection of $(Y_{n_1+1}, ..., Y_n)$.

Encoding We partition the first n_1 clients into $M = d/(2^b - 1)$ equal-sized groups $\mathcal{G}_1, ..., \mathcal{G}_M$ such that $|\mathcal{G}_m| = n_1\left(2^b - 1\right)/d$ and $\mathcal{G}_m = \{i \in [n_1] | \, i \equiv m \pmod{M}\}$. For client $i \in \mathcal{G}_m$, they only reports information about $j \in \left[(m-1)\cdot(2^b-1)+1 : m\cdot(2^b-1)\right]$, i.e. symbols that lie in the m-th block of $[d]$. Equivalently, for clients in group m, their encoding channel matrices are

$$W_i = \left[\begin{array}{ccc|c|ccc} e_{2^b} & \cdots & e_{2^b} & \underbrace{e_1 \quad e_2 \quad \cdots \quad e_{2^b-1}}_{(m-1)(2^b-1)+1\,:\,m(2^b-1)} & e_{2^b} & \cdots & e_{2^b} \end{array}\right], \tag{10}$$

where $e_\ell \in \{0,1\}^{2^b \times 1}$ is the ℓ-th coordinate vector.

Decoding Due to our construction of encoding functions, we see that as long as

$$Y_i \neq 0 \Longleftrightarrow X_i \in \left[(m-1)\cdot(2^b-1)+1 : m\cdot(2^b-1)\right]$$

(recall that m is the index of the block containing i), the server can specify X_i upon observing Y_i by computing $X_i = (m-1)\cdot(2^b-1)+Y_i$. Therefore, defining

$$\hat{X}\left(Y_i\right) \triangleq= \begin{cases} X_i, \text{ if } X_i \in \left[(m-1)\cdot(2^b-1)+1 : m\cdot(2^b-1)\right] \\ \texttt{null}, \text{ otherwise,} \end{cases}$$

then we can estimate $\mathcal{J}_\alpha$ by $\hat{\mathcal{J}}_\alpha \triangleq \left\{\hat{X}\left(Y_i\right) \middle| i \in [n_1]\right\}$, that is, $\hat{\mathcal{J}}_\alpha$ is the collection of all observed and successfully decoded symbols from the first n_1 clients. The details of the encoding and decoding algorithms are given in Algorithm 4 and Algorithm 5 in Section D.

Appendix C. Tree-based Interactive Scheme (Detailed)

Encoding Let $\hat{\mathcal{J}}_{\alpha,0} = \emptyset$. At round $t > 0$, let $\mathcal{G}_t \triangleq \{i \in [n_1] \,|\, i \equiv t(\mathsf{mod}\ \log d)\}$ be the participated clients, and let $\mathsf{C}\left(\hat{\mathcal{J}}_{\alpha,t-1}\right)$ be the set of all candidates of length-t prefixes, that is,

$$\mathsf{C}\left(\hat{\mathcal{J}}_{\alpha,t-1}\right) \triangleq \left\{\mathsf{append}(v,0), \mathsf{append}(v,1) \,\middle|\, v \in \hat{\mathcal{J}}_{\alpha,t-1}\right\}.$$

If round $t-1$ succeeds, then $\left|\mathsf{C}\left(\hat{\mathcal{J}}_{\alpha,t-1}\right)\right| \leq 2\cdot s$. We then partition $\mathsf{C}\left(\hat{\mathcal{J}}_{\alpha,t-1}\right)$ into $M \triangleq \frac{\left|\mathsf{C}\left(\hat{\mathcal{J}}_{\alpha,t-1}\right)\right|}{2^b-1}$ blocks $\mathcal{B}_1, ..., \mathcal{B}_M$, each contains $2^b - 1$ possible length-t prefixes. We also partition $\mathcal{G}_t$ into M groups $\mathcal{K}_{t,1}, ..., \mathcal{K}_{t,M}$ by setting $\mathcal{K}_{t,m} \triangleq \left\{i \in \mathcal{G}_t \,\middle|\, \frac{i-t}{\log d} \equiv m(\mathsf{mod}\ M)\right\}$. When client i in $\mathcal{K}_{t,m}$ observes X_i with $\mathsf{prefix}_t(X_i) \in \mathcal{B}_m$, it reports the index of $\mathsf{prefix}_t(X_i)$ in $\mathcal{B}_m$ (where we index elements in $\mathcal{B}_m$ by 1 to 2^b-1), and otherwise it reports 0. Then the message can be encoded in b bits. Formally, we have

$$Y_i = \begin{cases} \mathsf{index}\left(\mathcal{B}_m, \mathsf{prefix}_t(X_i)\right), & \text{if } \mathsf{prefix}_t(X_i) \in \mathcal{B}_m, \\ 0, & \text{else.} \end{cases}$$

Decoding Since each client is assigned to each group $\mathcal{K}_{t,m}$ deterministically, from the client's index i and $\hat{\mathcal{J}}_{\alpha,t-1}$, one can compute the group index $(t(i), m(i))$ explicitly. Therefore, upon observing Y_i, we can estimate $\mathsf{prefix}_t(X_i)$ by

$$\hat{\mathsf{prf}}_t(Y_i) \triangleq \begin{cases} \text{the } Y_i + (m(i)-1)\cdot\left(2^b-1\right) \text{-th element in } \mathsf{C}\left(\hat{\mathcal{J}}_{\alpha,t-1}\right), & \text{if } Y_i \neq 0, \\ \texttt{null}, & \text{else.} \end{cases}$$

Finally, we estimate $\mathcal{J}_{\alpha,t}$ by $\hat{\mathcal{J}}_{\alpha,t} \triangleq \left\{\hat{\mathsf{prf}}_t(Y_i) \,\middle|\, i \in \mathcal{G}_t\right\}$.

Appendix D. Algorithms

Algorithm 3: Decoding for estimation phase

```
Input: $Y^{n_2}$, $b \in \mathbb{N}$, $\hat{\mathcal{J}}_\alpha$
Output: $\hat{p}$
Set $\hat{p} = \mathbf{0}$;
for $j \in \hat{\mathcal{J}}_\alpha$ do
    $N_j(Y^{n_2}) \triangleq |\{i \in [n_1+1:n] : h_i(j) = Y_i\}|$;
end
for $j \in \hat{\mathcal{J}}_\alpha$ do
    $\hat{p}_j\left(Y^n\right) = \frac{N_j\left(2^b-1\right)}{n_2 \cdot 2^b} - \frac{1}{2^b}$;
end
return $\hat{p}$
```

Algorithm 4: Localization via uniform grouping: encoding

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

Algorithm 5: Localization via uniform grouping: decoding

```
Input: $Y^{n_1} \in [2^b]$
Output: $\hat{\mathcal{J}}_\alpha$
Initialize $\hat{\mathcal{J}}_\alpha = \emptyset$;
Compute $M = d/\left(2^b - 1\right)$;
for $i \in [n_1]$ do
    $m \leftarrow i \,\mathsf{mod}\, M$;
    if $Y_i \neq 0$ then
        Add $\hat{X} \triangleq m \cdot \left(2^b - 1\right) + Y_i$ into $\hat{\mathcal{J}}_\alpha$;
    end
end
return $\hat{\mathcal{J}}_\alpha$
```

Algorithm 6: Localization via random hash: decoding

```
Input: $Y^{n_1} \in [2^b]$
Output: $\hat{\mathcal{J}}_\alpha$
Initialize consist $\leftarrow$ True;
Let $\mathcal{C}_1, ..., \mathcal{C}_N$ be an enumerate of all $N \triangleq \binom{d}{s}$ size-$s$ subsets of $[d]$;
for $\ell \in [N]$ do
    consist $\leftarrow \bigwedge_{i\in[n_1]} \bigvee_{j\in\mathcal{C}_\ell} \mathbb{1}_{\{Y_i = h_i(j)\}}$;                    // check consistency
    if consist then
        $\hat{\mathcal{J}}_\alpha \leftarrow \mathcal{C}_\ell$;
        break;
    end
end
return $\hat{\mathcal{J}}_\alpha$
```

Algorithm 7: Localization via NCGT: 1-bit encoding

```
Input: $X_i \in [d]$, $M \in \{0,1\}^{T\times d}$
$t(i) \leftarrow i \,\mathsf{mod}\, T$;
$Y_i \leftarrow M_{t(i),X_i}$;
return $Y_i$
```

Algorithm 9: Localization via NCGT: b-bit encoding

```
Input: $X_i \in [d]$, $M_{\mathsf{KS}} \in \{0,1\}^{Cs^2\log^2 d\times d}$
$t \leftarrow i \,\mathsf{mod}\, \lceil \frac{T}{2^b-1} \rceil$;
$Y_i \leftarrow M_{(t-1)(2^b-1)+1:t(2^b-1),X_i}$;                    // Can be represented in $b$ bits
return $Y_i$
```

Algorithm 8: Localization via NCGT: 1-bit decoding

```
Input: $Y^{n_1}$, $M$
Initialize $\hat{\mathcal{J}}_\alpha = \emptyset$;
for $\tau \in [T]$ do
    $\hat{Z}_\tau \leftarrow \bigvee_{i:i\equiv\tau(\bmod T)} Y_i$ ;          // simulate the $\tau$-th group test
end
$\boldsymbol{Z} \leftarrow \left[\hat{Z}_1, ..., \hat{Z}_T\right]^\intercal$;
for $j \in [d]$ ;                                      // run the cover decoder
 do
    if $\mathsf{supp}\left(M_j\right) \subseteq \mathsf{supp}\left(\boldsymbol{Z}\right)$ then
        Add $j$ to $\hat{\mathcal{J}}_\alpha$
    end
end
return $\hat{\mathcal{J}}_\alpha$
```

Algorithm 10: Localization via NCGT: b-bit decoding

```
Input: $Y^{n_1}$, $M_{\mathsf{KS}}$
Initialize $\hat{\mathcal{J}}_\alpha = \emptyset$;
for $\tau \in \lceil \frac{T}{2^b-1} \rceil$ do
    for $\kappa \in [2^b - 1]$ do
        $\hat{Z}_\tau(\kappa) \leftarrow \bigvee_{i:i\equiv\tau\left(\bmod \lceil \frac{T}{2^b-1} \rceil\right)} Y_i(\kappa)$ ;    // simulate the $\tau$-th group test
    end
end
$\boldsymbol{Z} \leftarrow \left[\hat{Z}_1, ..., \hat{Z}_T\right]^\intercal$;
for $j \in [d]$ ;                                      // run the cover decoder
 do
    if $\mathsf{supp}\left(M_{\mathsf{KS}}(j)\right) \subseteq \mathsf{supp}\left(\boldsymbol{Z}\right)$ then
        Add $j$ to $\hat{\mathcal{J}}_\alpha$
    end
end
return $\hat{\mathcal{J}}_\alpha$
```

Appendix E. Missing Proofs

E.1. Proof of Lemma 17

Proof Notice that $N_j \sim \mathsf{Binom}\left(n_2, b_j\right)$, so for all $j \in \hat{\mathcal{J}}_\alpha$, $\hat{p}_j\left(Y^n\right)$ yields an unbiased estimator on p_j. Moreover,

$$\mathbb{E}\left[\left(\hat{p}_j - p_j\right)^2 \middle| j \in \hat{\mathcal{J}}_\alpha\right] = \mathsf{Var}\left(\hat{p}_j \middle| j \in \hat{\mathcal{J}}_\alpha\right) \leq \frac{1}{n_2^2}\mathsf{Var}\left(\mathsf{Binom}\left(n_2, b_j\right)\right) = \frac{b_j}{n_2}.$$

Summing over j, we obtain

$$\sum_{j \in \hat{\mathcal{J}}_\alpha} \mathbb{E}\left[\left(\hat{p}_j - p_j\right)^2 \middle| j \in \hat{\mathcal{J}}_\alpha\right] \leq \frac{1 + \frac{s}{2^b}}{n_2},$$

with probability 1 since $\left|\hat{\mathcal{J}}_\alpha\right| \leq s$ almost surely. This implies

$$\mathbb{E}\left[\sum_{j \in \mathcal{J}_\alpha} \left(\hat{p}_j - p_j\right)^2 \middle| \mathcal{E}_g\right] \leq \frac{s + 2^b}{n_2 2^b}.$$

Together with the fact that

$$\mathbb{E}\left[\sum_{j \notin \mathcal{J}_\alpha} p_j^2 \middle| \mathcal{E}_g\right] \leq s\alpha^2,$$

we establish Lemma 17. ■

E.2. Proof of Lemma 7

Proof Observe that

$$\begin{aligned}
\mathbb{E}\left[\sum_{j \in [d]} \left(\hat{p}_j - p\right)^2\right] &= \mathbb{P}\left\{\mathcal{E}_g^c\right\} \cdot \mathbb{E}\left[\sum_{j \in [d]} \left(\hat{p}_j\left(Y^n\right) - p_j\right)^2 \middle| \mathcal{E}_g^c\right] + \mathbb{P}\left\{\mathcal{E}_g\right\} \cdot \mathbb{E}\left[\sum_{j \in [d]} \left(\hat{p}_j\left(Y^n\right) - p_j\right)^2 \middle| \mathcal{E}_g\right] \\
&\leq 2\mathbb{P}\left\{\mathcal{E}_g^c\right\} + \mathbb{E}\left[\sum_{j \in [d]} \left(\hat{p}_j\left(Y^n\right) - p_j\right)^2 \middle| \mathcal{E}_g\right] \\
&\leq 2\mathbb{P}\left\{\mathcal{E}_g^c\right\} + s\alpha^2 + \frac{s}{n_2 2^b} + \frac{1}{n_2},
\end{aligned}$$

where the last inequality is due to Lemma 17. ■

E.3. Proof of Lemma 8

Proof If $j \in \left[(m-1)\cdot(2^b-1)+1 : m\cdot(2^b-1)-1\right] \cap \mathcal{J}_\alpha$, symbol j can be successfully included in $\hat{\mathcal{J}}_\alpha$ only when at least one of clients in $\mathcal{G}_m$ observes it. Therefore, let $\mathcal{E}_j \triangleq \{X_j \neq j, \forall i \in \mathcal{G}_m\}$ be the event of failing to include symbol j, then

$$\mathbb{P}\left\{\hat{\mathcal{J}}_\alpha \subseteq \mathcal{J}_\alpha\right\} = 1 - \mathbb{P}\left\{\bigcup_{j\in\mathcal{J}_\alpha} \mathcal{E}_j\right\} \geq 1 - |\mathcal{J}_\alpha|\cdot(1-\alpha)^{n'} \geq 1 - s\cdot e^{-n'\alpha},$$

where $n' \triangleq |\mathcal{G}_m| = \frac{n_1(2^b-1)}{d}$. ■

E.4. Proof of Lemma 10

Proof Let $\mathcal{E}_\ell$ be the event such that 1) $\mathcal{J}_\alpha \not\subset \mathcal{C}_\ell$ for some $\ell \in [N]$ and 2) $\mathcal{C}_\ell$ is consistent with (W^n, Y^n). Then the error $\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\ell$ can happen only if $\bigcup_{\ell\in[N]} \mathcal{E}_\ell$ occurs. Hence it suffices to control the probability of $\mathcal{E}_\ell$ and then apply union bound.

To bound $\mathcal{E}_\ell$, denote $j \in \mathcal{J}_\alpha \setminus \mathcal{C}_\ell$. Note that as long as the server ensures a client observes symbol j, they can rule out $\mathcal{C}_\ell$ from the candidates set. However, since each client only reports the hash value of their observation, the server cannot directly obtain such information. To address the difficulty, the following definition describes the condition that makes a channel W "good" with respect to $\mathcal{C}_\ell$ and j:

Definition 18 *We call a channel W* distinguishes *j under H_ℓ, if for all $j' \in \mathcal{C}_\ell$, we have $W(\cdot|j') \neq W(\cdot|j)$*[7].

Then $\mathcal{E}_\ell$ cannot happen if there exist a client i 1) who observes j and 2) whose channel W_i distinguishes j. Notice that due to our construction of localization channels,

$$\begin{aligned}
\mathbb{P}\left\{W_i \text{ distinguishes } j|H_\ell\right\} &= \mathbb{P}\left\{\forall j' \in \mathcal{C}_\ell,\, W_i\left(\cdot|j\right) \neq W_i\left(\cdot|j'\right)\right\} \\
&\overset{\text{(a)}}{=} \sum_{y\in[2^b]} \mathbb{P}\left\{L_{i,j} = y\right\}\cdot \prod_{j'\in\mathcal{C}_\ell} \mathbb{P}\left\{L_{i,j} \neq y\right\} \\
&\geq \frac{2^b-1}{s}\cdot\left(1-\frac{1}{s}\right)^s \\
&\overset{\text{(b)}}{\geq} \frac{2^b-1}{4s},
\end{aligned} \tag{11}$$

where $L_{i,j}$ in (a) is defined in (5), and (b) is due to the fact that $f(s) \triangleq \left(1-\frac{1}{s}\right)^s$ increasing in s and $s \geq 2$. We also have $\mathbb{P}\left\{X_i = j\right\} \geq \alpha$, and since we generate W_i independently,

$$\mathbb{P}\left\{\{X_i = j\} \cap \{W_i \text{ distinguishes } j\}\right\} \geq \alpha\cdot\frac{2^b-1}{4s}.$$

Thus we can upper bound the probability of error by

$$\mathbb{P}\left\{\mathcal{E}_\ell\right\} \leq \left(1-\alpha\cdot\frac{2^b-1}{4s}\right)^{n_1} \leq \exp\left(-n_1\alpha\frac{2^b-1}{4s}\right).$$

[7] Note that due to our construction, W is *deterministic*, i.e. $W(\cdot|j) = e_l$ for some $l \in [2^b]$.

Finally applying union bound, we arrive at

$$\mathbb{P}\left\{\bigcup_{\ell\in[N]}\mathcal{E}_\ell\right\} \leq \binom{d}{s}\cdot\exp\left(-n_1\alpha\frac{2^b-1}{4s}\right) \leq \exp\left(-n_1\alpha\frac{2^b-1}{4s}+C_0 s\log\left(\frac{d}{s}\right)\right),$$

establishing the lemma. ■

E.5. Proof of Lemma 11

Proof

Observe that the condition

$$n \geq 4\cdot f_1^2\cdot\max\left(f_2^2, 16\log^2(f_1)\right) \tag{12}$$

implies

$$\frac{n}{4}\geq f_1^2\cdot f_2^2 \Longrightarrow \frac{\sqrt{n}}{2}\geq f_1\cdot f_2.$$

(12) also implies

$$n\geq 64\cdot f_1^2\cdot\log(f_1)^2 \overset{\text{(b)}}{\Longrightarrow} \frac{\sqrt{n}}{\log n}\geq\frac{8\cdot f_1\cdot\log f_1}{2\log f_1+2\log\log f_1+2\log 64}\geq 2f_1 \Longrightarrow \frac{\sqrt{n}}{2}\geq f_1\cdot\log n,$$

where (b) holds since 1) $\frac{\sqrt{n}}{\log n}$ is increasing when $n\geq 10$ (note that by assumption, $f_1\geq 300$, so $4\cdot f_1^2\cdot\log f_1^2 > 10$), and 2) $\log f_1+\log\log f_1+\log 64\leq 2\log f_1$ when $f_1\geq 300$. Therefore we have

$$\sqrt{n}\geq f_1\cdot\log n+f_1\cdot f_2 \Longrightarrow \exp\left(-\frac{\sqrt{n}}{f_1}+f_2\right)\leq -\frac{1}{n}.$$

■

E.6. Proof of Lemma 13

Proof Recall that $\mathcal{G}_\tau$ is the τ-th bin of clients. Notice that as long as the decoding succeeds (i.e. $\left\{\mathcal{J}_\alpha\not\subset\hat{\mathcal{J}}_\alpha\right\}$) if for all $\tau\in[T]$, all symbols in $\mathcal{J}_\alpha$ appear at least once in $\mathcal{G}_\tau$'s observation. Let $\mathcal{E}_\tau$ denotes the error event $\mathcal{E}_\tau\triangleq\{\mathcal{J}_\alpha\not\subset\{X_i|i\in\mathcal{G}_\tau\}\}$. Then we have

$$\mathbb{P}\left\{\mathcal{J}_\alpha\not\subset\hat{\mathcal{J}}_\alpha\right\}\leq\mathbb{P}\left\{\bigcup_{\tau\in[T]}\mathcal{E}_\tau\right\}\leq\sum_{\tau\in[T]}\mathbb{P}\left\{\mathcal{E}_\tau\right\},$$

so it suffices to lower bound the probability of $\mathcal{E}_\tau$.

Recall that for each symbol $j\in\mathcal{J}_\alpha$, $\mathbb{P}\{X_i=j\}\geq\alpha$, and we also have $|\mathcal{J}_\alpha|\leq s$. Therefore by union bound,

$$\mathbb{P}\left\{\mathcal{E}_\tau\right\}\leq|\mathcal{J}_\alpha|\cdot(1-\alpha)^{\frac{n_1}{T}}\leq\exp\left(-\frac{n_1\alpha}{T}+\log s\right).$$

Finally, applying union bound on $\tau\in[T]$ again, we arrive at the desired bound. ■

E.7. Proof of Lemma 14

Proof The proof is the same as in Lemma 13, except for replacing $|\mathcal{G}_\tau|$ from n_1/T to $n_1 2^b/T_{\mathsf{KS}}$. ■

E.8. Proof of Lemma 16

Proof Let $T \triangleq \log d$ and $\mathcal{E}_t$ be the event that round t succeeds for all $1 \leq t \leq T$. Then $\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} = \mathcal{E}_T^c$. By union bound, we have

$$\mathbb{P}\left\{\mathcal{E}_T^c\right\} \leq \mathbb{P}\left\{\mathcal{E}_1\right\} + \sum_{t=2}^{T} \mathbb{P}\left\{\mathcal{E}_t^c \cap \mathcal{E}_{t-1}\right\} \leq \sum_{t=1}^{T} \mathbb{P}\left\{\mathcal{E}_t^c | \mathcal{E}_{t-1}\right\}, \tag{13}$$

where $\mathcal{E}_0$ always holds since the length-0 prefix set is always empty. Therefore, it suffices to control $\mathbb{P}\left\{\mathcal{E}_t^c|\mathcal{E}_{t-1}\right\}$.

Note that according to our decoding rule, every element in $\hat{\mathcal{J}}_{\alpha,t}$ must be a length-t prefix of symbols in $\mathsf{supp}(p)$, so we always have $\left|\hat{\mathcal{J}}_{\alpha,t}\right| \leq |\mathsf{supp}(p)| \leq s$. Hence to bound the failure probability $\mathbb{P}\left\{\mathcal{E}_t^c|\mathcal{E}_{t-1}\right\}$, we only need to control $\mathbb{P}\left\{\mathcal{J}_{\alpha,t} \not\subset \hat{\mathcal{J}}_{\alpha,t} \middle| \mathcal{J}_{\alpha,t-1} \subseteq \hat{\mathcal{J}}_{\alpha,t-1}\right\}$.

If $\mathcal{J}_{\alpha,t-1} \subseteq \hat{\mathcal{J}}_{\alpha,t-1}$ holds, then $\mathcal{J}_{\alpha,t}$ must be contained in the candidate set $\mathsf{C}\left(\hat{\mathcal{J}}_{\alpha,t}\right)$, so the only way that $\left\{\mathcal{J}_{\alpha,t} \not\subset \hat{\mathcal{J}}_{\alpha,t}\right\}$ can happen is that there is some symbol in $\mathcal{B}_m \cap \mathcal{J}_{\alpha,t}$ that is not observed by $\mathcal{K}_{t,m}$. Therefore, we have

$$\begin{aligned}\mathbb{P}\left\{\mathcal{E}_t^c|\mathcal{E}_{t-1}\right\} &= \mathbb{P}\left\{\mathcal{J}_{\alpha,t} \not\subset \hat{\mathcal{J}}_{\alpha,t} \middle| \mathcal{J}_{\alpha,t-1} \subseteq \hat{\mathcal{J}}_{\alpha,t-1}\right\} \leq \mathbb{P}\left\{\exists j \in \mathcal{B}_m \text{ s.t. } \forall i \in \mathcal{K}_{m,t}\, X_i \neq j\right\}\\ &\leq |\mathcal{B}_m| \cdot (1-\alpha)^{|\mathcal{K}_{m,t}|} \leq \frac{2s}{2^b-1} \cdot \exp\left(-\frac{n_1 \alpha}{2s \log d}\right),\end{aligned} \tag{14}$$

where the last inequality follows from $|\mathcal{K}_{m,t}| = \frac{|\mathcal{G}_t|}{M} \geq n_1 \frac{2^b-1}{2s\log d}$. Plugging into (13), we obtain

$$\mathbb{P}\left\{\mathcal{E}_T^c\right\} \leq T \cdot \frac{2s}{2^b-1} \cdot \exp\left(-\frac{n_1\alpha}{2s\log d}\right) = \exp\left(-\frac{n_1(2^b-1)\alpha}{2s\log d} + \left(\log\log d + \log\left(\frac{2s}{2^b-1}\right)\right)\right).$$

■

E.9. Proof of Theorem 4

Proof We prove that if we first randomly permute all of n clients with shared randomness, then all of previous schemes apply to distribution-free setting. We start with introducing a few notation. Let $f_1, f_2, ..., f_d \in [n]$ be the empirical frequencies of each symbol, i.e. $f_j \triangleq n\pi_j$, and let σ be a n-permutation drawn uniformly at random from the permutation group $\mathcal{S}_n$ with shared randomness. We set $n_1 = n_2 = n/2$ in the two-stage generic scheme Algorithm 1, and use $F_1, ..., F_d$ to denote the empirical frequency of the second half of samples, i.e. $F_j \triangleq \sum_{i=n/2+1}^{n} \mathbb{1}_{\left\{X_{\sigma(i)}=j\right\}}$. Notice that F_j is a random variable (since σ is random) with hyper-geometric distribution $\mathsf{HG}\left(n, \frac{n}{2}, f_j\right)$. Finally, let $\Pi_j \triangleq \frac{2F_j}{n}$ be the empirical distribution of the second half of clients.

Let $\hat{\mathcal{J}}_\alpha\left(Y^{n_1}\right)$ be an estimate of $\mathcal{J}_\alpha \triangleq \{j \in [d] \,|\, \pi_j \geq \alpha\}$ and $\hat{\Pi}_j\left(Y_2^n\right)$ be an estimator of Π_j (both will be explicitly defined later), then the final estimator $\hat{p}_j$ is defined in the same way as (9), i.e.

$$\hat{\pi}_j \triangleq \hat{\Pi}_j \cdot \mathbb{1}_{\left\{j \in \hat{\mathcal{J}}_\alpha\right\}}.$$

Now, consider using the same estimation scheme defined in Section A (i.e. Algorithm 2 and Algorithm 3) with clients' index being replaced by $\sigma(i)$. Then $\hat{\Pi}_j\left(Y^{n_2}\right)$ is

$$\hat{\Pi}_j\left(Y^{n_2}\right) = \frac{2\cdot 2^b}{n\left(2^b-1\right)} \sum_{i=\frac{n}{2}+1}^{n} \mathbb{1}_{\{Y_i = h_i(j)\}} - \frac{1}{2^b}. \tag{15}$$

Notice that condition on σ, $\mathbb{1}_{\{Y_i=h_i(j)\}}$ follows distribution $\mathbb{1}_{\{X_i=j\}} + \mathbb{1}_{\{X_i\neq j\}} \cdot \mathsf{Ber}\left(\frac{1}{2^b}\right)$, so

$$\sum_{i=\frac{n}{2}+1}^{n} \mathbb{1}_{\{Y_i=h_i(j)\}} \sim F_j + \mathsf{Binom}\left(\frac{n}{2} - F_j, \frac{1}{2^b}\right).$$

Thus $\hat{\Pi}_j\left(Y_2^n\right)$ yields an unbiased estimator on Π_j with variance bounded by

$$\begin{aligned}\mathsf{Var}\left(\hat{\Pi}_j \middle| \sigma\right) &= \mathsf{Var}\left(\frac{2\cdot 2^b}{n\left(2^b-1\right)}\mathsf{Binom}\left(\frac{n}{2} - F_j, \frac{1}{2^b}\right)\middle|\sigma\right)\\ &\leq \frac{4\cdot 4^b}{n^2\left(2^b-1\right)^2}\cdot\left(\frac{n}{2}-F_j\right)\cdot\frac{1}{2^b} \leq \frac{4}{n2^b}.\end{aligned} \tag{16}$$

Next, we control the estimation errors by

$$\mathbb{E}\left[\sum_{j\in[d]}\left(\hat{\pi}_j - \pi_j\right)^2\right] \leq 2\mathbb{E}\left[\sum_{j\in[d]}\left(\hat{\pi}_j - \Pi_j\right)^2\right] + 2\mathbb{E}\left[\sum_{j\in[d]}\left(\Pi_j - \pi_j\right)^2\right]. \tag{17}$$

To bound the first term, consider two cases $j \in \mathcal{J}_\alpha$ and $j \notin \mathcal{J}_\alpha$. For $j \in \mathcal{J}_\alpha$, we have

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{\pi}_j - \Pi_j\right)^2\middle|\sigma\right]\\ =&\mathbb{P}\left\{\mathcal{J}_\alpha \subseteq \hat{\mathcal{J}}_\alpha\middle|\sigma\right\}\cdot\mathbb{E}\left[\left(\hat{\Pi}_j - \Pi_j\right)^2\middle|\mathcal{J}_\alpha \subseteq \hat{\mathcal{J}}_\alpha, \sigma\right] + \mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\middle|\sigma\right\}\cdot\mathbb{E}\left[\left(\hat{\pi}_j - \Pi_j\right)^2\middle|\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha, \sigma\right]\\ \overset{\text{(a)}}{=}&\mathbb{P}\left\{\mathcal{J}_\alpha \subseteq \hat{\mathcal{J}}_\alpha\middle|\sigma\right\}\cdot\mathbb{E}\left[\left(\hat{\Pi}_j - \Pi_j\right)^2\middle|\sigma\right] + \mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\middle|\sigma\right\}\cdot\mathbb{E}\left[\left(\hat{\pi}_j - \Pi_j\right)^2\middle|\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha, \sigma\right]\\ \leq&\mathbb{E}\left[\left(\hat{\Pi}_j - \Pi_j\right)^2\middle|\sigma\right] + 2\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\middle|\sigma\right\}\cdot\mathbb{E}\left[\hat{\pi}_j^2 + \Pi_j^2\middle|\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha, \sigma\right],\end{aligned}$$

where (a) holds since conditioned on σ, $\hat{\Pi}$ is independent with $\hat{\mathcal{J}}_\alpha$. Summing over j, we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{j\in\mathcal{J}_\alpha}\left(\hat{\pi}_j - \Pi_j\right)^2\middle|\sigma\right] &\leq \sum_{j\in\mathcal{J}_\alpha}\mathsf{Var}\left(\hat{\Pi}_j\middle|\sigma\right) + 2\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\middle|\sigma\right\}\sum_{j\in\mathcal{J}_\alpha}\mathbb{E}\left[\hat{\pi}_j^2 + \Pi_j^2\middle|\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha, \sigma\right]\\ &\leq \frac{4s}{n2^b} + 4\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\middle|\sigma\right\},\end{aligned} \tag{18}$$

where the last inequality follows from (16) and the fact that $\sum_j \Pi_j^2 \leq 1$ and $\sum_j \hat{\pi}_j^2 \leq 1$. Similarly for $j \notin \mathcal{J}_\alpha$, we have

$$\mathbb{E}\left[(\hat{\pi}_j - \Pi_j)^2 \middle| \sigma\right] \leq \mathbb{P}\left\{j \in \hat{\mathcal{J}}_\alpha \middle| \sigma\right\} \cdot \mathbb{E}\left[\left(\hat{\Pi}_j - \Pi_j\right)^2 \middle| \sigma\right] + \mathbb{P}\left\{j \notin \hat{\mathcal{J}}_\alpha \middle| \sigma\right\} \cdot \Pi_j^2.$$

Summing over j, we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{j \notin \mathcal{J}_\alpha} (\hat{\pi}_j - \Pi_j)^2 \middle| \sigma\right] &\leq \frac{4}{n2^b} \sum_{j \notin \mathcal{J}_\alpha} \mathbb{P}\left\{j \in \hat{\mathcal{J}}_\alpha \middle| \sigma\right\} + \sum_{j \notin \mathcal{J}_\alpha} \Pi_j^2 \\
&\leq \frac{4}{n2^b} \mathbb{E}\left[\sum_{j \in [d]} \mathbb{1}_{\left\{j \in \hat{\mathcal{J}}_\alpha\right\}}\right] + \sum_{j \notin \mathcal{J}_\alpha} \Pi_j^2 \\
&\leq \frac{4s}{n2^b} + \sum_{j \notin \mathcal{J}_\alpha} \Pi_j^2, \qquad (19)
\end{aligned}$$

where in the last inequality we use the fact that $\left|\hat{\mathcal{J}}_\alpha\right| \leq s$ with probability 1. Plugging (18) and (19) into (17) yields

$$\begin{aligned}
\mathbb{E}\left[\sum_{j \in [d]} (\hat{\pi}_j - \pi_j)^2\right] &\leq \frac{8s}{n2^b} + 4\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} + \mathbb{E}\left[\sum_{j \notin \mathcal{J}_\alpha} \Pi_j^2\right] + 2\mathbb{E}\left[\sum_{j \in [d]} (\Pi_j - \pi_j)^2\right] \\
&\overset{\text{(a)}}{\leq} \frac{8s}{n2^b} + 4\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} + \sum_{j \notin \mathcal{J}_\alpha} \left(\pi_j^2 + \mathsf{Var}\left(\Pi_j\right)\right) + \sum_{j \in [d]} \mathsf{Var}\left(\Pi_j\right) \\
&\overset{\text{(b)}}{\leq} \frac{8s}{n2^b} + 4\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} + s\alpha^2 + 2\sum_{j \in [d]} \mathsf{Var}\left(\Pi_j\right) \\
&\overset{\text{(c)}}{\leq} \frac{8s}{n2^b} + 4\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} + s\alpha^2 + \frac{1}{n}, \qquad (20)
\end{aligned}$$

where (a) holds since 1) $\Pi_j \sim \frac{2}{n}\mathsf{HG}\left(n, \frac{n}{2}, f_j\right)$ so $\mathbb{E}\left[\Pi_j\right] = \pi_j$, and 2) $\mathbb{E}\left[X^2\right] = \left(\mathbb{E}X\right)^2 + \mathsf{Var}\left(X\right)$, (b) holds since by definition, for all $j \notin \mathcal{J}_\alpha$, $\pi_j \leq \alpha$, and (c) follows from the fact $\mathsf{Var}\left(\Pi_j\right) \leq \frac{\pi_j}{2n}$. Hence it remains to bound $\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\}$.

Next, we prove that the localization schemes (Scheme A, B, D) yields the same (oder-wise) bound of failure probability.

Scheme A (uniform grouping) We slightly modify the encoding scheme Algorithm 4 such that each client i in the localization stage (i.e. i satisfies $\sigma(i) < n/2$) is assigned to a randomly selected group $\mathcal{G}_m$ with $m \in \mathsf{uniform}(M)$ chosen by shared randomness, where $M = \frac{d}{2^b-1}$. Follow the same analysis as in Lemma 8, for any $j \in \mathcal{J}_\alpha$, it can be successful localized if one of symbol j in the first half sequence is assigned to $\mathcal{G}_m$ with $j \in [(m-1)(2^b-1)+1 : m(2^b-1)]$. Denote such event by $\mathcal{E}_j$. Since there are $F_j' \triangleq f_j - F_j$ clients in the first stage observing j, the probability that symbol j is not detected is

$$\mathbb{P}\left\{\mathcal{E}_j^c \middle| \sigma\right\} \leq \left(1 - \frac{1}{M}\right)^{F_j'} \leq e^{-\frac{F_j'}{M}} = e^{-\frac{F_j' \cdot (2^b-1)}{d}}. \qquad (21)$$

To further bound it, we apply Hoeffding's inequality on hyper-geometric distribution (notice that $F_j' \sim \mathsf{HG}\left(n, \frac{n}{2}, f_j\right)$).

Lemma 19 (Hyper-geometric tail bound) *Let $F \sim \mathsf{HG}(n,k,f)$ and define $p = k/n$. We have*

$$\mathbb{P}\left\{F \leq (p-t)f\right\} \leq e^{-2t^2 f},$$

for all $t \in (0,p)$.

Applying Lemma 19 on (21) with $t = \frac{1}{4}$, we have

$$\mathbb{E}\left[\mathbb{P}\left\{\mathcal{E}_j^c | \sigma\right\}\right] \leq \mathbb{P}\left\{F_j' > \frac{1}{4}f_j\right\} \cdot e^{-\frac{f_j\cdot(2^b-1)}{4d}} + \mathbb{P}\left\{F_j' \leq \frac{1}{4}f_j\right\} \leq e^{-\frac{f_j\cdot(2^b-1)}{4d}} + e^{-\frac{f_j}{2}}.$$

Therefore

$$\mathbb{P}\left\{\bigcup_{j\in\mathcal{J}_\alpha} \mathcal{E}_j^c\right\} \leq \sum_{j\in\mathcal{J}_\alpha} e^{-\frac{f_j\cdot(2^b-1)}{4d}} + e^{-\frac{f_j}{2}} \leq se^{-\frac{n\alpha(2^b-1)}{4d}} + se^{-\frac{n\alpha}{2}}.$$

pick $\alpha = \frac{1}{\sqrt{n2^b}}$ and by the same argument as in (3), the ℓ_2 estimation error can be bounded by $O\left(\frac{s}{n2^b} \vee \frac{1}{n}\right)$, as long as $n = \Omega\left(\frac{d^2 \log d^2}{2^b}\right)$.

Scheme B: random hashing Let $\mathcal{E}_\ell$ be defined as in the proof of Lemma 10, and $j \in \mathcal{J}_\alpha \backslash \mathcal{C}_\ell$. Note that symbol $j \in \mathcal{J}_\alpha$ can be detected if a client i in localization stage observes j and W_i distinguishes j under hypothesis H_ℓ. Since every channels are generated identically and independently, such probability can be controlled by (11). Also notice that there are F_j' clients in the first stage who observe j, the failure probability can be controlled by

$$\mathbb{P}\left\{\mathcal{E}_\ell|\sigma\right\} \leq \left(1 - \frac{2^b-1}{4s}\right)^{F_j'} \leq \exp\left(-F_j' \frac{2^b-1}{4s}\right),$$

applying Lemma 19 gives us

$$\mathbb{E}\left[\mathbb{P}\left\{\mathcal{E}_\ell|\sigma\right\}\right] \leq \exp\left(-nf_j\frac{2^b-1}{16s}\right) + e^{-\frac{f_j}{2}} \leq \exp\left(-n\alpha\frac{2^b-1}{32s}\right) + e^{-\frac{n\alpha}{2}}.$$

Taking union bound over $\ell \in [N]$, we obtain

$$\mathbb{P}\left\{\bigcup_{\ell\in[N]} \mathcal{E}_\ell\right\} \leq \exp\left(-n\alpha\frac{2^b-1}{32s} + C_0 s\log\left(\frac{d}{s}\right)\right) + e^{-\frac{n\alpha}{2} + C_0 s \log\left(\frac{d}{s}\right)}.$$

Selecting $\alpha = \frac{1}{\sqrt{n2^b}}$ and as in Section 4.2, we conclude that as long as $n = \Omega\left(\frac{s^4 \log^2\left(\frac{d}{s}\right)}{2^b}\right)$, the ℓ_2 error (7) is $O\left(\frac{s}{n2^b} \vee \frac{1}{n}\right)$. Scheme A and Scheme B establishes the non-interactive part of Theorem 4.

Scheme D: tree-based recovery As in Scheme A, we replace every deterministic grouping with a randomized one. That is, in the localization step in Scheme D, instead of set client i with $\sigma(i) \equiv t(\text{ mod } \log d)$ into $\mathcal{G}_t$, we assign it to group $t_i \sim \mathsf{uniform}(\log d)$. Similarly, they will later be assigned to $\mathcal{K}_{t,m}$ with $m \sim \mathsf{uniform}(M)$. Then conditioned on σ, all of the analysis in the proof of Lemma 16 holds, except that we replace (14) with

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{E}_t^c|\mathcal{E}_{t-1},\sigma\right\} &= \mathbb{P}\left\{\mathcal{J}_{\alpha,t} \not\subset \hat{\mathcal{J}}_{\alpha,t}\middle|\mathcal{J}_{\alpha,t-1} \subseteq \hat{\mathcal{J}}_{\alpha,t-1},\sigma\right\}\\
&\leq \mathbb{P}\left\{\exists j \in \mathcal{B}_m \text{ s.t. } \forall i \in \mathcal{K}_{m,t}\, X_i \neq j|\sigma\right\}\\
&\leq |\mathcal{B}_m|\,\mathbb{P}\left\{\forall i \in \mathcal{K}_{m,t}\, X_i \neq j|\sigma\right\}\\
&\overset{\text{(a)}}{\leq} \frac{2s}{2^b-1}\cdot\left(1-\frac{1}{MT}\right)^{F'_j}\\
&\leq \frac{2s}{2^b-1}\exp\left(\frac{F'_j(2^b-1)}{2s\log d}\right).
\end{aligned}$$

where (a) holds because we assign each client uniformly at random in $[T]\times[M]$. Applying Lemma 19, we have

$$\mathbb{P}\left\{\mathcal{E}_t^c|\mathcal{E}_{t-1}\right\} \leq \frac{2s}{2^b-1}\exp\left(\frac{n\alpha(2^b-1)}{8s\log d}\right) + \exp\left(-\frac{n\alpha}{2}\right).$$

Plugging into (13) and assume $\frac{8s}{2^b-1} > 2$, we obtain

$$\mathbb{P}\left\{\mathcal{E}_T^c\right\} \leq T\cdot\frac{4s}{2^b-1}\cdot\exp\left(-\frac{n\alpha}{8s\log d}\right) = \exp\left(-\frac{n(2^b-1)\alpha}{8s\log d} + \left(\log\log d + \log\left(\frac{4s}{2^b-1}\right)\right)\right).$$

Finally, picking $n_1 = \frac{n}{2}$ and $\alpha = \frac{1}{\sqrt{n2^b}}$ and combining Lemma 16 and Lemma 7, we arrive at

$$\begin{aligned}
\mathbb{E}\left[\sum_{j\in[d]}(\hat{p}_j - p)^2\right] &\leq \exp\left(-\frac{n(2^b-1)\alpha}{8s\log d} + \left(\log\log d + \log\left(\frac{4s}{2^b-1}\right)\right)\right) + \frac{3s}{n2^b} + \frac{2}{n}\\
&\leq C_0\cdot\left(\frac{s}{n2^b}\vee\frac{1}{n}\right),
\end{aligned}$$

where the last inequality follows from Lemma 11 and the assumption

$$n = \Omega\left(\frac{s^2\log^2 d}{2^b}(\log s + \log\log d)^2\right).$$

This establishes the interactive part of Theorem 4.

■

E.10. Proof of Theorem 5

Proof Let $\mathcal{S}$ be the set of symbols with s-highest probabilities, that is, $\mathcal{S} = \left\{j \in [d] : p_j \geq p_{(s)}\right\}$[8] and let $\beta \triangleq p_{(s+1)}$. As in previous section, write $\mathcal{J}_\alpha \triangleq \left\{i : p_j \geq \max\left(\alpha, p_{(s)}\right)\right\}$[9] where $\alpha \geq 0$ will be specify later. Let $\hat{\mathcal{J}}_\alpha$ be the output of the localization step in Scheme B.

[8] In case that $p_{(s)} = p_{(s+1)}$, we define $\mathcal{S}$ to be an arbitrary set such that $|\mathcal{S}| = s$ and $p_j \geq p_{(s)}\ \forall i \in \mathcal{S}$.

[9] Again, with a slight abuse of notation, if $\alpha = p_{(s)} = p_{(s+1)}$, we require $\mathcal{J}_\alpha = \mathcal{S}$.

Claim E.1

$$\mathbb{P}\left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\} \leq n_1\frac{2^b-1}{4s}\left(1-P_{\mathcal{S}}\right) + \exp\left(-n_1\alpha\frac{2^b-1}{4s}+C_0 s\log d\right).$$

Moreover, $\left|\hat{\mathcal{J}}_\alpha\right| < s$ *almost surely.*

Proof Recall the in the decoding step of Scheme B, we define all the candidate supports as $\mathcal{C}_\ell$ for $\ell \in [N]$ (where $N \triangleq \binom{d}{s}$) and reduce the problem into multiple hypothesis $H_\ell : \mathcal{J}_\alpha \subseteq \mathcal{C}_\ell$. Now, without s-sparse assumption, we need a more detailed analysis and carefully bounding the failure probability.

To begin with, define $\mathcal{B} \subset [N]$ to be the indices of "incorrect" hypotheses, that is, $\mathcal{B} \triangleq \{\ell \in [N] : \mathcal{J}_\alpha \not\subset \mathcal{C}_\ell\}$. Then observe that the error event $\mathcal{E}_f \triangleq \left\{\mathcal{J}_\alpha \not\subset \hat{\mathcal{J}}_\alpha\right\}$ occurs when 1) there exist one incorrect but consistent[10] candidates $\ell \in \mathcal{B}$, i.e. for $\ell \in \mathcal{B}$, define

$$\mathcal{F}_\ell \triangleq \left\{\mathcal{C}_\ell \text{ is consistent with } \left(Y^{n_1}, W^{n_1}\right)\right\};$$

and 2) for all $\ell \notin \mathcal{B}$, $\mathcal{C}_\ell$ does not consistent with (Y^{n_1}, W^{n_1}), i.e. for all $\ell \notin \mathcal{B}$,

$$\mathcal{G}_\ell \triangleq \left\{\mathcal{C}_\ell \text{ is \textit{not} consistent with } \left(Y^{n_1}, W^{n_1}\right)\right\}.$$

Therefore, we can bound the failure event by

$$\mathcal{E}_f \subseteq \left(\bigcup_{\ell\in\mathcal{B}} \mathcal{F}_\ell\right) \cup \left(\bigcap_{\ell\in[N]\setminus\mathcal{B}} \mathcal{G}_\ell\right).$$

Bounding $\mathcal{F}_\ell$ We bound $\mathcal{F}_\ell$ as in previous section: observe that $\mathcal{F}_\ell$ cannot happen if there exist a client i 1) who observes $j \in \mathcal{J}_\alpha \setminus \mathcal{C}_\ell$ and 2) whose channel W_i distinguishes j. Thus we have

$$\begin{aligned}\mathbb{P}\left\{\mathcal{F}_\ell\right\} &\leq \left(1-\mathbb{P}\left\{\{X_i=j\}\cap\{W_i \text{ distinguishes } j\}\right\}\right)^{n_1}\\ &\leq \left(1-\alpha\cdot\frac{2^b-1}{4s}\right)^{n_1} \leq \exp\left(-n_1\alpha\frac{2^b-1}{4s}\right).\end{aligned} \tag{22}$$

Bounding $\mathcal{G}_\ell$ Let ℓ^* be the index such that $\mathcal{C}_{\ell^*} \triangleq \mathcal{S}$. Then

$$\left(\bigcap_{\ell\in[N]\setminus\mathcal{B}} \mathcal{G}_\ell\right) \subseteq \mathcal{G}_{\ell^*}. \tag{23}$$

[10] Recall that the consistency is defined in Def 9

Observe that $\mathcal{G}_{\ell^*}$ happens if there exist a client i who 1) observes $j \neq \mathcal{S}$ and 2) W_i distinguishes j. Therefore

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{G}_{\ell^*}\right\} &\leq \mathbb{P}\left\{\bigcup_{i\in[n]}\bigcup_{j\in[d]\setminus\mathcal{S}}\left\{X_i = j\right\}\cap\left\{W_i \text{ distinguishes} j\right\}\right\}\\
&\overset{\text{(a)}}{\leq} n\sum_{j\in[d]\setminus\mathcal{S}}\mathbb{P}\left\{\left\{X_i = j\right\}\cap\left\{W_i \text{ distinguishes} j\right\}\right\}\\
&\overset{\text{(b)}}{\leq} n\frac{2^b-1}{4s}\left(\sum_{j\in[d]\setminus\mathcal{S}}\mathbb{P}\left\{X_i = j\right\}\right)\\
&= n\frac{2^b-1}{4s}\left(1-P_{\mathcal{S}}\right), \qquad (24)
\end{aligned}$$

where (a) is due to union bound, and (b) is because we generate W_i independently with X_i.

Finally, combining (22), (23) and (24), we obtain

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{E}_f\right\} &\leq \mathbb{P}\left\{\left(\bigcup_{\ell\in\mathcal{B}}\mathcal{F}_\ell\right)\cup\left(\bigcap_{\ell\in[N]\setminus\mathcal{B}}\mathcal{G}_\ell\right)\right\}\\
&\leq n_1\frac{2^b-1}{4s}\left(1-P_{\mathcal{S}}\right)+\exp\left(-n_1\alpha\frac{2^b-1}{4s}+C_0 s\log d\right),
\end{aligned}$$

where the last inequality is due to union bound over $[N] = \binom{d}{s}$. ■

As in previous section, picking $\alpha = \frac{1}{\sqrt{n\cdot 2^b}}$ and by Claim E.1, we can show that

$$\begin{aligned}
\mathbb{E}\left[\sum_{j\in[d]}\left(\hat{p}_j - p\right)^2\right] \leq \left(n_1\frac{2^b-1}{4s}+1\right)\left(1-P_{\mathcal{S}}\right)+\exp\left(-\sqrt{\frac{n_1^2\left(2^b-1\right)^2}{n\cdot 2^b}}\frac{1}{4s}+C_0 s\log\left(\frac{d}{s}\right)\right)\\
+\frac{2s}{(n-n_1)2^b}+\frac{1}{n-n_1}.
\end{aligned}$$

Finally, picking $n_1 = C_1\cdot\sqrt{n\log n}\cdot s\cdot\log\left(\frac{d}{s}\right)$, we have

$$\mathbb{E}\left[\sum_{j\in[d]}\left(\hat{p}_j - p\right)^2\right] \leq C_2\cdot\left(\frac{s}{n\cdot 2^b}\vee\frac{1}{n}\right)+C_3\cdot\sqrt{n\log n}\cdot 2^b\cdot\left(1-P_{\mathcal{S}}\right).$$

■