Communication-Constrained Inference and the Role of Shared Randomness

Jayadev Acharya [1] **Clément Canonne** [2] **Himanshu Tyagi** [3]

Abstract

A central server needs to perform statistical inference based on samples that are distributed over multiple users who can each send a message of limited length to the center. We study problems of distribution learning and identity testing in this distributed inference setting and examine the role of shared randomness as a resource. We propose a general purpose *simulate-and-infer* strategy that uses only private-coin communication protocols and is sample-optimal for distribution learning. This general strategy turns out to be sample-optimal even for distribution testing among private-coin protocols. Interestingly, we propose a public-coin protocol that outperforms simulate-and-infer for distribution testing and is, in fact, sample-optimal. Underlying our public-coin protocol is a random hash that when applied to the samples minimally contracts the chi-squared distance of their distribution from the uniform distribution.

1. Introduction

Sample-optimal statistical inference has taken center-stage in modern data analytics where the number of samples can be comparable to the dimensions of the data. In many emerging applications, especially those arising in sensor networks and the Internet of Things (IoT), we are not only constrained in the number of samples but are also given access to only limited communication about the samples. We consider such a distributed inference setting and seek sample-optimal algorithms for inference under communication constraints.

In our setting, n players get independent samples from an unknown k-ary distribution and each can send only ℓ bits about their observed sample to a central referee using a simultaneous message passing (SMP) protocol for communication. The referee uses communication from the players to accomplish an inference task $\mathcal{P}$.

[*]Equal contribution [1]Cornell University [2]Stanford University [3]Indian Institute of Science Institute of Technology. Correspondence to: Clément Canonne <ccanonne@cs.stanford.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Question 1.1. *What is the minimum number of players n required by an SMP protocol that successfully accomplishes $\mathcal{P}$, as a function of k, ℓ, and the relevant parameters of $\mathcal{P}$?*

Our first contribution is a general *simulate-and-infer* strategy for inference under communication constraints where we use the communication to simulate samples from the unknown distribution at the referee. To describe this strategy, we introduce a natural notion of *distributed simulation*: n players observing an independent sample each from an unknown k-ary distribution $\mathbf{p}$ can send ℓ-bits each to a referee. A distributed simulation protocol consists of an SMP protocol and a randomized decision map that enables the referee to generate a sample from $\mathbf{p}$ using the communication from the players. Clearly, when[1] $\ell \geq \log k$ such a sample can be obtained by getting the sample of any one player. But what can be done in the communication-starved regime of $\ell < \log k$?

We first show that perfect simulation is impossible using any finite number of players in the communication-starved regime. But perfect simulation is not even required for our application. When we allow a small probability of declaring failure, namely admit Las Vegas simulation schemes, we obtain a distributed simulation scheme that requires an optimal $O\big(k/2^{\ell}\big)$ players to simulate k-ary distributions using ℓ bits of communication per player. Thus, our proposed simulate-and-infer strategy can accomplish $\mathcal{P}$ with a blow-up in sample-complexity by an extra factor of $O\big(k/2^{\ell}\big)$.

The specific inference tasks we consider are those of *distribution learning*, where we seek to estimate the unknown k-ary distribution to an accuracy of ε in total variation distance, and *identity testing* where we seek to know if the unknown distribution is $\mathbf{q}$ or ε-far from it in total variation distance. For distribution learning, the simulate-and-infer strategy matches the lower bound from (**?**) and is therefore sample-optimal. For identity testing, the plot thickens.

Recently, a lower bound for the sample complexity of identity testing using only private-coin protocols was established (**?**). The simulate-and-infer protocol is indeed a private-coin protocol and it attains this lower bound. When public coins (shared randomness) are available, (**?**) derived a different, more relaxed lower bound. The performance of simulate-and-infer is far from this lower bound. Our second contribution is a public-coin protocol for identity testing

[1]We assume throughout that $\log k$ is an integer.

that not only outperforms simulate-and-infer but matches the lower bound in (**?**) and is sample-optimal.

We provide a concrete description of our results in the next section, followed by an overview of our proof techniques in the subsequent section. To put our results in context, we provide a brief overview of the literature as well.

1.1. Main results

We begin by summarizing our distributed simulation results.

Theorem 1.2. *For every $k, \ell \geq 1$, there exists a private-coin protocol with ℓ bits of communication per player for distributed simulation over $[k]$ and expected number of players $O\big((k/2^\ell) \vee 1\big)$. Moreover, this expected number is optimal, up to constant factors, even when public-coin and interactive communication protocols are allowed.*

The proposed algorithm is a Las Vegas algorithm,[2] which produces a sample from the unknown distribution when they terminate, but they may never terminate. In fact, we can show that distributed simulation is impossible, unless we allow for such algorithms.

Theorem 1.3. *For $k \geq 1$, $\ell < \log k$, and any $N \in \mathbb{N}$, there does not exist a SMP protocol with N players and ℓ bits of communication per player for distributed simulation over $[k]$. Furthermore, the result continues to hold even for public-coin and interactive communication protocols.*

The proof is delegated to Section 4.1.

Since the distributed simulation protocol in Theorem 1.3 is a private-coin protocol, we can use it to generate the desired number of samples from the unknown distribution at the center to obtain the following result.

Theorem 1.4 (Informal)**.** *For any inference task $\mathcal{P}$ over k-ary distributions with sample complexity s in the non-distributed model, there exists a private-coin protocol for $\mathcal{P}$ using ℓ bits of communication per player and requiring $n = O(s \cdot (k/2^\ell \vee 1))$ players.*

Instantiating this general statement for distribution learning and identity testing leads to the following results.

Corollary 1.5. *For every $k, \ell \geq 1$, simulate-and-infer can accomplish distribution learning over $[k]$, with ℓ bits of communication per player and $n = O\Big(\frac{k^2}{(2^\ell \wedge k)\varepsilon^2}\Big)$ players.*

Corollary 1.6. *For every $k, \ell \geq 1$, simulate-and-infer can accomplish identity testing over $[k]$ using ℓ bits of communication per player and $n = O\Big(\frac{k^{3/2}}{(2^\ell \wedge k)\varepsilon^2}\Big)$ players.*

Using the lower bound in (**?**) (see, also, (**?**)), we obtain that simulate-and-infer is sample-optimal for distribution learning even when public-coin protocols are allowed. In fact, the sample complexity of simulate-and-infer for identity testing matches the lower bound for private-coin protocols in (**?**), rendering it sample-optimal.

Our most striking result is the next one which shows that public-coin protocols can outperform the sample complexity of private-coin protocols for identity testing by a factor of $\sqrt{k/2^\ell}$.

Theorem 1.7. *For every $k, \ell \geq 1$, there exists a public-coin protocol for identity testing over $[k]$ using ℓ bits of communication per player and $n = O\Big(\frac{k}{\sqrt{2^\ell \wedge k}\varepsilon^2}\Big)$ players.*

We further note that our protocol is remarkably simple to describe and implement: We generate a random partition of $[k]$ into 2^ℓ parts and report which part each sample lies in. Although, as stated, our protocol seems to require $\Omega(\ell \cdot k)$ bits of shared randomness, an immediate inspection of the proof shows that 4-wise independent shared randomness suffice, drastically reducing the number of random bits required.

Our results are summarized in the table below.

Distribution Learning		Identity Testing	
Public-Coin	Private-Coin	Public-Coin	Private-Coin
$\frac{k}{\varepsilon^2} \cdot \frac{k}{2^\ell}$		$\frac{\sqrt{k}}{\varepsilon^2} \cdot \sqrt{\frac{k}{2^\ell}}$	$\frac{\sqrt{k}}{\varepsilon^2} \cdot \frac{k}{2^\ell}$

Table 1. Summary of the sample complexity of distributed learning and testing, under private and public randomness. All results are order optimal.

Interestingly, this shows that public randomness, despite allowing a significant sample complexity improvement for identity testing, is not helpful for distribution learning. A high-level heuristic to explain this discrepancy can be obtained by focusing on the uniform distribution. For testing, we are given a fixed (unknown) distribution at distance ε, and public randomness helps as it allows focusing on the appropriate direction to separate this distribution from the uniform one. However, for learning, ones needs to distinguish the uniform distribution from *all* distributions at distance ε – i.e., in all directions at once, thereby making public randomness useless.

1.2. Proof techniques

We now provide a high-level description of the proofs of our main results.

Distributed simulation. The upper bound of Theorem 1.3 uses a rejection sampling based approach; see Section 5 for details. The lower bound follows by relating distributed simulation to communication constrained distribution learning and using the lower bound for sample complexity of latter from (**??**).

Distributed identity testing. Using a reduction due Goldreich (**?**), we note first that it suffices to consider uniformity testing. To test whether an unknown distribution $\mathbf{p}$ is uni-

[2] Or, roughly equivalently, when one is allowed to abort with a special symbol with small constant probability.

form using at most ℓ bits to describe each sample, a natural idea is to randomly partition the alphabet into $L := 2^\ell$ parts, and send to the referee independent samples from the L-ary distribution $\mathbf{q}$ induced by $\mathbf{p}$ on this partition. For a random balanced partition (i.e., where every part has cardinality k/L), clearly the uniform distribution $\mathbf{u}_k$ is mapped to the uniform distribution $\mathbf{u}_L$. Thus, one can hope to reduce the problem of testing uniformity of $\mathbf{p}$ (over $[k]$) to that of testing uniformity of $\mathbf{q}$ (over $[L]$). The latter task would be easy to perform, as every player can simulate one sample from $\mathbf{q}$ and communicate it fully to the referee with $\log L = \ell$ bits of communication. Hence, the key issue is to argue that this random "flattening" of $\mathbf{p}$ would somehow preserve the distance to uniformity; namely, that if $\mathbf{p}$ is ε-far from $\mathbf{u}_k$, then (with a constant probability over the choice of the random partition) $\mathbf{q}$ will remain ε'-far from $\mathbf{u}_L$, for some ε' depending on ε, L, and k. If true, then it is easy to see that this would imply a very simple protocol with $O(\sqrt{L}/\varepsilon'^2)$ players, where all agree on a random partition and send the induced samples to the referee, who then runs a centralized uniformity test. Therefore, in order to apply the aforementioned natural recipe, it suffices to derive a "random flattening" structural result for $\varepsilon' \asymp \sqrt{(L/k)}\varepsilon$.

An issue with this approach, unfortunately, is that the total variation distance (that is, the ℓ_1 distance) does not behave as desired under these random flattenings, and the validity of our desired result remains unclear. Fortunately, an analogous statement with respect to the ℓ_2 distance turns out to be much more manageable and suffices for our purposes. In more detail, we show that a random flattening of $\mathbf{p}$ does preserve, with constant probability, the ℓ_2 distance to uniformity; in our case, by Cauchy–Schwarz the original ℓ_2 distance will be at least $\gamma \asymp \varepsilon/\sqrt{k}$, which implies using known ℓ_2 testing results that one can test uniformity of the "randomly flattened" $\mathbf{q}$ with $O(1/(\sqrt{L}\gamma^2)) = O(k/(2^{\ell/2}\varepsilon^2))$ samples. This yields the desired guarantees on the protocol. However, the proposed algorithm suffers one drawback: The amount of public randomness required for the players to agree on a random balanced partition is $\Omega(k \log L) = \Omega(k \cdot \ell)$, which in cases with large alphabet size k can be prohibitive.

1.3. Related prior work

Distribution learning problem is finite-dimensional parametric learning problem, and the identity testing problem is a specific goodness-of-fit problem. Both these problems have a long history in statistics. However, the sample-optimal setting of interest to us has received a lot of attention in the past decade, especially in the computer science literature; see (**???**) for survey. Most pertinent to our work is uniformity testing (**???**), the prototypical distribution testing problem for which the sample complexity was established to be $\Theta(\sqrt{k}/\varepsilon^2)$ in **??**.

Distributed hypothesis testing and estimation problems were first studied in information theory, although in a different setting than what we consider (**???**). The focus in that line of work has been to characterize the trade-off between asymptotic error exponent and communication rate per sample.

Closer to our work is distributed parameter estimation and functional estimation that has gained significant attention in recent years (see e.g., (**????**)). In these works, much like our setting, independent samples are distributed across players, which deviates from the information theory setting described above where each player observes a fixed dimension of each independent sample. However, the communication model in these results differs from ours, and the communication-starved regime we consider has not been studied in these works.

The problem of distributed density estimation, too, has gathered recent interest in various statistical settings (**????????**). Our work is closest to two of these: The aforementioned (**??**) and (**?**). The latter considers both ℓ_1 (total variation) and ℓ_2 losses, although in a different setting than ours. Specifically, they study an interactive model where the players do not have any individual communication constraint, but instead the goal is to bound the total number of bits communicated over the course of the protocol. This difference in the model leads to incomparable results and techniques (for instance, the lower bound for learning k-ary distributions in our model is higher than the upper bound in theirs).

Our current work further deviates from this prior literature, since we consider distribution testing as well and examine the role of public-coin for SMP protocols. Additionally, a central theme here is the connection to distribution simulation and its limitation in enabling distributed testing. In contrast, the prior work on distribution estimation, in essence, establishes the optimality of simple protocols that rely on distributed simulation for inference. (We note that although recent work of (**?**) considers both communication complexity and distribution testing, their goal and results are very different – indeed, they explain how to leverage on negative results in the standard SMP model of communication complexity to obtain sample complexity lower bounds in collocated distribution testing.)

Problems related to joint simulation of probability distributions have been the object of focus in the information theory and computer science literature. Starting with the works of Gács and Körner (**?**) and Wyner (**?**) where the problem of generating shared randomness from correlated randomness and vice-versa, respectively, were considered, several important variants have been studied such as correlated sampling (**????**) and non-interactive simulation (**???**). Yet, our problem of exact simulation of a single (unknown) distribution with communication constraints from multiple parties has not been studied previously to the best of our knowledge.

1.4. Organization

We begin by setting notation and recalling some useful definitions and results in Section 2, before formally introducing our distributed model in Section 3. Next, Section 4 introduces the question of distributed simulation and contains our protocols and impossibility results for this problem. In Section 5, we consider the relation between distributed simulation and private-coin distribution inference. The subsequent section, Section 6, focuses on the problem of uniformity testing and contains the proofs of the upper and lower bounds of Theorem 1.7. Due to lack of space, we only provide proof outlines and the details are relegated to the appendix.

2. Preliminaries

We write $\log$ (resp. $\ln$) for the binary (resp. natural) logarithm, and $[k]$ for the set of integers $\{1, 2, \ldots, k\}$. Given a fixed (and known) discrete domain $\mathcal{X}$ of size k, we denote by $\Delta_{\mathcal{X}}$ the set of probability distributions over $\mathcal{X}$, i.e.,

$$\Delta_{\mathcal{X}} = \{\, \mathbf{p} \colon \mathcal{X} \to [0,1] \,:\, \|\mathbf{p}\|_1 = 1 \,\}\,.$$

A *property of distributions* over $\mathcal{X}$ is a subset $\mathcal{P} \subseteq \Delta_{\mathcal{X}}$. Given $\mathbf{p} \in \Delta_{\mathcal{X}}$ and a property $\mathcal{P}$, the distance from $\mathbf{p}$ to the property is defined as

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}) := \inf_{\mathbf{q} \in \mathcal{P}} \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \tag{1}$$

where $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \sup_{S \subseteq \mathcal{X}} (\mathbf{p}(S) - \mathbf{q}(S))$ for $\mathbf{p}, \mathbf{q} \in \Delta_{\mathcal{X}}$, is the *total variation distance* between $\mathbf{p}$ and $\mathbf{q}$. For a given parameter $\varepsilon \in (0, 1]$, we say that $\mathbf{p}$ is *ε-close* to $\mathcal{P}$ if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}) \leq \varepsilon$; otherwise, we say that $\mathbf{p}$ is *ε-far* from $\mathcal{P}$. For a discrete set $\mathcal{X}$, we write $\mathbf{u}_{\mathcal{X}}$ for the uniform distribution on $\mathcal{X}$, and will sometimes omit the subscript when the domain is clear from context. We indicate by $x \sim \mathbf{p}$ that x is a sample drawn from the distribution $\mathbf{p}$.

In addition to total variation distance, we shall rely in some of our proofs on the χ^2 and Kullback–Leibler (KL) divergences between discrete distributions $\mathbf{p}, \mathbf{q} \in \Delta_{\mathcal{X}}$, defined respectively as $\chi^2(\mathbf{p}, \mathbf{q}) := \sum_{x \in \mathcal{X}} \frac{(\mathbf{p}_x - \mathbf{q}_x)^2}{\mathbf{q}_x(1-\mathbf{q}_x)}$ and $D(\mathbf{p}\|\mathbf{q}) := \sum_{x \in \mathcal{X}} \mathbf{p}_x \ln \frac{\mathbf{p}_x}{\mathbf{q}_x}$.

We use the standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$; and will sometimes write $a_n \lesssim b_n$ to indicate that there exists an absolute constant $c > 0$ such that $a_n \leq c \cdot b_n$ for all n. Finally, we will denote by $a \wedge b$ and $a \vee b$ the minimum and maximum of two numbers a and b, respectively.

3. Communication, Simulation, and Inference Protocols

We set the stage by describing the communication protocols we study for both the distributed simulation and the distributed inference problems. Throughout the paper, we restrict to simultaneous communication models with private and public randomness. We remark that simultaneous communication does not mean that the messages are sent at the same time. It is a formalism that implies that the messsages from any user cannot be used by others in their protocols.

Formally, n players observe samples $X_1, \ldots, X_n$ with player i given access to X_i. The samples are assumed to be generated independently from an unknown distribution $\mathbf{p}$. In addition, player i has access to uniform randomness U_i such that $(U_1, \ldots, U_n)$ is jointly independent of $(X_1, \ldots, X_n)$. An ℓ-bit *simultaneous message-passing* (SMP) communication protocol π for the players consists of $\{0,1\}^\ell$-valued mappings $\pi_1, \ldots, \pi_n$ where player i sends the message $M_i = \pi_i(X_i, U_i)$. The message $M = (M_1, \ldots, M_n)$ sent by the players is received by a common referee. Based on the assumptions on the availability of the randomness $(U_1, \ldots, U_n)$ to the referee and the players, three natural classes of protocols arise:

1. *Private-coin protocols:* $U_1, \ldots, U_n$ are mutually independent and unavailable to the referee.
2. *Public-coin protocols:* All player and the referee have access to $U_1, \ldots, U_n$.

For the ease of presentation, we represent the private randomness communication $f_i(x_i, U_i)$ using a channel $W_i \colon \mathcal{X} \to \{0,1\}^\ell$ where player i upon observing x_i declares y with probability $W_i(y|x_i)$. Also, for public-coin protocols, we can assume without loss of generality that $U_1 = U_2 = \cdots = U_n$.

Distributed simulation protocols. An ℓ-bit *simulation* $\mathcal{S} = (\pi, \delta)$ of k-ary distributions using n players consists of an ℓ-bit SMP protocol π and a decision map δ comprising mappings $\delta_x \colon (M, U) \mapsto [0,1]$ such that for each message m and randomness u,

$$\sum_x \delta_x(m, u) \leq 1.$$

Upon observing the message $M = (M_1, \ldots, M_n)$ and (depending on the type of protocol) randomness $U = (U_1, \ldots, U_n)$, the referee declares the random sample $\hat{X} = x$ with probability $\delta_x(M, U)$ or declares an abort symbol $\perp$ if no x is selected. For concreteness, we assume that the random variable $\hat{X}$ takes values in $\mathcal{X} \cup \{\perp\}$ with $\{\hat{X} = \perp\}$ corresponding to the abort event. When π is a private or public-coin protocol, respectively, the simulation $\mathcal{S}$ is called private or public-coin simulation.

A simulation $\mathcal{S}$ is an *α-simulation* if for every $\mathbf{p}$

$$\Pr_{\mathbf{p}}\Big[\, \hat{X} = x \mid \hat{X} \neq \perp \,\Big] = \mathbf{p}_x, \quad \forall\, x \in \mathcal{X},$$

and the abort probability satisfies $\Pr_{\mathbf{p}}\Big[\, \hat{X} = \perp \,\Big] \leq \alpha$. When the probability of abort is *zero*, $\mathcal{S}$ is termed a *perfect simulation*.

Distributed inference protocols. We give a general definition of distributed inference protocols that is applicable beyond the use-cases considered in this work. An inference problem $\mathcal{P}$ can be described by a tuple $(\mathcal{C}, \mathcal{X}, \mathcal{E}, L)$ where $\mathcal{C}$ denotes a family of distributions on the alphabet $\mathcal{X}$, $\mathcal{E}$ a class of allowed estimates for elements of $\mathcal{C}$ (or their functions), and $L\colon \mathcal{C} \times \mathcal{E} \to \mathbb{R}_+^q$ is a loss function that evaluates the accuracy of our estimate $e \in \mathcal{E}$ when $\mathbf{p} \in \mathcal{C}$ was the ground truth.

An ℓ-bit *distributed inference protocol* $\mathcal{I} = (\pi, e)$ for the inference problem $(\mathcal{C}, \mathcal{X}, \mathcal{E}, L)$ consists of an ℓ-bit SMP protocol π and an estimator e available to the referee who, upon observing the message $M = \pi(X^n, U)$ and the randomness U, estimates the unknown $\mathbf{p}$ as $e(M, U) \in \mathcal{E}$. As before, we say that a private-, or public-coin inference protocol, respectively, uses a private- or public-coin communication protocol π.

For $\vec{\gamma} \in \mathbb{R}_+^q$, an inference protocol (π, e) is a *$\vec{\gamma}$-inference protocol* if

$$\mathbb{E}_{\mathbf{p}}[L_i(\mathbf{p}, e(M, U))] \leq \gamma_i, \quad \forall 1 \leq i \leq q.$$

We instantiate the abstract definition above in two illustrative questions that we will pursue in this paper.

Example 3.1 (**Distribution learning**). Consider the problem $\mathcal{L}_k(\varepsilon, \delta)$ of estimating a k-ary distribution $\mathbf{p}$ by observing independent samples from it, namely the finite alphabet distribution learning problem. This problem is obtained from the general formulation above by setting $\mathcal{X}$ to be $[k]$, $\mathcal{C}$ and $\mathcal{E}$ both to be the $(k-1)$-dimensional probability simplex $\mathcal{C}_k$, and $L(\mathbf{p}, \hat{\mathbf{p}})$ as follows:

$$L(\mathbf{p}, \hat{\mathbf{p}}) = \mathbb{1}_{\left\{\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \hat{\mathbf{p}}) > \varepsilon\right\}}.$$

For this case, we term the δ-inference protocol an ℓ-bit (k, ε, δ)-*learning protocol* for n player. In this case, γ is equal to δ, the probability of error.

Example 3.2 (**Uniformity testing**). In the uniformity testing problem $\mathcal{T}_k(\varepsilon, \delta)$, our goal is to determine whether $\mathbf{p}$ is the uniform distribution $\mathbf{u}_k$ over $[k]$ (null hypothesis H_0) or if it satisfies $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$ (alternative hypothesis H_1). This can be obtained as a special case of our general formulation by setting $\mathcal{X} = [k]$, $\mathcal{C}$ to be the set containing $\mathbf{u}_k$ and all $\mathbf{p}$ satisfying $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$, $\mathcal{E} = \{0, 1\}$, and the loss function L to be

$$L(\mathbf{p}, b) = b \cdot \mathbb{1}_{\{\mathbf{p} = \mathbf{u}_k\}} + (1 - b) \cdot \mathbb{1}_{\{\mathbf{p} \neq \mathbf{u}_k\}}, \quad b \in \{0, 1\},$$

where b denotes the output of the test (i.e., declaring hypothesis H_b).

For this case, we term the δ-inference protocol an ℓ-bit (k, ε, δ)-*uniformity testing protocol* for n players. Further, for simplicity we will refer to $(k, \varepsilon, 1/3)$-uniformity testing protocols simply as (k, ε)-*uniformity testing protocols*.

Note that distributed variants of several other inference problems such as that of estimating functionals of distributions and parametric estimation problems can be included as instantiations of the distributed inference problem described above.

We close by noting that while we have restricted to the SMP model of communication, the formulation can be easily extended to include interactive communication protocols where the communication from each player can be heard by all the other players (and the referee), and in its turn, a player communicates using its local observation and the communication received from all the other players in the past. A formal description of such a protocol can be given in the form of a multiplayer protocol tree *à la* (**?**). However, such considerations are beyond the scope of this paper.

A note on the parameters. It is immediate to see that for $\ell \geq \log k$ the distributed and centralized settings are equivalent, as the players can simply send their input sample to the referee (thus, both upper and lower bounds from the centralized setting carry over).

4. Distributed Simulation

In this section, we consider the distributed simulation problem described in the previous section. The proof of impossibility of perfect simulation (Theorem 1.2) when $\ell < \log k$ and $n < \infty$ is given in Section 4.1. We now consider α-simulation for constant $\alpha \in (0, 1)$ and exhibit an ℓ-bit α-simulation of k-ary distributions using $O(k/2^\ell)$ players. In fact, by drawing on a reduction from distributed distribution learning, we will show in the next section that this is the least number of players required (up to a constant factor) for α-simulation for any $\alpha \in (0, 1)$. The sample complexity of our simulation algorithm for a general α can be shown to be $O(k/2^\ell \log(1/\alpha))$; we omit the argument here due to space constraints and defer it to the full version of the paper (**?**).

We now establish Theorem 1.3 and provide α-simulation protocols for k-ary distributions using $n = O(k/2^\ell)$ players. We first present the protocol for the case $\ell = 1$, before extending it to general ℓ. The proof of lower bound for the number of players required for α-simulation of k-ary distributions is based on the connection between distributed simulation and distributed distribution learning and will be provided in the next section where this connection is discussed in detail.

For ease of presentation, we allow a slightly different class of protocols where we have an infinitely long sequence of players, each with access to one independent sample from the unknown $\mathbf{p}$. The referee's protocol entails checking each player's message and deciding either to declare an output $\hat{X} = x$ and stop, or see the next player's output. We assume that with probability one the referee uses finitely many players and declares an output. The cost of maximum number of players of the previous setting is now replaced

with the expected number of players used to declare an output. By an application of Markov's inequality, this can be easily related to our original setting of private-coin α-simulation.

Theorem 4.1. *There exists a* 1*-bit private-coin protocol that outputs a sample* $x \sim \mathbf{p}$ *using messages of at most* $20k$ *players in expectation.*

Proof Sketch. We describe the base version of the protocol below, and the delegate the description of the complete protocol and the detailed proof to **??**.

The scheme, base version. Consider a protocol with $2k$ players where the 1-bit communication from players $(2i-1)$ and $(2i)$ just indicates if their observation is i or not, namely $\pi_{2i-1}(x) = \pi_{2i}(x) = \mathbb{1}_{\{x=i\}}$.

On receiving these $2k$ bits, the referee $\mathcal{R}$ acts as follows:

- if exactly one of the bits $M_1, M_3, \ldots, M_{2k-1}$ is equal to one, say the bit M_{2i-1}, and the corresponding bit M_{2i} is zero, then the referee outputs $\hat{X} = i$;
- otherwise, it outputs $\perp$.

In the above, the probability $\rho_{\mathbf{p}}$ that some $i \in [k]$ is declared as the output (and not $\perp$) is

$$\rho_{\mathbf{p}} := \sum_{i=1}^{k} (1-\mathbf{p}_i) \cdot \mathbf{p}_i \prod_{j \neq i} (1-\mathbf{p}_j) = \prod_{j=1}^{k} (1-\mathbf{p}_j),$$

so that

$$\rho_{\mathbf{p}} = \exp \sum_{j=1}^{k} \ln(1-\mathbf{p}_j) = \exp\left(-\sum_{t=1}^{\infty} \frac{\|\mathbf{p}\|_t^t}{t} \right)$$
$$\geq \exp\left(-\left(1 + \sum_{t=2}^{\infty} \frac{\|\mathbf{p}\|_2^t}{t} \right) \right) = \frac{1-\|\mathbf{p}\|_2}{e^{1-\|\mathbf{p}\|_2}}$$

which is bounded away from 0 as long as $\mathbf{p}$ is far from being a point mass (i.e., $\|\mathbf{p}\|_2$ is not too close to 1).

Further, for any fixed $i \in [k]$, the probability that $\mathcal{R}$ outputs i is

$$\mathbf{p}_i \cdot \prod_{j=1}^{k} (1-\mathbf{p}_j) = \mathbf{p}_i \rho_{\mathbf{p}} \propto \mathbf{p}_i \,.$$

The full scheme now requires some modifications to this approach, esp. to handle this "point mass" issue; we provide the entire proof in **??**, establishing the stated bound of $20k$ players (in expectation). □

The extension for general ℓ is given in **??**.

5. The Simulate-and-Infer Strategy

In this section, we focus on the connection between distributed simulation and (private-coin) distributed inference. We first describe the implications of the results from Section 4 for *any* distributed inference task; before considering the natural question this general connection prompts: "Are the resulting protocols optimal?"

Having a distributed simulation protocol at our disposal, a natural protocol for distributed inference entails using distributed simulation to generate independent samples from the underlying distribution, as many as warranted by the sample complexity of the underlying problem, before running a sample inference algorithm (for the centralized setting) at the referee. The resulting protocol will require a number of players roughly equal to the sample complexity of the inference problem when the samples are centralized times $\left(k/2^{\ell}\right)$, the number of players required to simulate each independent sample at the referee. We refer to such protocols that first simulate samples from the underlying distribution and then use a standard sample-optimal inference algorithm at the referee as *simulate-and-infer* protocols. Formally, we have the following result.

Theorem 5.1. *Let* $\mathcal{P}$ *be an inference problem for distributions over a domain of size* k *that is solvable using* $\psi(\mathcal{P}, k)$ *samples with error probability at most 1/3. Then, the simulate-and-infer protocol for* $\mathcal{P}$ *requires at most* $O\left(\psi(\mathcal{P}, k) \cdot \frac{k}{2^{\ell}}\right)$ *players, with each player sending at most* ℓ *bits to the referee and the overall error probability at most* $2/5$.

Proof. The reduction is quite straightforward, and works in the following steps: (i) Partition the players into blocks of size $54k/2^{\ell}$; (ii) run the distributed simulation protocol on each block; and (iii) run the centralized algorithm over the simulated samples. Recall from the previous section that we have a Las Vegas protocol for distributed simulation using $27k/2^{\ell}$ players in expectation. Thus, by Markov's inequality, each block in the above protocol simulates a sample with probability at least 1/2. If the number of samples simulated is larger than $\psi(\mathcal{P}, k)$, then the algorithm has error at most 1/3. Denoting the number of blocks by B, the number of samples produced has expectation at least $B/2$, and variance at most $B/4$. By Chebychev's inequality, the probability that the number of samples simulated being less than $B/2 - \sqrt{B/4}\sqrt{15}$ is at most 1/15. If $B > 4\psi(\mathcal{P}, k) + 8$, then $B/2 - \sqrt{B}\sqrt{15/4} > \psi(\mathcal{P}, k)$. As $1/3 + 1/15 = 2/5$, the result follows from a union bound. □

As immediate corollaries of the result, we obtain distributed inference protocols for distribution learning and uniformity testing. Specifically, using the well-known result that $\Theta\left(k/\varepsilon^2\right)$ samples are sufficient to learn a distribution over $[k]$ to within a total variation distance ε with probability 2/3, we obtain **??**.

Next, from the existence of uniformity testing algorithms using $O(\sqrt{k}/\varepsilon^2)$ samples (**???**), we obtain Corollary 1.5

for uniformity testing. The result for identity testing follows using the reduction from (**?**).

Interestingly, a byproduct of this connection between simulate-and-infer and distribution learning (more precisely, of **??**) is that our α-simulation protocol requires the optimal number of players, up to constants.

Corollary 5.2. *Let $\ell \in \{1, \dots, \log k\}$, and $\alpha \in (0, 1)$. Then, any ℓ-bit public-coin (possibly adaptive) α-simulation protocol for k-ary distributions must have $n = \Omega(k/2^\ell)$ players.*

Remark 5.3. We note that the learning upper bound of **??** coincides with the one reported in (**?**), although the latter was obtained using a different technique. The authors of (**?**) also describe a distributed protocol for distribution learning, but their criterion is the ℓ_2 distance instead of total variation.[3] Finally, the learning lower bound we invoke in the proof of Corollary 5.4 is established by adapting a similar lower bound from (**?**) which, too, applied to learning in the ℓ_2 metric.

6. Public-Coin Uniformity Testing

In this section, we consider public-coin protocols for (k, ε)-uniformity testing and establish the following upper and lower bounds for the required number of players.

Theorem 6.1. *For $1 \le \ell \le \log k$, there exists an ℓ-bit public-coin (k, ε)-uniformity testing protocol for $n = O\big(\frac{k}{2^{\ell/2}\varepsilon^2}\big)$ players.*

Note that this is much fewer than the $O(k^{3/2}/(2^\ell\varepsilon^2))$ players required by simulate-and-infer, and indeed by any private-coin using the private-coin uniformity testing lower bound from (**?**). In fact, public-coin uniformity testing lower bound from (**?**) shows that the required number of players is optimal up to constant factors.

We establish Theorem 6.1 below. Before delving into the proof, we note that the results for uniformity testing imply similar upper and lower bounds for the more general question of *identity testing*, where the goal is to test whether the unknown distribution $\mathbf{p}$ is equal to (versus ε-far from) a reference distribution $\mathbf{q}$ known to all the players.

Corollary 6.2. *For $1 \le \ell \le \log k$, and for any fixed $\mathbf{q} \in \Delta_{[k]}$, there exists an ℓ-bit public-coin $(k, \varepsilon, \mathbf{q})$-identity testing protocol for $n = O\big(\frac{k}{2^{\ell/2}\varepsilon^2}\big)$ players. Further, any ℓ-bit public-coin $(k, \varepsilon, \mathbf{q})$-identity testing protocol must have $\Omega\big(\frac{k}{2^{\ell/2}\varepsilon^2}\big)$ players (in the worst case over $\mathbf{q}$).*

We describe this reduction (similar to that in the non-distributed setting) in Appendix A, further detailing how it actually leads to the stronger notion of "instance-optimal" identity testing in the sense of (**?**).

[3] We note that, based on a preliminary version of our manuscript on arXiv, the ℓ_2 learning upper bound of (**?**) was updated to use a "simulate-and-infer" protocol as well.

We now prove Theorem 6.1. Interestingly, the corresponding protocol is remarkably simple, and, moreover, is "smooth" – that is, no player's output depends too much on any particular symbol from $[k]$ (this in turn could be a desirable feature in some cases, for instance, in privacy-minded settings, to control the sensitivity of the algorithm; or for extensions where a quantization of the samples had to be performed, and one seeks an algorithm robust to the specific choice of quantization). Before delving into the details of this protocol, we mention (as briefly evoked in the introduction) that it can actually be implemented in a *randomness-efficient* way. Indeed, although it at first glance appears to require a significant amount of public randomness, namely $\Theta(k \cdot \ell) = \Omega(k)$ bits, we note that the analysis only relies on properties of the second and fourth moments of some suitable random variables; as such, correctness of the protocol only requires 4-wise independent random bits. This in turn can be implemented with only $O(\log k)$ bits of public randomness.

The protocol will rely on a generalization of the following observation: *if $\mathbf{p}$ is ε-far from uniform, then for a subset $S \subseteq [k]$ of size $\frac{k}{2}$ generated uniformly at random, we have $\mathbf{p}(S) = \frac{1}{2} \pm \Omega(\varepsilon/\sqrt{k})$, with constant probability.* Of course, if $\mathbf{p}$ is uniform, then $\mathbf{p}(S) = \frac{1}{2}$ with probability one. Further, note that this fact is qualitatively tight: for the specific case of $\mathbf{p}$ assigning probability $(1 \pm \varepsilon)/k$ to each element, the bias obtained will be $\frac{1}{2} \pm \Theta(\varepsilon/\sqrt{k})$ with high probability.

As a warm-up, we observe that the above claim immediately suggests a protocol for the case $\ell = 1$: The n players, using their shared randomness, agree on a uniformly random subset $S \subseteq [k]$ of size $k/2$, and send to the referee the bit indicating whether their sample fell into this set. Indeed, if $\mathbf{p}$ is ε-far from uniform, with constant probability all corresponding bits will be $(\varepsilon/\sqrt{k})$-biased, and in this case the referee can detect it with $n = O(k/\varepsilon^2)$ players.[4]

The claim in question, although very natural, is already non trivial to establish due to the dependencies between the different elements randomly assigned to the set S. We refer the reader to Corollary 15 in (**?**) for a proof involving anticoncentration of a suitable random variable, $Z := \sum_{i \in [k]} (\mathbf{p}_i - 1/k) X_i$, with $X_1, \dots, X_k$ being (correlated) Bernoulli random variables summing to $k/2$. At a high-level, the argument goes by analyzing the second and fourth moments of Z, and applying the Paley–Zygmund inequality.

For our purposes, we need to show a generalization of the aforementioned claim, considering balanced partitions into $L := 2^\ell$ pieces instead of 2. To do so, we first set up some notation. Let $L < k$ be an integer; for simplicity and with little loss of generality, assume that L divides k.

[4] To handle the small constant probability, it suffices to repeat this independently constantly many times, on disjoint sets of $O(k/\varepsilon^2)$ players.

Further, with $Y_1, \ldots, Y_k$ independent and uniform random variables on $[L]$, let random variables $X_1, \ldots, X_k$ have the same distribution as $Y_1, \ldots, Y_k$ conditioned on the event that for every $r \in [L]$, $\sum_{i=1}^{k} \mathbb{1}_{\{Y_i = r\}} = \frac{k}{L}$. Note that each X_i, too, is uniform on $[L]$, but X_is are not independent. For $\mathbf{p} \in \Delta_{[k]}$, define random variables $Z_1, \ldots, Z_L$ as follows:

$$Z_r := \sum_{i=1}^{k} \mathbf{p}_i \mathbb{1}_{\{X_i = r\}} \,. \tag{2}$$

Equivalently, $(Z_1, \ldots, Z_L)$ correspond to the probabilities $(\mathbf{p}(S_1), \ldots, \mathbf{p}(S_L))$ where $S_1, \ldots, S_L$ is a uniformly random partition of $[k]$ into L sets of equal size.

Theorem 6.3. *For the (random) distribution $\mathbf{q} = (Z_1, \ldots, Z_L)$ over $[L]$ induced by $(Z_1, \ldots, Z_L)$ above, the following holds: (i) if $\mathbf{p} = \mathbf{u}$, then $\|\mathbf{q} - \mathbf{u}_L\|_2 = 0$ with probability one; and (ii) if $\ell_1(\mathbf{p}, \mathbf{u}) > \varepsilon$, then*

$$\Pr\Big[\|\mathbf{q} - \mathbf{u}_L\|_2^2 > \varepsilon^2/k \Big] \geq c \,,$$

for some absolute constant $c > 0$.

The proof of this theorem is quite technical and is deferred to Appendix C. We now explain how it yields a protocol with the desired guarantees (i.e., matching the bounds of Theorem 6.1). By Theorem 6.4, setting $L = 2^\ell$ we get that with constant probability the induced distribution $\mathbf{q}$ on $[L]$ is either uniform (if $\mathbf{p}$ was), or at ℓ_2 distance at least ε' from uniform, where $\varepsilon' := \sqrt{\varepsilon^2/k}$.[5] However, testing uniformity vs. $(\gamma/\sqrt{L})$-farness from uniformity in ℓ_2 distance, over $[L]$, has sample complexity $O(\sqrt{L}/\gamma^2)$ (see e.g., Proposition 3.1 of (**?**) or Theorem 2.10 of (**?**)), and for our choice of $\gamma := \sqrt{L}\varepsilon' \in (0, 1)$, we have

$$\frac{\sqrt{L}}{\gamma^2} = \frac{\sqrt{L}}{L\varepsilon'^2} = \frac{k}{\sqrt{L}\varepsilon^2} = \frac{k}{2^{\ell/2}\varepsilon^2}, \tag{3}$$

giving the bound we sought. This is the idea underlying the following result:

Corollary 6.4. *For $1 \leq \ell \leq \log k$, there exists an ℓ-bit public-coin (k, ε)-uniformity testing protocol for $n = O\big(\frac{k}{2^{\ell/2}\varepsilon^2}\big)$ players, which uses $O(\ell k)$ bits of randomness.*

Proof. The protocol proceeds as follows: Let $m = \Theta(1)$ be an integer such that $(1 - c)^m \leq 1/6$, where c is the constant from Theorem 6.4; define $\delta := 1/(6m)$. Let $N = \Theta(k/(2^{\ell/2}\varepsilon^2))$ be the number of samples sufficient to test $(\varepsilon/\sqrt{k})$-farness in ℓ_2 distance from the uniform distribution over $[L]$, with failure probability δ (as guaranteed by (7)). Finally, let $n := mN = \Theta(k/(2^{\ell/2}\varepsilon^2))$. Given n players, the protocol divides them into m disjoint batches of N players, and each group acts independently as follows:

- Using their shared randomness, the players choose uniformly at random a partition Π of $[k]$ into subsets of size $k/2^\ell$.
- Next, they send to the referee the ℓ bits indicating in which part of the partition their observed sample fell.

The referee, receiving these N messages (which correspond to N independent samples of the distribution $\mathbf{q} \in \Delta_{[2^\ell]}$ induced by $\mathbf{p}$ on Π) runs the ℓ_2 uniformity test, with failure probability δ and distance parameter $\varepsilon/\sqrt{k}$. After running these m tests, the referee rejects if any of the batch is rejected, and accepts otherwise.

By a union bound, all these m tests will be correct with probability at least $1 - m\delta = 5/6$. If $\mathbf{p} = \mathbf{u}_k$, then all m batches generate samples from the uniform distribution on $[L]$, and the referee returns accept with probability at least $5/6$. However, if $\mathbf{p}$ is ε-far from uniform then with probability at least $1 - (1 - c)^m \geq 5/6$ at least one of the m groups will choose a partition such that the corresponding induced distribution on $[L]$ is at ℓ_2 distance at least $\varepsilon/\sqrt{k}$ from uniform; by a union bound, this implies the referee will return reject with probability at least $1 - 2 \cdot 1/6 = 2/3$.

The bound on the total amount of randomness required comes from the fact that $m = \Theta(1)$ independent partitions of $[k]$ into $L := 2^\ell$ are chosen and each such partition can be specified using $O(\log(L^k)) = O(k \cdot \ell)$ bits. □

Acknowledgments. JA is supported by NSF-CCF-1846300 (CAREER), CC is supported by a Motwani Fellowship, and HT is supported in part by the Bosch Research and Technology Centre, Bangalore, India under the Project E-Sense. The authors would like to thank the organizers of the 2018 Information Theory and Applications Workshop (ITA), where the collaboration leading to this work started.

[5] Note that here ℓ_2 and χ^2 distances are equivalent, as the reference distribution is the uniform one. With this in mind, the result we establish can be seen as a random hashing of the k-ary alphabet into L elements, which preserves the χ^2 distance to uniform of each distribution with constant probability.

A. From uniformity to parameterized identity testing

In this appendix, we explain how the existence of any distributed protocol for uniformity testing implies the existence of one for identity testing with roughly the same parameters, and further even implies one for identity testing in the *massively parameterized* sense[6] ("instance-optimal" in the vocabulary of Valiant and Valiant, who introduced it (**?**)). These two results will be seen as a straightforward consequence of (**?**), which establishes the former reduction in the standard non-distributed setting; and of (**?**), which implies that massively parameterized identity testing reduces to "worst-case" identity testing. Specifically, we show the following:

Proposition A.1. *Suppose that there exists an ℓ-bit protocol π for testing uniformity of k-ary distributions, with number of players $n(k,\ell,\varepsilon)$ and failure probability $1/3$. Then there exists an ℓ-bit protocol π' for testing identity against a fixed k-ary distribution $\mathbf{q}$ (known to all players), with number of players $n(5k,\ell,\frac{16}{25}\varepsilon)$ and failure probability $1/3$.*

Furthermore, this reduction preserves the setting of randomness (i.e., private-coin protocols are mapped to private-coin protocols).

Proof. We rely on the result of Goldreich (**?**), which describes a randomized mapping $F_{\mathbf{q}}\colon \Delta_{[k]} \to \Delta_{[5k]}$ such that $F_{\mathbf{q}}(\mathbf{q}) = \mathbf{u}_{[5k]}$ and $\mathrm{d}_{\mathrm{TV}}\big(F_{\mathbf{q}}(\mathbf{p}), \mathbf{u}_{[5k]}\big) > \frac{16}{25}\varepsilon$ for any $\mathbf{p} \in \Delta_{[k]}$ ε-far from $\mathbf{q}$.[7] In more detail, this mapping proceeds in two stages: the first allows one to assume, at essentially no cost, that the reference distribution $\mathbf{q}$ is "grained," i.e., such that all probabilities $\mathbf{q}(i)$ are a multiple of $1/m$ for some $m = O(k)$. Then, the second mapping transforms a given m-grained distribution to the uniform distribution on an alphabet of slightly larger cardinality. The resulting $F_{\mathbf{q}}$ is the composition of these two mappings.

Moreover, a crucial property of $F_{\mathbf{q}}$ is that, given the knowledge of $\mathbf{q}$, a sample from $F_{\mathbf{q}}(\mathbf{p})$ can be efficiently simulated from a sample from $\mathbf{p}$; this implies the proposition. □

Remark A.2. The result above crucially assumes that every player has explicit knowledge of the reference distribution $\mathbf{q}$ to be tested against, as this knowledge is necessary for them to simulate a sample from $F_{\mathbf{q}}(\mathbf{p})$ given their sample from the unknown $\mathbf{p}$. If only the referee $\mathcal{R}$ is assumed to know $\mathbf{q}$, then the above reduction does not go through, although one can still rely on any testing scheme based on distributed simulation.

The previous reduction enables a distributed test for any identity testing problem using at most, roughly, as many players as that required for distributed uniformity testing. However, we can expect to use fewer players for specific distributions. Indeed, in the standard, non-distributed setting, Valiant and Valiant in (**?**) introduced a refined analysis termed the instance-optimal setting and showed that the sample complexity of testing identity to $\mathbf{q}$ is essentially captured by the $2/3$-quasinorm of a sub-function of $\mathbf{q}$ obtained as follows: Assuming without loss of generality $\mathbf{q}_1 \geq \mathbf{q}_2 \geq \ldots \mathbf{q}_k \geq 0$, let $t \in [k]$ be the largest integer that $\sum_{i=t+1}^{k} q_i \geq \varepsilon$, and let $\mathbf{q}_\varepsilon = (\mathbf{q}_2, \ldots, \mathbf{q}_t)$ (i.e., removing the largest element and the "tail" of $\mathbf{q}$). The main result in (**?**) shows that the sample complexity of testing identity to $\mathbf{q}$ is upper and lower bounded by $\max(\|\mathbf{q}_{\varepsilon/16}\|_{2/3}/\varepsilon^2, 1/\varepsilon)$ and $\max(\|\mathbf{q}_\varepsilon\|_{2/3}/\varepsilon^2, 1/\varepsilon)$, respectively.

However, it is not clear if the aforementioned reduction between identity and uniformity of Goldreich preserves this parameterization of sample complexity for identity testing; in particular, the $2/3$-quasinorm characterization does not seem to be amenable to the same type of analysis as that underlying Proposition A.1. Interestingly, a different instance-optimal characterization due to Blais, Canonne, and Gur (**?**) admits such a reduction, enabling us to obtain the analogue of Proposition A.1 for this massively parameterized setting.

To state the result as parameterized by $\mathbf{q}$ (instead of k), we will need the following definition of $\Phi(\mathbf{p},\gamma)$; see Section 6 of (**?**) for a discussion on basic properties of $\Phi(\mathbf{p},\gamma)$ and how it relates to notions such as the sparsity of $\mathbf{p}$ and the functional $\|\mathbf{p}_\gamma^{-\max}\|$ defined in (**?**). For $a \in \ell_2(\mathbb{N})$ and $t \in (0,\infty)$, let

$$\kappa_a(t) := \inf_{a'+a''=a} \left(\|a'\|_1 + t\|a''\|_2\right)$$

and, for $\mathbf{p} \in \Delta_{\mathbb{N}}$ and any $\gamma \in (0,1)$, let

$$\Phi(\mathbf{p},\gamma) := 2\kappa_{\mathbf{p}}^{-1}(1-\gamma)^2\,. \tag{4}$$

It can be seen that, if $\mathbf{p}$ is supported on at most k elements, $\Phi(\mathbf{p},\gamma) \leq 2k$ for all $\gamma \in (0,1)$. We are now in a position to state our general reduction.

Proposition A.3. *Suppose that there exists an ℓ-bit protocol π for testing uniformity of k-ary distributions, with number of players $n(k,\ell,\varepsilon)$ and failure probability $1/3$. Then there exists an ℓ-bit protocol π' for testing identity against a fixed distribution $\mathbf{p}$ (known to all players), with number of players $O\big(n(\Phi(\mathbf{q},\frac{\varepsilon}{9}),\ell,\frac{\varepsilon}{18}))\big)$ and failure probability $2/5$.*

Further, this reduction preserves the setting of randomness

[6] Massively parameterized setting, a terminology borrowed from property testing, refers here to the fact that the sample complexity depends not only on a single parameter k but a k-ary distribution $\mathbf{q}$.

[7] In (**?**), Goldreich exhibits a randomized mapping that converts the problem from testing identity over domain of size k with proximity parameter ε to testing uniformity over a domain of size $k' := k/\alpha^2$ with proximity parameter $\varepsilon' := (1-\alpha)^2\varepsilon$, for every fixed choice of $\alpha \in (0,1)$. This mapping further preserves the success probability of the tester. Since the resulting uniformity testing problem has sample complexity $\Theta\Big(\sqrt{k'}/\varepsilon'^2\Big)$, the blowup factor $1/(\alpha(1-\alpha)^4)$ is minimized by $\alpha = 1/5$.

(i.e., private-coin protocols are mapped to private-coin protocols).

Proof. This strengthening of Proposition A.1 stems from the algorithm for identity testing given in (**?**), which at a high-level reduces testing identity to $\mathbf{q}$ to three tasks: (i) computing the $(\varepsilon/3)$-effective support[8] of $\mathbf{q}$, $S_{\mathbf{q}}(\varepsilon)$, which can be done easily given explicit knowledge of $\mathbf{q}$; (ii) testing that the unknown distribution $\mathbf{p}$ puts mass at most $\varepsilon/2$ outside of $S_{\mathbf{q}}(\varepsilon)$ (which only requires $O(1/\varepsilon)$ players to be done with a high constant probability, say $1/30$); and (iii) testing identity of $\mathbf{p}$ and $\mathbf{q}$ conditioned on $S_{\mathbf{q}}(\varepsilon)$ with parameter $\varepsilon/18$, which can be done using rejection sampling and Proposition A.1 with $O\big(n(|S_{\mathbf{q}}(\varepsilon)|, \ell, \frac{\varepsilon}{18})\big)$ players and success probability, say $2/3 - 1/30$, where the additional $1/30$ error probability comes from rejection sampling. See Fig. 1 for an illustration.

As shown in Section 7.2 of (**?**), we have $|S_{\mathbf{q}}(\varepsilon)| \le \Phi(\mathbf{q}, \frac{\varepsilon}{9})$, and thereby the claimed result, since it follows that the approach above indeed yields an algorithm which is instance-optimal. Technically, the claimed bound is obtained upon recalling that $n(\Phi(\mathbf{q}, \frac{\varepsilon}{9}), \ell, \frac{\varepsilon}{18})) = \Omega(1/\varepsilon)$ using the trivial lower bound of $\Omega(1/\varepsilon)$ on uniformity testing, so that $n(\Phi(\mathbf{q}, \frac{\varepsilon}{9}), \ell, \frac{\varepsilon}{18})) + O(1/\varepsilon) = O\big(n(\Phi(\mathbf{q}, \frac{\varepsilon}{9}), \ell, \frac{\varepsilon}{18})\big)$. □

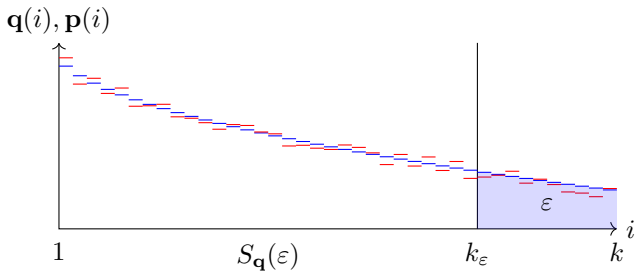

Figure 1. The reference distribution $\mathbf{q}$ (in blue; assumed non-increasing without loss of generality) and the unknown distribution $\mathbf{p}$ (in red). By the reduction above, testing equality of $\mathbf{p}$ to $\mathbf{q}$ is tantamount to (i) determining $S_{\mathbf{q}}(\varepsilon)$, which depends only on $\mathbf{q}$; (ii) testing identity for the conditional distributions of $\mathbf{p}$ and $\mathbf{q}$ given $S_{\mathbf{q}}(\varepsilon)$, and (iii) testing that $\mathbf{p}$ assigns at most $O(\varepsilon)$ probability to the complement of $S_{\mathbf{q}}(\varepsilon)$.

B. Impossibility of perfect simulation when $\ell < \log k$

We begin with a proof of impossibility which shows that any simulation that works for all points in the interior of the $(k-1)$-dimensional probability simplex must fail for a distribution on the boundary. Our main result of this section is the following:

[8] Recall the *ε-effective support* of a distribution $\mathbf{q}$ is the minimal set of elements accounting for at least $1-\varepsilon$ probability mass of $\mathbf{q}$.

Theorem B.1. *For any $n \ge 1$, there exists no ℓ-bit public-coin perfect simulation of k-ary distributions using n players unless $\ell \ge \log k$.*

Proof. Let $\mathcal{S} = (\pi, \delta)$ be an ℓ-bit perfect simulation for k-ary distributions using n players. Suppose that $\ell < \log k$. We show a contradiction for any such public-coin simulation $\mathcal{S}$. Fix a realization $U = u$ of the public randomness. By the pigeonhole principle we can find a message vector $m = (m_1, \ldots, m_n)$ and distinct elements $x_i, x_i' \in [k]$ for each $i \in [n]$ such that

$$\pi_i(x_i, u) = \pi_i(x_i', u) = m_i.$$

Note that the probability of declaring $\perp$ for a public-coin simulation must be 0 for every k-ary distribution. Therefore, since the message m occurs with a positive probability under a distribution $\mathbf{p}$ with $\mathbf{p}_{x_i} > 0$ for all i, the referee must declare an output $x \in [k]$ with positive probability when it receives m, i.e., there exists $x \in [k]$ such that $\delta_x(m, u) > 0$. Also, since x_i and x_i' are distinct for each i, we can assume without loss of generality that $x_i \ne x$ for each i. Now, consider a distribution $\mathbf{p}$ such that $\mathbf{p}_x = 0$ and $\mathbf{p}_{x_i} > 0$ for each i. For this case, the referee must never declare $\mathbf{p}_x$, i.e., $\Pr\Big[\hat{X} = x\Big] = 0$. In particular, $\Pr\Big[\hat{X} = x \;\Big|\; U = u\Big]$ must be 0, which can only happen if $\Pr[\, M = m \mid U = u\,] = 0$. But since $\mathbf{p}_{x_i} > 0$ for each i,

$$\Pr[\, M = m \mid U = u\,] \ge \prod_{i=1}^{n} \mathbf{p}_{x_i} > 0\,,$$

which is a contradiction. □

Note that the proof above shows, as stated before, that any perfect simulation that works for every $\mathbf{p}$ in the interior of the $(k-1)$-dimensional probability simplex, must fail at one point on the boundary of the simplex. In fact, a much stronger impossibility result holds. We show next that for $k = 3$ and $\ell = 1$, we cannot find a perfect simulation that works in the neighborhood of any point in the interior of the simplex.

Theorem B.2. *For any $n \ge 1$, there does not exist any ℓ-bit perfect simulation of 3-ary distributions unless $\ell \ge 2$, even under the promise that the input distribution comes from an open set in the interior of the probability simplex.*

Before we prove the theorem, we show that there is no loss of generality in restricting to *deterministic* protocols, namely protocols where each player uses a deterministic function of its observation to communicate. The high-level argument is relatively simple: By replacing player j by two players j_1, j_2, each with a suitable deterministic strategy, the two 1-bit messages received by the referee will allow him to simulate player j's original randomized mapping.

Lemma B.3. *For $\mathcal{X} = \{0,1,2\}$, suppose there exists a 1-bit perfect simulation $S' = (\pi', \delta')$ with n players. Then, there is a 1-bit perfect simulation $S = (\pi, \delta)$ with $2n$ players such that, for each $j \in [2n]$, the communication π is deterministic, i.e., for each realization u of public randomness*

$$\pi_j(x_j, u) = \pi_j(x), \qquad x \in \mathcal{X}\,.$$

Proof. Consider a mapping $f\colon \{0,1,2\}\times\{0,1\}^* \to \{0,1\}$. We will show that we can find mappings $g_1\colon \{0,1,2\} \to \{0,1\}$, $g_2\colon \{0,1,2\} \to \{0,1\}$, and $h\colon \{0,1\} \times \{0,1\} \times \{0,1\}^* \to \{0,1\}$ such that for every u

$$\Pr[\, f(X,u) = 1\,] = \Pr[\, h(g_1(X_1), g_2(X_2), u) = 1\,], \tag{5}$$

where random variables X_1, X_2, X are independent and identically distributed and take values in $\{0,1,2\}$. We can then use this construction to get our claimed simulation S using $2n$ players as follows: Replace the communication $\pi'_j(x,u)$ from player j with communication $\pi_{2j-1}(x_{2j-1})$ and $\pi_{2j}(x_{2j})$, respectively, from two players $2j-1$ and $2j$, where π_{2j-1} and π_{2j} correspond to mappings g_1 and g_2 above for $f = \pi'_j$. The referee can then emulate the original protocol using the corresponding mapping h and using $h(\pi_{2j-1}(x_{2j-1}), \pi_{2j}(x_{2j}), u)$ in place of communication from player j in the original protocol (recall that, the protocol being known to all parties, the referee knows the mapping $f = \pi'_j$ and thus can implement this strategy). Then, since the probability distribution of the communication does not change, we retain the performance of S', but using only deterministic communication now.

Therefore, it suffices to establish (2). For convenience, denote $\alpha_u := \mathbb{1}_{\{f(0,u)=1\}}$, $\beta_u := \mathbb{1}_{\{f(1,u)=1\}}$, and $\gamma_u := \mathbb{1}_{\{f(2,u)=1\}}$. Assume without loss of generality that $\alpha_u \le \beta_u + \gamma_u$; then, $(\beta_u + \gamma_u - \alpha_u) \in \{0,1\}$. Let $g_i(x) = \mathbb{1}_{\{x=i\}}$ for $i \in \{1,2\}$. Consider the mapping h given by

$$\begin{aligned} h(0,0,u) = \alpha_u,\ h(1,0,u) = \beta_u,\\ h(0,1,u) = \gamma_u,\ h(1,1,u) = (\beta_u + \gamma_u - \alpha_u)\,. \end{aligned}$$

Then, for every u,

$$\begin{aligned} &\Pr[\, h(g_1(X_1), g_2(X_2), u) = 1\,]\\ &= \alpha_u(1-\mathbf{p}_1)(1-\mathbf{p}_2) + \beta_u(1-\mathbf{p}_1)\mathbf{p}_2\\ &\qquad + \gamma_u\mathbf{p}_1(1-\mathbf{p}_2) + (\beta_u + \gamma_u - \alpha_u)\mathbf{p}_1\mathbf{p}_2\\ &= \alpha_u(1-\mathbf{p}_1-\mathbf{p}_2) + \beta_u\mathbf{p}_2 + \gamma_u\mathbf{p}_1\\ &= \Pr[\, f(X,u) = 1\,], \end{aligned}$$

which completes the proof. □

We now prove Theorem 4.2, but in view of our previous observation, we only need to consider deterministic communication.

Proof of Theorem 4.2. Suppose by contradiction that there exists such a 1-bit perfect simulation protocol $S = (\pi, \delta)$ for n players on $\mathcal{X} = \{0,1,2\}$ such that $\pi(x,u) = \pi(x)$. Assume that this protocol is correct for all distributions $\mathbf{p}$ in the neighborhood of some $\mathbf{p}^*$ in the interior of the simplex. Consider a partition the players into three sets $\mathcal{S}_0$, $\mathcal{S}_1$, and $\mathcal{S}_2$, with

$$\mathcal{S}_i := \{\, j \in [n] :\ \pi_j(i) = 1\,\}\,, \qquad i \in \mathcal{X}\,.$$

Note that for deterministic communication the message M is independent of public randomness U. Then, by the definition of perfect simulation, it must be the case that

$$\begin{aligned} \mathbf{p}_x &= \mathbb{E}_U \sum_{m\in\{0,1\}^n} \delta_x(m,U)\Pr[\, M = m \mid U\,] \qquad (6)\\ &= \mathbb{E}_U \sum_m \delta_x(m,U)\Pr[\, M = m\,]\\ &= \sum_m \mathbb{E}_U[\delta_x(m,U)]\Pr[\, M = m\,] \qquad (7) \end{aligned}$$

for every $x \in \mathcal{X}$, which with our notation of $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2$ can be re-expressed as

$$\begin{aligned} &\mathbf{p}_x\\ &= \sum_{m\in\{0,1\}^n} \mathbb{E}_U[\delta_x(m,U)] \prod_{i=0}^{2}\prod_{j\in\mathcal{S}_i} (m_j\mathbf{p}_i + (1-m_j)(1-\mathbf{p}_i))\\ &= \sum_{m\in\{0,1\}^n} \mathbb{E}_U[\delta_x(m,U)] \prod_{i=0}^{2}\prod_{j\in\mathcal{S}_i} (1 - m_j + (2m_j-1)\mathbf{p}_i)\,, \end{aligned}$$

for every $x \in \mathcal{X}$. But since the right-side above is a polynomial in $(\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2)$, it can only be zero in an open set in the interior if it is identically zero. In particular, the constant term must be zero:

$$\begin{aligned} 0 &= \sum_{m\in\{0,1\}^n} \mathbb{E}_U[\delta_x(m,U)] \prod_{i=0}^{2}\prod_{j\in\mathcal{S}_i} (1-m_j)\\ &= \sum_{m\in\{0,1\}^n} \mathbb{E}_U[\delta_x(m,U)] \prod_{j=1}^{n} (1-m_j)\,. \end{aligned}$$

Noting that every summand is non-negative, this implies that for all $x \in \mathcal{X}$ and $m \in \{0,1\}^n$, $\mathbb{E}_U[\delta_x(m,U)]\prod_{j=1}^n(1-m_j) = 0$. In particular, for the all-zero message $\mathbf{0}^n$, we get $\mathbb{E}_U[\delta_x(\mathbf{0}^n, U)] = 0$ for all $x \in \mathcal{X}$, so that again by non-negativity we must have $\delta_x(\mathbf{0}^n, u) = 0$ for all $x \in \mathcal{X}$ and randomness u. But the message $\mathbf{0}^n$ will happen with probability

$$\begin{aligned} \Pr[\, M = \mathbf{0}^n\,] &= \prod_{i=0}^{2}\prod_{j\in\mathcal{S}_i}(1-\mathbf{p}_i)\\ &= (1-\mathbf{p}_0)^{|\mathcal{S}_0|}(1-\mathbf{p}_1)^{|\mathcal{S}_1|}(1-\mathbf{p}_2)^{|\mathcal{S}_2|} > 0, \end{aligned}$$

where the inequality holds since $\mathbf{p}$ lies in the interior of the simplex. Therefore, for the output $\hat{X}$ of the referee we have

$$\begin{aligned}\Pr\Big[\,\hat{X} \neq \perp\,\Big] &= \sum_{m}\sum_{x\in\mathcal{X}} \mathbb{E}_U[\delta_x(m,U)]\cdot \Pr[\,M=m\,] \\ &= \sum_{m\neq \mathbf{0}^n} \Pr[\,M=m\,] \sum_{x\in\mathcal{X}} \mathbb{E}_U[\delta_x(m,U)] \\ &\leq \sum_{m\neq \mathbf{0}^n} \Pr[\,M=\mathbf{0}^n\,] \\ &= 1-\Pr[\,M=\mathbf{0}^n\,] < 1\,,\end{aligned}$$

contradicting the fact that π is a perfect simulation protocol. □

Remark B.4. It is unclear how to extend the proof of Theorem 4.2 arbitrary k, ℓ. In particular, the proof of Lemma 4.3 does not extend to the general case. A plausible proof-strategy is a black-box application of the $k = 3$, $\ell = 1$ result to obtain the general result using a direct-sum-type argument.

We close this section by noting that perfect simulation is impossible even when the communication from each player is allowed to depend on that from the previous ones. Specifically, we show that availability of such an interactivity can at most bring an exponential improvement in the number of players.

Lemma B.5. *For every $n \geq 1$, if there exists an interactive public-coin ℓ-bit perfect simulation of k-ary distributions with n players, then there exists a public-coin ℓ-bit perfect simulation of k-ary distributions with $2^{\ell n+1}$ players that uses only SMP.*

Proof. Consider an interactive communication protocol π for distributed simulation with n players and ℓ bits of communication per player. We can view the overall protocol as a (2^ℓ)-ary tree of depth n where player j is assigned all the nodes at depth j. An execution of the protocol is a path from the root to the leaf of the tree. Suppose the protocol starting at the root has reached a node at depth j, then the next node at depth $j+1$ is determined by the communication from player j. Thus, this protocol can be simulated non-interactively using at most $((2^\ell)^n - 1)/(2^\ell - 1) < 2^{\ell n+1}$ players, where players $(2^{j-1}+1)$ to 2^j send all messages correspond to nodes at depth j in the tree. Then, the referee receiving all the messages can output the leaf by following the path from root to the leaf. □

Corollary B.6. *Theorems 4.1 and 4.2 extend to interactive protocols as well.*

C. Distributed Simulation with one bit

Proof of Theorem 4.7. To help the reader build heuristics for the proof, we describe the protocol and analyze its performance in steps. We begin by describing the basic idea and building blocks; we then build upon it to obtain a full-fledged protocol, but with potentially unbounded expected number of players used. Finally, we describe a simple modification which yields our desired bound for expected number of player's accessed.

The scheme, base version. Consider a protocol with $2k$ players where the 1-bit communication from players $(2i-1)$ and $(2i)$ just indicates if their observation is i or not, namely $\pi_{2i-1}(x) = \pi_{2i}(x) = \mathbb{1}_{\{x=i\}}$.

On receiving these $2k$ bits, the referee $\mathcal{R}$ acts as follows:

- if exactly one of the bits $M_1, M_3, \ldots, M_{2k-1}$ is equal to one, say the bit M_{2i-1}, and the corresponding bit M_{2i} is zero, then the referee outputs $\hat{X} = i$;
- otherwise, it outputs $\perp$.

In the above, the probability $\rho_{\mathbf{p}}$ that some $i \in [k]$ is declared as the output (and not $\perp$) is

$$\rho_{\mathbf{p}} := \sum_{i=1}^{k} (1-\mathbf{p}_i)\cdot \mathbf{p}_i \prod_{j\neq i}(1-\mathbf{p}_j) = \prod_{j=1}^{k}(1-\mathbf{p}_j),$$

so that

$$\begin{aligned}\rho_{\mathbf{p}} &= \exp\sum_{j=1}^{k}\ln(1-\mathbf{p}_j) = \exp\left(-\sum_{t=1}^{\infty}\frac{\|\mathbf{p}\|_t^t}{t}\right) \\ &\geq \exp\left(-\left(1+\sum_{t=2}^{\infty}\frac{\|\mathbf{p}\|_2^t}{t}\right)\right) = \frac{1-\|\mathbf{p}\|_2}{e^{1-\|\mathbf{p}\|_2}}\end{aligned}$$

which is bounded away from 0 as long as $\mathbf{p}$ is far from being a point mass (i.e., $\|\mathbf{p}\|_2$ is not too close to 1).

Further, for any fixed $i \in [k]$, the probability that $\mathcal{R}$ outputs i is

$$\mathbf{p}_i \cdot \prod_{j=1}^{k}(1-\mathbf{p}_j) = \mathbf{p}_i\rho_{\mathbf{p}} \propto \mathbf{p}_i\,.$$

The scheme, medium version. The (almost) full protocol proceeds as follows. Divide the countably infinitely many players into successive, disjoint batches of $2k$ players each, and apply the base scheme to each of these runs. Execute the base scheme to each of the batch, one at a time and moving to the next batch only when the current batch declares a $\perp$; else declare the output of the batch as $\hat{X}$.

It is straightforward to verify that the distribution of the output $\hat{X}$ is exactly $\mathbf{p}$, and moreover that on expectation $1/\rho_{\mathbf{p}}$ runs are considered before a sample is output. Therefore, the expected number of players accessed (i.e., bits considered by the referee) satisfies

$$\frac{2k}{\rho_{\mathbf{p}}} \leq 2k\cdot\frac{e^{1-\|\mathbf{p}\|_2}}{1-\|\mathbf{p}\|_2}\,. \tag{8}$$

The scheme, final version. The protocol described above can have the expected number of players blowing to infinity when $\mathbf{p}$ has ℓ_2 norm close to one. To circumvent this

difficulty, we modify the protocol as follows: Consider the distribution $\mathbf{q}$ on $[2k]$ defined by

$$\mathbf{q}_{2i} = \mathbf{q}_{2i-1} = \frac{\mathbf{p}_i}{2}, \qquad i \in [k]\,.$$

Clearly, $\|\mathbf{q}\|_2 = \|\mathbf{p}\|_2/\sqrt{2} \le 1/\sqrt{2}$, and therefore by (4) the expected number of players required to simulate $\mathbf{q}$ using our previous protocol is at most

$$4k \cdot \frac{e^{1-1/\sqrt{2}}}{1-1/\sqrt{2}} \le 20k.$$

But we can simulate a sample from $\mathbf{p}$ using a sample from $\mathbf{q}$ simply by mapping $(2i-1)$ and $2i$ to i. The only thing remaining now is to simulate samples from $\mathbf{q}$ using samples from $\mathbf{p}$. This, too, is easy. Every 2 players in a batch that declare 1 on observing symbols $(2i-1)$ and $(2i)$ from $\mathbf{q}$ declare 1 when they see i from $\mathbf{p}$. The referee then simply flips each of this 1 to 0, thereby simulating the communication corresponding to samples from $\mathbf{q}$. In summary, we modified the original protocol for $\mathbf{p}$ by replacing each player with two identical copies and modifying the referee to flip 1 received from these players to 0 independently with probability $1/2$; the output is declared in a batch only when there is exactly one 1 in the modified messages, in which case the output is the element assigned to the player that sent 1. Thus, we have a simulation for k-ary distributions that uses at most $20k$ players, completing the proof of the theorem. □

D. Distributed Simulation for any ℓ

Proof of Theorem 1.3. For simplicity, assume that $2^\ell - 1$ divides k. We can then extend the previous protocol by considering a partition of domain into $m = k/(2^\ell - 1)$ parts and assigning one part of size $2^\ell - 1$ each to a player. Each player then sends the all-zero sequence of length ℓ when it does not see an element from its assigned set, or indicates the precise element from its assigned set that it observed. For each batch, the referee, too, proceeds as before and declares an output if exactly one player in the batch sends a 1 – the declared output is the element indicated by the player that sent a 1; else it moves to the next batch. To bound the number of players, consider the analysis of the base protocol. The probability that an output is declared for a batch (a $\perp$ is not declared in the base protocol) is given by

$$\begin{aligned}
\rho_{\mathbf{p}} &:= \sum_{i=1}^{m} (1 - \mathbf{p}(S_i)) \cdot \sum_{\ell \in S_i} \mathbf{p}_\ell \prod_{j \neq i} (1 - \mathbf{p}(S_j)) \\
&= \prod_{j=1}^{m} (1 - \mathbf{p}(S_j)) \cdot \sum_{i=1}^{m} \sum_{\ell \in S_i} \mathbf{p}_\ell \\
&= \prod_{j=1}^{m} (1 - \mathbf{p}(S_j))\,,
\end{aligned}$$

where $\{S_1, \ldots, S_m\}$ denotes the partition used. Then, writing $\mathbf{p}^{(S)}$ for the distribution on $[m]$ given by $\mathbf{p}^{(S)}(j) = \mathbf{p}(S_j)$, by proceeding as in the $\ell = 1$ case we obtain

$$\rho_{\mathbf{p}} \ge \frac{1 - \|\mathbf{p}^{(S)}\|_2}{e^{1-\|\mathbf{p}^{(S)}\|_2}}\,.$$

Once again, this quantity may be unbounded and we circumvent this difficulty by replacing each player with two players that behave identically and flipping their communicated 1's to 0's randomly at the referee; the output is declared in a batch only when there is exactly one 1 in the modified messages, in which case the output is the element indicated by the player that sent 1. The analysis can be completed exactly in the manner of the $\ell = 1$ case proof by noticing that the protocol is tantamount to simulating $\mathbf{q}$ with $\|\mathbf{q}^{(S)}\|_2 \le 1/\sqrt{2}$ and accesses messages from at most $20m$ players in expectation. □

E. Proof of Theorem 6.4

In this appendix, we prove Theorem 6.4, stating that taking a random balanced partition of the domain in $L \ge 2$ parts preserves the ℓ_2 distance between distributions with constant probability. Note that the special case of $L = 2$ was proven in (**?**). In fact, the proof for general L is similar to the proof in (**?**), but requires some additional work. We provide a self-contained proof here for easy reference.

We begin by recall the Paley–Zigmund inequality, a key tool we shall rely upon.

Theorem E.1 (Paley–Zygmund)**.** *Suppose U is a non-negative random variable with finite variance. Then, for every $\theta \in [0, 1]$,*

$$\Pr[\, U > \theta \mathbb{E}[U]\,] \ge (1-\theta)^2 \frac{\mathbb{E}[U]^2}{\mathbb{E}[U^2]}\,.$$

We will prove a more general version of Theorem 6.4, showing that the ℓ_2 distance to any fixed distribution $\mathbf{q} \in \Delta_{[k]}$ is preserved with a constant probability.[9] Let random variables $X_1, \ldots, X_k$ be as in Theorem 6.4; in particular, each X_i is distributed uniformly on $[L]$ and for every $r \in [L]$, $\sum_{i=1}^{k} \mathbb{1}_{\{X_i = r\}} = \frac{k}{L}$.

Theorem E.2. *Suppose $2 \le L < k$ is an integer dividing k, and fix $\delta \in \mathbb{R}^k$ such that $\sum_{i \in [k]} \delta_i = 0$. For random variables $X_1, ..., X_k$ above, let $Z = (Z_1, \ldots, Z_L) \in \mathbb{R}^L$ with*

$$Z_r := \sum_{i=1}^{k} \delta_i \mathbb{1}_{\{X_i = r\}}, \qquad r \in [L]\,.$$

Then, there exists a constant $c > 0$ such that

$$\Pr\left[\, \|Z\|_2 > \frac{1}{2} \cdot \|\delta\|_2 \,\right] \ge c.$$

[9] For this application, one should read the theorem statement with $\delta := \mathbf{p} - \mathbf{q}$.

Proof of Theorem C.2. As in Theorem 14 of (**?**), the gist of the proof is to consider a suitable non-negative random variable (namely, $\|Z\|_2^2$) and bound its expectation and second moment in order to apply the Paley–Zygmund inequality to argue about anticoncentration around the mean. The difficulty, however, lies in the fact that bounding the moments of $\|Z\|_2$ involves handling the products of correlated L-valued random variables X_i's, which is technical even for the case $L = 2$ considered in (**?**). For ease of presentation, we have divided the proof into smaller results.

Lemma E.3 (Each part has the right expectation)**.** *For every $r \in [L]$,*

$$\mathbb{E}[Z_r] = 0\,.$$

Proof. By linearity of expectation,

$$\mathbb{E}[Z_r] = \sum_{i=1}^{k} \delta_i \mathbb{E}\big[\mathbb{1}_{\{X_i=r\}}\big] = \frac{1}{L}\sum_{i=1}^{k} \delta_i = 0.$$

□

Lemma E.4 (The ℓ_2^2 distance to uniform of the flattening has the right expectation)**.** *For every $r \in [L]$,*

$$\begin{aligned}\operatorname{Var} Z_r &= \mathbb{E}\big[Z_r^2\big] \\ &= \frac{1}{L}\|\delta\|_2^2\left(1 - \frac{1}{L} + \frac{L-1}{L(k-1)}\right) \geq \frac{1}{2L}\|\delta\|_2^2\,.\end{aligned}$$

In particular, the expected squared ℓ_2 norm of Z is

$$\mathbb{E}\Big[\|Z\|_2^2\Big] = \mathbb{E}\left[\sum_{r=1}^{L} Z_r^2\right] \geq \frac{1}{2}\|\delta\|_2^2\,.$$

Proof. For a fixed $r \in [L]$, using the definition of Z, the fact that $\sum_{i=1}^{k} \mathbb{1}_{\{X_i=r\}} = \frac{k}{L}$, and Lemma C.3, we get that

$$\begin{aligned}&\operatorname{Var}[Z_r] \\ &= \mathbb{E}\big[Z_r^2\big] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{k} \delta_i \mathbb{1}_{\{X_i=r\}}\right)^2\right] \\ &= \sum_{1\leq i,j\leq k} \delta_i\delta_j \mathbb{E}\big[\mathbb{1}_{\{X_i=r\}}\mathbb{1}_{\{X_j=r\}}\big] \\ &= \sum_{i=1}^{k} \delta_i^2 \mathbb{E}\big[\mathbb{1}_{\{X_i=r\}}\big] + 2\sum_{1\leq i<j\leq k} \delta_i\delta_j \mathbb{E}\big[\mathbb{1}_{\{X_i=r\}}\mathbb{1}_{\{X_j=r\}}\big]\,.\end{aligned}$$

Since the X_i's – while not independent – are identically distributed, it is enough by symmetry to compute $\mathbb{E}\big[\mathbb{1}_{\{X_k=r\}}\big]$ and $\mathbb{E}\big[\mathbb{1}_{\{X_{k-1}=r\}}\mathbb{1}_{\{X_k=r\}}\big]$. The former is $1/L$; for the latter, note that

$$\begin{aligned}&\mathbb{E}\big[\mathbb{1}_{\{X_{k-1}=r\}}\mathbb{1}_{\{X_k=r\}}\big] && (9)\\ &= \mathbb{E}\big[\mathbb{E}\big[\,\mathbb{1}_{\{X_{k-1}=r\}}\mathbb{1}_{\{X_k=r\}} \;\big|\; \mathbb{1}_{\{X_k=r\}}\,\big]\big] && (10)\\ &= \frac{1}{L}\Pr[\,X_{k-1}=r \mid X_k=r\,] \\ &= \frac{1}{L}\Pr\left[X_{k-1}=r \;\middle|\; \sum_{i=1}^{k-1}\mathbb{1}_{\{X_i=r\}} = \frac{k}{L}-1\right] && (11)\\ &= \frac{1}{L^2}\cdot\frac{k-L}{k-1}, && (12)\end{aligned}$$

where the final identity uses symmetry once again, along with the observation that

$$\sum_{i=1}^{k-1}\mathbb{E}\left[\mathbb{1}_{\{X_i=r\}} \;\middle|\; \sum_{j=1}^{k-1}\mathbb{1}_{\{X_j=r\}} = \frac{k}{L}-1\right] = \frac{k}{L}-1.$$

Putting it together, we get the result as follows:

$$\begin{aligned}\operatorname{Var}[Z_r] &= \frac{1}{L}\sum_{i=1}^{k}\delta_i^2 + \frac{1}{L^2}\cdot\frac{k-L}{k-1}\cdot 2\sum_{1\leq i<j\leq k}\delta_i\delta_j \\ &= \frac{1}{L}\|\delta\|_2^2 - \frac{1}{L^2}\left(1-\frac{L-1}{k-1}\right)\|\delta\|_2^2 \\ &= \frac{1}{L}\|\delta\|_2^2\left(1-\frac{1}{L}+\frac{L-1}{L(k-1)}\right).\end{aligned}$$

□

Lemma E.5 (The ℓ_2^2 distance to uniform of the flattening has the required second moment)**.** *There exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\Big[\|Z\|_2^4\Big] \leq C\|\delta\|_2^4\,.$$

Proof of Lemma C.5. Expanding the square, we have

$$\mathbb{E}\Big[\|Z\|_2^4\Big] = \mathbb{E}\left[\left(\sum_{r=1}^{L} Z_r^2\right)^2\right] = \sum_{r=1}^{L}\mathbb{E}\big[Z_r^4\big] + 2\sum_{r<r'}\mathbb{E}\big[Z_r^2 Z_{r'}^2\big] \tag{13}$$

We will bound both terms separately. For the first term, we note that using Equation(21) of (**?**) with $\mathbb{1}_{\{X_i=r\}}$ in the role of X_i there, each term $\mathbb{E}\big[Z_r^4\big]$ is bounded above by $19\|\delta\|_2^4/L$ whereby

$$\sum_{r=1}^{L}\mathbb{E}\big[Z_r^4\big] \leq 19\|\delta\|_2^4. \tag{14}$$

However, we need additional work to handle the second term comprising roughly L^2 summands. In particular, to complete the proof we show that each summand in the second term is less than a constant factor times $\|\delta\|_2^4/L^2$.

Claim E.6. *There exists an absolute constant $C' > 0$ such that*

$$\sum_{r<r'} \mathbb{E}\big[Z_r^2 Z_{r'}^2\big] \le C' \|\delta\|_2^4 .$$

Proof. Fix any $r \neq r'$. As before, we expand

$$\begin{aligned}
&\mathbb{E}\big[Z_r^2 Z_{r'}^2\big] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^k \delta_i \mathbb{1}_{\{X_i=r\}}\right)^2 \left(\sum_{i=1}^k \delta_i \mathbb{1}_{\{X_i=r'\}}\right)^2\right] \\
&= \sum_{1\le a,b,c,d\le k} \delta_a\delta_b\delta_c\delta_d \mathbb{E}\big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r\}}\mathbb{1}_{\{X_c=r'\}}\mathbb{1}_{\{X_d=r'\}}\big] .
\end{aligned}$$

Using symmetry once again, note that the term $\mathbb{E}\Big[\tilde{X}_a\tilde{X}_b\tilde{X}_c\tilde{X}_d\Big]$ depends only on the number of distinct elements in the multiset $\{a,b,c,d\}$, namely the cardinality $|\{a,b,c,d\}|$. The key observation here is that if $\{a,b\} \cap \{c,d\} \neq \emptyset$, then $\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r\}}\mathbb{1}_{\{X_c=r'\}}\mathbb{1}_{\{X_d=r'\}} = 0$. This will be crucial as it implies that the expected value can only be non-zero if $|\{a,b,c,d\}| \ge 2$, yielding a $1/L^2$ dependence for the leading term in place of $1/L$.

$$\begin{aligned}
&\mathbb{E}\big[Z_r^2 Z_{r'}^2\big] \qquad (15)\\
&= \sum_{|\{a,b,c,d\}|=2} \delta_a^2\delta_b^2 \mathbb{E}\big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r'\}}\big] \\
&+ \sum_{|\{a,b,c,d\}|=3} \delta_a^2\delta_b\delta_c \mathbb{E}\big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r'\}}\mathbb{1}_{\{X_c=r'\}}\big] \\
&+ \sum_{|\{a,b,c,d\}|=3} \delta_a\delta_b\delta_c^2 \mathbb{E}\big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r\}}\mathbb{1}_{\{X_c=r'\}}\big] \\
&+ \sum_{|\{a,b,c,d\}|=4} \delta_a\delta_b\delta_c\delta_d \mathbb{E}\big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r\}}\mathbb{1}_{\{X_c=r'\}}\mathbb{1}_{\{X_d=r'\}}\big] \qquad (16)
\end{aligned}$$

The first term, which we will show dominates, is bounded as

$$\begin{aligned}
&\sum_{|\{a,b,c,d\}|=2} \delta_a^2\delta_b^2 \mathbb{E}\big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r'\}}\big] \\
&= \mathbb{E}\big[\mathbb{1}_{\{X_{k-1}=r\}}\mathbb{1}_{\{X_k=r'\}}\big] \|\delta\|_2^4 \le \frac{2}{L^2}\|\delta\|_2^4
\end{aligned}$$

where the inequality uses

$$\mathbb{E}\big[\mathbb{1}_{\{X_{k-1}=r\}}\mathbb{1}_{\{X_k=r'\}}\big] = \frac{1}{L^2}\cdot\frac{k}{k-1} \le \frac{2}{L^2},$$

which in turn is obtained in the manner of (15).

For the second and the third terms, noting that

$$\mathbb{E}\big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r'\}}\mathbb{1}_{\{X_c=r'\}}\big] = \big|\delta_a^2\delta_b\delta_c\big|\cdot\frac{1}{L^3}\frac{k(k-L)}{(k-1)(k-2)},$$

and that

$$\sum_{|\{a,b,c,d\}|=3} \delta_a^2\delta_b\delta_c = \sum_{1\le a,b,c\le k} \delta_a^2\delta_b\delta_c - \sum_{a\neq b}\delta_a^2\delta_b^2 - 2\sum_{a\neq b}\delta_a^3\delta_b$$

with $\sum_{1\le a,b,c\le k} \delta_a^2\delta_b\delta_c = \left(\sum_{a=1}^k \delta_a^2\right)\left(\sum_{a=1}^k \delta_a\right)^2 = 0$, $\sum_{a\neq b}\delta_a^2\delta_b^2 \le \sum_{1\le a,b\le k}\delta_a^2\delta_b^2 = \|\delta\|_2^4$, and $\sum_{a\neq b}\delta_a^3|\delta_b| \le \sum_{1\le a,b\le k}\delta_a^3|\delta_b| \le \|\delta\|_\infty\|\delta\|_3^3 \le \|\delta\|_2^4$, we get

$$\begin{aligned}
&-\frac{6}{L^3}\|\delta\|_2^4 \\
&\le \sum_{|\{a,b,c,d\}|=3} \delta_a^2\delta_b\delta_c \mathbb{E}\big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r'\}}\mathbb{1}_{\{X_c=r'\}}\big] \\
&\le \frac{6}{L^3}\|\delta\|_2^4 .
\end{aligned}$$

Finally, as $\mathbb{E}\big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r\}}\mathbb{1}_{\{X_c=r'\}}\mathbb{1}_{\{X_d=r'\}}\big] = \frac{1}{L^4}\frac{k^2(k-L)^2}{(k-1)(k-2)(k-3)(k-4)} \le \frac{10}{L^4}$, similar manipulations yield

$$\begin{aligned}
&-\frac{\alpha}{L^4}\|\delta\|_2^4 \\
&\le \sum_{|\{a,b,c,d\}|=4} \delta_a\delta_b\delta_c\delta_d \mathbb{E}\big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r\}}\mathbb{1}_{\{X_c=r'\}}\mathbb{1}_{\{X_d=r'\}}\big] \\
&\le \frac{\alpha}{L^4}\|\delta\|_2^4
\end{aligned}$$

for some absolute constant $\alpha > 0$. Gathering all this in (18), we get that there exists some absolute constant $C' > 0$ such that

$$\sum_{r<r'} \mathbb{E}\big[Z_r^2 Z_{r'}^2\big] \le C' \sum_{r<r'} \frac{1}{L^2}\|\delta\|_2^4 \le \frac{C'}{2}\|\delta\|_2^4 .$$

□

The lemma follows by combining the previous claim with (17). □

We are now ready to establish Theorem 6.4. By Lemmas C.4 to C.5, we have $\mathbb{E}\Big[\|Z\|_2^2\Big] \ge \frac{1}{2}\|\delta\|_2^2$ and $\mathbb{E}\Big[\|Z\|_2^4\Big] \le C\|\delta\|_2^4$, for some absolute constant $C > 0$. Therefore, by the Payley–Zygmund inequality (Theorem C.1) applied to $\|Z\|_2^2$ for $\theta = 1/2$,

$$\begin{aligned}
\Pr\left[\|Z\|_2^2 > \frac{1}{4}\|\delta\|_2^2\right] &\ge \Pr\left[\|Z\|_2^2 > \frac{1}{2}\mathbb{E}\Big[\|Z\|_2^2\Big]\right] \\
&\ge \frac{1}{4}\frac{\mathbb{E}\Big[\|Z\|_2^2\Big]^2}{\mathbb{E}\Big[\|Z\|_2^4\Big]} \ge \frac{1}{16C} .
\end{aligned}$$

This concludes the proof. □