Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 79 (2018)

# Distributed Simulation and Distributed Inference

Jayadev Acharya* Clément L. Canonne† Himanshu Tyagi‡

August 10, 2018

**Abstract**

Independent samples from an unknown probability distribution $\mathbf{p}$ on a domain of size $k$ are distributed across $n$ players, with each player holding one sample. Each player can communicate $\ell$ bits to a central referee in a simultaneous message passing model of communication to help the referee infer a property of the unknown $\mathbf{p}$. What is the least number of players for inference required in the communication-starved setting of $\ell < \log k$? We begin by exploring a general *simulate-and-infer* strategy for such inference problems where the center simulates the desired number of samples from the unknown distribution and applies standard inference algorithms for the collocated setting. Our first result shows that for $\ell < \log k$ perfect simulation of even a single sample is not possible. Nonetheless, we present next a Las Vegas algorithm that simulates a single sample from the unknown distribution using no more than $O(k/2^\ell)$ samples in expectation. As an immediate corollary, it follows that simulate-and-infer attains the optimal sample complexity of $\Theta(k^2/2^\ell\varepsilon^2)$ for learning the unknown distribution to an accuracy of $\varepsilon$ in total variation distance.

For the prototypical testing problem of identity testing, simulate-and-infer works with $O(k^{3/2}/2^\ell\varepsilon^2)$ samples, a requirement that seems to be inherent for all communication protocols not using any additional resources. Interestingly, we can break this barrier using public coins. Specifically, we exhibit a public-coin communication protocol that accomplishes identity testing using $O(k/\sqrt{2^\ell}\varepsilon^2)$ samples. Furthermore, we show that this is optimal up to constant factors. Our theoretically sample-optimal protocol is easy to implement in practice. Our proof of lower bound entails showing a contraction in $\chi^2$ distance of product distributions due to communication constraints and may be of interest beyond the current setting.

*Cornell University. Email: acharya@cornell.edu.

†Stanford University. Email: ccanonne@cs.stanford.edu. Supported by a Motwani Postdoctoral Fellowship.

‡Indian Institute of Science. Email: htyagi@iisc.ac.in.

ISSN 1433-8092

# Contents

# 1 Introduction

A set of sensor nodes are deployed in an active volcano to measure seismic activity. They are connected to a central server over a low bandwidth communication link, but owing to their very limited battery, they can only send a fixed number of short packets. The server seeks to determine if the distribution of the quantized measurements have changed significantly from the one on record. How many sensors must be deployed?

This situation is typical in many emerging sensor network applications as well as other distributed learning scenarios where the data is distributed across multiple clients with limited communication capability. The question above is an instance of the *distributed inference* problem where independent samples from an unknown distribution are given to physically separated players who can only send a limited amount of communication to a central referee. The referee uses the communication to infer some properties of the generating distribution of the samples. A variant of this problem where each player gets different (correlated) coordinates of the independent samples has been studied extensively in the information theory literature (cf. [AC86, Han87, HA98]). The problem described above has itself received significant attention lately in various communities (see, for instance, [BPC$^+$11, BBFM12, Sha14, DGL$^+$17, HÖW18a]), with the objective of performing parameter or density estimation while minimizing the number of players (or, equivalently, the amount of data). Of particular interest to us are the results in [DGL$^+$17, HÖW18a], which consider distribution learning problems. Specifically, it is shown that roughly trivial schemes quantizing and compressing each sample separately turn out to be optimal for *simultaneous message passing* (SMP) communication protocols. One of our goals in this research is to formalize this heuristic and explore its limitations.

To formalize the question, 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 $\ell$-bits each to a referee. A distributed simulation protocol consists of an SMP 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 $\ell < \log k$?

This problem of distributed simulation is connected innately to the aforementioned distributed inference problems where the generating distribution $\mathbf{p}$ is unknown and the referee uses the communication from the players to accomplish a specific inference task $\mathcal{P}$.

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

The formulation above encompasses both density and parameter estimation, as well as distribution testing (see e.g. [Rub12, Can15] and [Gol17] for a survey of property and distribution testing).

Equipped with a distributed simulation, we can accomplish any distributed inference task by simulating as many samples as warranted by the sample complexity of the inference problem. Our objective is to understand when such a *simulate-and-infer* strategy is optimal. We study the distributed simulation problem and apply it to distribution learning and distribution testing. Our results are the most striking for distribution testing, which, to the best of our knowledge, has not been studied in the distributed setting prior to our work.

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

Starting with the distributed simulation problem, we first establish that perfect simulation is impossible using any finite number of players in the communication-starved regime. This establishes an interesting dichotomy where communication from a single party suffices for perfect simulation when $\ell \geq \log k$ and no finite number of parties can accomplish it when $\ell < \log k$. If we allow a small probability of declaring failure, namely Las Vegas schemes, distributed simulation is possible with finitely many players. Indeed, we present such a Las Vegas distributed simulation scheme that requires optimal (up to constant factors) number of players to simulate $k$-ary distributions using $\ell$ bits of communication per player. Moving to the connection between distributed simulation and distributed inference, we exhibit an instance when simulate-and-infer is optimal. Perhaps more interestingly, we even exhibit a case where a simple simulate-and-infer scheme is far from optimal if we allow the communication protocol to use public randomness. As a byproduct, we characterize the least number of players required for distributed uniformity testing in the SMP model. 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 an overview of the literature on distribution learning as well.

### 1.1 Main results

Our first theorem shows that perfect distributed simulation with a finite number of players is impossible:

**Theorem 1.2.** *For every $k \geq 1$ and $\ell < \log k$, there exists no SMP with $\ell$ bits of communication per player for distributed simulation over $[k]$ with finite number of players. Furthermore, the result continues to hold even when public-coin and interactive communication protocols are allowed.*

In light of this impossibility result, one can ask if distributed estimation is still possible by relaxing the requirement of finiteness in the worst-case for the number of players. We demonstrate that this is indeed the case and describe a protocol with finite expected number of players.[2]

**Theorem 1.3.** *For every $k, \ell \geq 1$, there exists a private-coin protocol with $\ell$ 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.*

We use this distributed simulation result to derive protocols for *any* distributed inference task:

**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 $\mathcal{P}$, with $\ell$ bits of communication per player, and $n = O(s \cdot k/2^{\ell})$ players.*

Instantiating this general statement for the prototypical distribution testing problem of uniformity testing leads to:

**Corollary 1.5.** *For every $k, \ell \geq 1$, there exists a private-coin protocol for testing uniformity over $[k]$, with $\ell$ bits of communication per player and $n = O\Big(\frac{k^{3/2}}{(2^{\ell} \wedge k)\varepsilon^2}\Big)$ players.*

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

The optimality of the simulate-and-infer strategy that generates $O(\sqrt{k})$ samples from the unknown $\mathbf{p}$ at the referee using private-coin protocols is open. Note that for a general inference problem even for $k$-ary observations the effective support-size can be much smaller. Thus, we can define the size of a problem as the least number of bits to which each samples can be compressed without increase in the number of compressed sample required to solve the problem (see Section 5.2 for a formal definition). An intriguing question ensues:

**Question 1.6** (The Flying Pony Question (Informal))**.** *Does the compressed simulate-and-infer scheme, which simulates independent samples compressed to the size of the problem using private-coin protocols and sends them to the referee who then infers from them, require the least number of players?*

For the problems considered in [DGL+17, HÖW18a], the answer to the question above is in the affirmative. However, we exhibit an example in Section 5.2 for which the answer is negative. Roughly, the problem we consider is that of testing if the distribution is uniform on $[k]$ or instead satisfies the following: for every $i \in [k]$, $\mathbf{p}_i$ is either $0$ or $2/k$. We show that the size of this problem remains $\log k$, whereby the simple simulate-and-infer scheme of the question above for $\ell = 1$ will require $O(k^{3/2})$ players. On the other hand, one can obtain a simple scheme to solve this task using $1$-bit communication from only $O(k)$ players. Interestingly, even this new scheme is of simulate-and-infer form, although it compresses below the size of the problem.

While the answer to the question above remains open for uniformity testing using private-coin protocols, it is natural to examine its scope and consider public-coin protocols for uniformity testing. As it turns out, here, too, the answer to the question is negative – public-coin protocols lead to an improvement in the required number of parties over the simple simulate-and-infer protocol described earlier by a factor of $\sqrt{k/2^\ell}$. Specifically, we provide a public-coin protocol for uniformity testing that requires roughly $O(k/2^{\ell/2})$ players and show that no public-coin protocol can work with fewer players.

**Theorem 1.7.** *For every $k, \ell \geq 1$, consider the problem of testing if the distribution is uniform or $\varepsilon$-far from uniform in total variation distance. There exists a public-coin protocol for uniformity testing with $\ell$ bits of communication per player and $n = O\left(\frac{k}{(2^{\ell/2} \wedge \sqrt{k})\varepsilon^2}\right)$ players. Moreover, this number is optimal up to constant factors.*

In fact, we provide two different protocols achieving this optimal guarantee. The first is remarkably simple to describe and requires $\Omega(\ell \cdot k)$ bits of shared randomness; the second is more randomness-efficient, requiring only $O(2^\ell \cdot \log k)$ bits of shared randomness,[3] but it is also more involved.

Before concluding this section, we emphasize that all our results for uniformity testing immediately imply the analogue for the more general question of identity testing, via a standard reduction argument. We detail this further in Section 6.

## 1.2 Proof techniques

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

[3] For our regime of interest, $\ell \ll \log k$, and so, $2^\ell \cdot \log k \ll \ell \cdot k$.

**Perfect and $\alpha$-simulation.** Our general impossibility result for perfect simulation with a finite number of players is based on simple heuristics. Observe that for any distribution $\mathbf{p}$ with $\mathbf{p}_i = 0$ for some $i$, the referee must not output $i$ for any sequence of received messages from the players. However, since $\ell < \log k$, by the pigeonhole principle one can find a sequence of messages $M = (M_1, ..., M_n)$ where each message $M_i$ has a positive probability of appearing from two different elements in $[k]$. Note that there exist distributions for which $M$ can occur with a positive probability, and not being able to abort with the symbol $\perp$, upon receiving this sequence the referee must output *some* element, say $i^*$. Then, for any distribution with $\mathbf{p}_{i^*} = 0$, this sequence $M$ must not be sent. But by construction each message in $M$ can be triggered by at least two elements in $[k]$. Thus, we can find a distribution with $\mathbf{p}_{i^*} = 0$ for which the sequence of messages $M$ will be sent with positive probability, which is a contradiction.

Next, we consider $\alpha$-simulation protocols, namely simulation protocols that are allowed to abort with probability less than $\alpha$. The proof of the positive result establishing the existence of $\alpha$-simulation proceeds by dividing the alphabet into $k/(2^\ell - 1)$ sets of size $2^\ell - 1$ and assigning each such set to two different players (each using their $\ell$ bits to indicate whether their sample fell in this subset, and if so on which element). If only one pair of players finds the sample in its assigned subset, the referee can declare this as the output, and it will have the desired probability. But it is possible that several pairs of players observe their assigned symbol and send conflicting messages. In this case, the referee cannot decide which of the elements to choose and must declare abort. However, we show that this happens with a probability that depends only on the $\ell_2$ norm of the unknown distribution $\mathbf{p}$; if we could assume this norm to be bounded away from 1, then our protocol would require $O(k)$ players. Unfortunately, this need not be the case. To circumvent this difficulty, we artificially duplicate every element of the domain and "split" each element $i \in [k]$ into two equiprobable elements $i_1, i_2 \in [2k]$. This has the effect of decreasing the $\ell_2$ norm of $\mathbf{p}$ by a factor $\sqrt{2}$, allowing us to instead apply our protocol to the resulting distribution $\mathbf{p}'$ on $[2k]$, for which the aforementioned probability of aborting can be bounded by a constant.

**Distributed uniformity testing.** To test whether an unknown distribution $\mathbf{p}$ is uniform using at most $\ell$ 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 $\varepsilon$-far from $\mathbf{u}_k$, then (with a constant probability over the choice of the random partition) $\mathbf{q}$ will remain $\varepsilon'$-far from $\mathbf{u}_L$, for some $\varepsilon'$ depending on $\varepsilon$, $L$, 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 $\ell_1$ distance) does not behave as desired under these random flattenings, and the validity of our de-

sired result remains unclear. Fortunately, an analogous statement with respect to the $\ell_2$ 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 $\ell_2$ distance to uniformity; in our case, by Cauchy–Schwarz the original $\ell_2$ distance will be at least $\gamma \asymp \varepsilon/\sqrt{k}$, which implies using known $\ell_2$ 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.

This leads us to our second protocol, whose main advantage is that it requires much fewer bits of randomness (specifically, $O_\varepsilon(2^\ell \log k)$); however, this comes at the price of some loss in simplicity. In fact, our second algorithm too pursues a natural, perhaps more greedy, approach: Pick uniformly at random a subset $S \subseteq [k]$ of size $s := 2^\ell - 1$ and communicate to the referee either an element in $S$ that equals the observed sample or indicate that the sample does not lie in $S$. If $\mathbf{p}$ is indeed uniform, then the probability $\mathbf{p}(S)$ of set $S$ satisfies $\mathbf{p}(S) = s/k$ and the conditional distribution $\mathbf{p}^S$ given that the sample lies in $S$ is uniform. On the other hand, it is not difficult to show that if $\mathbf{p}$ is $\varepsilon$-far from uniform in total variation distance, then the expected contribution of elements in $S$ to the $\ell_1$ distance of $\mathbf{p}$ to uniform is order $\varepsilon$. By an averaging argument, this implies that with probability at least $\varepsilon$ either (i) $\mathbf{p}(S)$ differs from $s/k$ by a $(1 \pm \Omega(\varepsilon))$ factor, or (ii) $\mathbf{p}_S$ is itself $\Omega(\varepsilon)$-far from uniform.

For a given $S$, detecting if (i) holds requires roughly $k/(s\varepsilon^2)$ samples (and hence as many players), while under (ii) one would need $(k/s) \cdot \sqrt{s}/\varepsilon^2 = k/(\sqrt{s}\varepsilon^2)$ players (the cost of rejection sampling, times that of uniformity testing on support size $s$) to test uniformity. When public randomness is available, the players can choose jointly the same random set $S$, so this protocol is valid. But there is a caveat. Since each choice of $S$ is only "good" with probability $\varepsilon$, to achieve a constant probability of success one needs to repeat the procedure outlined above for $\Omega(1/\varepsilon)$ different choices of $S$. This, along with the overhead cost of a union bound over all repetitions, leads to a bound of $O(k/(2^{\ell/2}\varepsilon^3) \cdot \log(1/\varepsilon))$ on the number of players – far from the optimal answer of $n = O(k/(2^{\ell/2}\varepsilon^2))$. To avoid the extra $1/\varepsilon$, we rely instead on *Levin's work investment strategy* (see e.g. [Gol14, Appendix A.2]), which by a more careful accounting enables us to avoid paying the cost of the naive averaging argument. Instead, by considering logarithmically many different possible "scales" $\varepsilon_j$ of distance between $\mathbf{p}_S$ and uniform, each with its own probability $\alpha_j$ of occurring for a random choice of $S$ and by keeping track of the various costs that ensue, we can get rid of this extra $1/\varepsilon$ factor. This only leaves us with an extra $\log(1/\varepsilon)$ factor to handle, which arises due to the union bound. To omit this extra factor, we refine our argument by allocating different failure probabilities to every different test conducted, depending on the respective scale used. By choosing these probabilities so that their sum is bounded by a constant (for instance, by setting $\delta_i \propto 1/j^2$), we can still ensure overall correctness with a high, constant probability, while the extra cost $\log(1/\delta_j)$ for the $j$-th scale considered is subsumed in the accounting using Levin's strategy. This finally yields the desired bound of $n = O(k/(2^{\ell/2}\varepsilon^2))$ for the number of players.

For the lower bound, we take recourse to Le Cam's two-point method. Specifically, we use the construction proposed by Paninski [Pan08] for proving the lower bound for sample complexity in the collocated setting. Roughly, we consider the problem of distinguishing the uniform distribution from a randomly selected element of the family of distributions consisting each of a

perturbation of uniform distribution where the probabilities conditioned on pairs of consecutive elements are changed from unbiased coins to coins of bias $\varepsilon$. However, Paninski's original treatment does not suffice now as we need to handle the total variation distance between the distribution induced on the message sequence $M$ under the uniform distribution and a uniform mixture of the pairwise perturbed distributions. This is further bounded above by the average distance between the message distribution under uniform input and under the pairwise perturbed input. In fact, treating public randomness as a common observation in both settings, it suffices to obtain a worst-case bound for deterministic inference protocols. Capitalizing on the fact that both distributions of messages are in this case product distributions, we can show that this average distance for deterministic protocols is bounded above by $\sqrt{n(2^\ell\varepsilon^2)/k}$, which leads to a lower bound of $n = \Omega(k/(2^\ell\varepsilon^2))$.

The bound obtained above is tight for $\ell = 1$, but is sub-optimal in general . To refine this bound further, instead of considering the average distance between the distributions, we need to carefully analyze the distance of uniform from the average. However, this quantity is not amenable to standard bounds for total variation distance in terms of Kullback–Leibler divergence and Hellinger distance devised to handle product distributions, as the average "no-distribution" is not itself a product distribution anymore. Instead, we take recourse to a technique used in [Pan08], building on [Pol03], that uses a $\chi^2$-distance bound and proceeds by expanding the product likelihood ratios in multilinear form. Obtaining the final bound requires a sub-Gaussian bound for a log-moment generating function, which is completed by using a standard transportation method technique.

### 1.3 Related prior work

For clarity, we divide our discussion of the relevant literature into three parts: The first discussing the literature in the collocated setting, and the next two the prior work concerned with distributed inference and simulation.

**Inference in the collocated setting.** The goodness-of-fit problem is a classic hypothesis testing problem with a long line of work in statistics, but the finite-alphabet variants of interest to us were first considered by Batu et al. [BFR+00] and Goldreich, Goldwasser, and Ron [GGR98] under distribution testing, which in turn evolved as a a branch of property testing [RS96, GGR98], a field of theoretical computer science focusing on "ultra-fast" (sublinear-time) algorithms for decision problems. Distribution testing has received much attention in the past decade, with considerable progress made and tight answers obtained for many distribution properties (see e.g. surveys [Rub12, Can15, BW18] and references within for an overview). Most pertinent to our work is uniformity testing [GR00, Pan08, DGPP17], the prototypical distribution testing problem with applications to many other property testing problems [BKR04, DKN15, Gol16, CDGR17].

Another inference question in the finite-alphabet setting that has received a lot of attention in recent years is that of functional estimation, where the goal is to estimate a function of the underlying distribution. Recent advances in this area have pinpointed the optimal rates for functionals such as entropy, support size, and many others (see for instance [Pan04, RRSS09, VV11, JVYW15, WY16, AOST17, JVHW17, ADOS17] for some of the most recent work).

The extreme case of functional estimation is the fundamental question of distribution learning, namely the classic density estimation problem in statistic where the goal is estimate the

entire distribution. With more than a century of history (see the books [Tsy09, DL01]), distribution learning has recently seen a surge of interest in the computer science community as well, with a focus on discrete domains (see e.g. [Dia16] for a survey of these recent developments).

**Inference in the distributed setting.** As previously mentioned, distributed hypothesis testing and estimation problems were first studied in information theory, albeit in a different setting than what we consider [AC86, Han87, HA98]. The focus in that line of work has been to characterize the trade-off between asymptotic error exponent and communication rate per sample. Recent extensions have considered interactive communication [XK13], more complicated communication models [WT16], and even no communication [Wat17]. The communication complexity of independence testing for fixed error has been considered recently in [ST18].

Closer to our work is distributed parameter estimation and functional estimation that has gained significant attention in recent years (see e.g. [DJW13, GMN14, BGM+16, Wat18]). 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.

Our communication model is the same as that considered in [HÖW18a], which establishes, under some mild assumptions, a general lower bound for estimation of model parameters under squared $\ell_2$ loss. Although the problems considered in our work differ from those in [HÖW18a] and the results are largely incomparable, we build on a result of theirs to establish one of our lower bounds.[4]

The problem of distributed density estimation, too, has gathered recent interest in various statistical settings [BPC+11, BBFM12, ZDJW13, Sha14, DGL+17, HÖW18a, XR17, ASZ18]. Our work is closest to two of these: The aforementioned [HÖW18a, HMÖW18] and [DGL+17]. The latter considers both $\ell_1$ (total variation) and $\ell_2$ 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 SMPs. 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 [BCG17] 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.)

[4] In fact, the same communication model was proposed in a different work, presented at the 2018 ITA workshop [HMÖW18]. In this talk, the authors described a protocol for learning discrete distributions under $\ell_1$ error, with a number of players that they showed to be optimal up to constant factors.

**Distributed simulation.** 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 [GK73] and Wyner [Wyn75] 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 [Bro97, KT02, Hol07, BGH+16] and non-interactive simulation [KA12, GKS16, DMN18]. 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. Section 4 introduces the question of distributed simulation and contains our protocols and impossibility results for this problem (specifically, Theorem 1.2 and Theorem 1.3 are proven in Section 4.1 and in Section 4.2). In Section 5, we consider the relation between distributed simulation and (private-coin) distribution inference. Namely, we explain in Section 5.1 how a distributed simulation protocol immediately implies protocols for every inference task (Theorem 1.4) and instantiate this result for two concrete examples of distribution learning and uniformity testing. Section 5.2 is concerned with Question 1.6: "Is inference via distributed simulation optimal in general?" After rigorously formalizing this question, we answer it in the negative in Theorem 5.10.

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 (as previously mentioned, we provide there two proofs of the upper bound using different protocols, with a simple, albeit randomness-heavy, protocol, and a more involved, randomness-savvy one).

Although we rely throughout on the formal description of our model given in Section 3, the other sections are self-contained and can be read independently.

## 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 $\varepsilon$*-close* to $\mathcal{P}$ if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}) \leq \varepsilon$; otherwise, we say that $\mathbf{p}$ is $\varepsilon$*-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 $\chi^2$ 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 number $a$ 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.

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 $\ell$-bit *simultaneous message-passing* (SMP) communication protocol $\pi$ 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. *Pairwise-coin protocols:* $U_1, \ldots, U_n$ are mutually independent and available to the referee.
3. *Public-coin protocols:* All player and the referee have access to $U_1, \ldots, U_n$.

In this paper, we focus only on private- and public-coin communication protocols; an interesting question is distinguishing pairwise-coin protocols from the other two. 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 $U_1 = U_2 = \cdots = U_n = U$.

**Distributed simulation protocols.** An $\ell$-bit *simulation* $\mathcal{S} = (\pi, \delta)$ of $k$-ary distributions using $n$ players consists of an $\ell$-bit SMP $\pi$ and a decision map $\delta$ 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 $\pi$ is a private, pairwise, or public-coin protocol, respectively, the simulation $\mathcal{S}$ is called private, pairwise, or public-coin simulation.

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

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

and the abort probability satisfies

$$\Pr_{\mathbf{p}}\left[\hat{X} = \perp\right] \leq \alpha.$$

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 $\ell$-bit *distributed inference protocol* $\mathcal{I} = (\pi, e)$ for the inference problem $(\mathcal{C}, \mathcal{X}, \mathcal{E}, L)$ consists of an $\ell$-bit SMP $\pi$ 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, pairwise, or public-coin inference protocol, respectively, uses a private, pairwise, or public-coin communication protocol $\pi$.

For $\vec{\gamma} \in \mathbb{R}_+^q$, an inference protocol $(\pi, e)$ 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 examples 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 $\delta$-inference protocol an $\ell$-bit $(k, \varepsilon, \delta)$-*learning protocol* for $n$ player.

*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 $[k]$ or if it satisfies $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$. This can be obtained as a special case of our general formulation by setting $\mathcal{X} = [k]$, $\mathcal{C}$ to be 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 loss function $L$ to be

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

For this case, we term the $\delta$-inference protocol an $\ell$-bit $(k, \varepsilon, \delta)$-*uniformity testing protocol* for $n$ players. Further, for simplicity we will refer to $(k, \varepsilon, 1/3)$-uniformity testing protocols simply as $(k, \varepsilon)$-*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* [KN97]. 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. We start by considering the more ambitious problem of perfect simulation and show that when $\ell < \log k$, perfect simulation using $n$ players is impossible using any $n$. Next, we consider $\alpha$-simulation for a constant $\alpha \in (0,1)$ and exhibit an $\ell$-bit $\alpha$-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 Section 5.1 that this is the least number of players required (up to a constant factor) for $\alpha$-simulation for any $\alpha \in (0,1)$.

## 4.1 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:

**Theorem 4.1.** *For any $n \geq 1$, there exists no $\ell$-bit perfect simulation of $k$-ary distributions using $n$ players unless $\ell \geq \log k$.*

*Proof.* Let $\mathcal{S} = (\pi, \delta)$ be an $\ell$-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 $x_i \neq x$ 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., $\Pr\left[\hat{X} = x\right] = 0$. In particular, $\Pr\left[\hat{X} = x \,\middle|\, U = u\right]$ must be $0$, which can only happen if $\Pr[M = m \mid U = u] = 0$. But since $\mathbf{p}_{x_i} > 0$ for each $i$,

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

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

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 4.3.** *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 $\pi$ 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 the 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 $u$

$$\Pr[\, f(X,u) = 1\,] = \Pr[\, h(g_1(X_1), g_2(X_2), u) = 1\,]\,, \tag{2}$$

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 $\pi_{2j-1}$ and $\pi_{2j}$ 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. 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 $\alpha_u \leq \beta_u + \gamma_u$; 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

$$h(0,0,u) = \alpha_u,\ \ h(1,0,u) = \beta_u,\ \ h(0,1,u) = \gamma_u,\ \ h(1,1,u) = (\beta_u + \gamma_u - \alpha_u)\,.$$

Then, for every $u$,

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

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

$$\begin{aligned}\mathbf{p}_x &= \mathbb{E}_U \sum_{m\in\{0,1\}^n} \delta_x(m,U)\Pr[\,M=m \mid U\,] = \mathbb{E}_U \sum_m \delta_x(m,U)\Pr[\,M=m\,] \\ &= \sum_m \mathbb{E}_U[\delta_x(m,U)]\Pr[\,M=m\,] \qquad (3)\end{aligned}$$

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:

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

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

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

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

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

contradicting the fact that $\pi$ is a perfect simulation protocol. □

*Remark* 4.4. It is unclear how to extend the proof of Theorem 4.2 arbitrary $k, \ell$. 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 4.5.** *For every $n \geq 1$, if there exists an interactive public-coin $\ell$-bit perfect simulation of $k$-ary distributions with $n$ players, then there exists a public-coin $\ell$-bit perfect simulation of $k$-ary distributions with $2^{\ell n+1}$ players that uses only SMP.*

*Proof.* Consider an interactive communication protocol $\pi$ for distributed simulation with $n$ players and $\ell$ bits of communication per player. We can view the overall protocol as a $(2^\ell)$-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 4.6.** *Theorems 4.1 and 4.2 extend to interactive protocols as well.*

## 4.2 An $\alpha$-simulation protocol using rejection sampling

In this section, we establish Theorem 1.3 and provide $\alpha$-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 $\ell$. The proof of lower bound for the number of players required for $\alpha$-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 $\alpha$-simulation.

**Theorem 4.7.** *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.* 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

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

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.

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

$$\frac{2k}{\rho_{\mathbf{p}}} \leq 2k \cdot \frac{e^{1 - \|\mathbf{p}\|_2}}{1 - \|\mathbf{p}\|_2} \,. \tag{4}$$

**The scheme, final version.** The protocol described above can have the expected number of players blowing to infinity when $\mathbf{p}$ has $\ell_2$ 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, $\|\mathbf{q}\|_2 = \|\mathbf{p}\|_2/\sqrt{2} \leq 1/\sqrt{2}$, and therefore by (4) the expected number of players required to simulate $\mathbf{q}$ using our previous protocol is at most

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

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. □

Moving now to the more general setting, we have the following result.

**Theorem 4.8.** *For any $\ell \geq 2$, there exists a $\ell$-bit private-coin protocol that outputs a sample $x \sim \mathbf{p}$ using messages of at most $20\left\lceil \frac{k}{2^\ell - 1} \right\rceil$ players in expectation.*

*Proof.* 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 $\ell$ 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

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

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

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

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 $\|\mathbf{q}^{(S)}\|_2 \leq 1/\sqrt{2}$ and accesses messages from at most $20m$ players in expectation. □

# 5 Distributed Simulation for Distributed Inference

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?"

## 5.1 Private-coin distributed inference via distributed simulation

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 $(k/2^\ell)$, 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 $\ell$ 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

1. Partition the players into blocks of size $54k/2^\ell$.
2. Run the distributed simulation protocol (Theorem 4.8) on each block.
3. Run the centralized algorithm over the simulated samples.

From Theorem 4.8, 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 $B/2-\sqrt{B}\sqrt{15/4} > \psi(\mathcal{P},k)$. Since $1/3+1/15 = 2/5$, 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 $\Theta(k/\varepsilon^2)$ samples are sufficient to learn a distribution over $[k]$ to within a total variation distance $\varepsilon$ with probability 2/3, we obtain:

**Corollary 5.2.** *Let $\ell \in \{1, \ldots, \log k\}$. Then, there exists an $\ell$-bit private-coin $(k, \varepsilon, 3/5)$-learning protocol for $O\left(\frac{k^2}{2^\ell \varepsilon^2}\right)$ players.*

From the existence of uniformity testing algorithms using $O(\sqrt{k}/\varepsilon^2)$ samples [Pan08, VV17, DGPP17], we obtain:

**Corollary 5.3.** *Let $\ell \in \{1, \ldots, \log k\}$. Then, there exists an $\ell$-bit private-coin $(k, \varepsilon, 3/5)$-uniformity testing protocol for $O\left(\frac{k^{3/2}}{2^\ell \varepsilon^2}\right)$ players.*

Interestingly, a byproduct of this "simulate-and-infer" connection (and, more precisely, of Corollary 5.2) is that the $\alpha$-simulation protocol from Theorem 4.8 has optimal number of players, up to constants.

**Corollary 5.4.** *Let $\ell \in \{1, \ldots, \log k\}$, and $\alpha \in (0, 1)$. Then, any $\ell$-bit public-coin (possibly adaptive) $\alpha$-simulation protocol for $k$-ary distributions must have $n = \Omega(k/2^\ell)$ players.*

*Proof.* Let $\pi$ be any $\ell$-bit $\alpha$-simulation protocol with $n$ players; by Theorem 5.1, and analogously to Corollary 5.2 we have that $\pi$ implies an $\ell$-bit $(k, \varepsilon, 1/3)$-learning protocol for $n' = O(n \cdot k/\varepsilon^2)$ players.[5] (Moreover, the resulting protocol is adaptive, private-, pairwise-, or public-coin, respectively, whenever $\pi$ is.) However, as shown in Appendix B (Theorem B.1), any $\ell$-bit public-coin (possibly adaptive) $(k, \varepsilon, 1/3)$-learning protocol must have $\Omega\left(k^2/(2^\ell \varepsilon^2)\right)$ players. It follows that $n$ must satisfy $n \gtrsim k/2^\ell$, as claimed. □

*Remark* 5.5. We note that the learning upper bound of Corollary 5.2 appears to be established in [HMÖW18] as well (with however, to the best of our knowledge, completely different techniques). The authors of [HÖW18a] also describe a distributed protocol for distribution learning, but their criterion is the $\ell_2$ distance instead of total variation.[6] Finally, our learning lower bound (Appendix B), invoked in the proof of Corollary 5.4 above, is established by adapting a similar lower bound from [HÖW18a] which again applies to learning in the $\ell_2$ metric.

## 5.2 Is distributed simulation essential for distributed inference?

In the previous subsection, we saw that it is easy to derive distributed learning and testing protocols from distributed sampling. However, the optimality of simulate-and-infer for uniformity testing using private-coin protocols is unclear. In fact, a natural question arises: *Is the simulate-and-infer approach always optimal?* Note that such an optimality would have appealing implementation consequences, where one need not worry about the target inference application when designing the communication protocol – the communication protocol will simply enable

[5] Improving the probability of success from $3/5$ to $1/3$ can be achieved by standard arguments, with at most a constant factor blowup in the number of players.

[6] We note that, based on a preliminary version of our manuscript on arXiv, the $\ell_2$ learning upper bound of [HÖW18a] was updated to use a "simulate-and-infer" protocol as well.

distributed simulation and the referee can implement the specific inference algorithm needed. A similar result, known as *Shannon's source-channel separation theorem*, has for instance allowed for development of compression algorithms separately from the error-correcting codes for noisy channels. Unfortunately, optimality of simulate-and-infer can be refuted by the following simple example:

**Observation 5.6.** In the distributed setting model ($n$ players, and $\ell = 1$ bit of communication per player to the referee), testing whether a distribution over $[k]$, promised to be monotone, is uniform vs. $\varepsilon$-far from uniform can be done with $n = O(1/\varepsilon^2)$ (moreover, this is optimal).

*Sketch.* Optimality is trivial, since that many samples are required in the non-distributed setting. To see why this is enough, recall that a monotone distribution $\mathbf{p} \in \Delta([k])$ is $\varepsilon$-far from uniform if, and only if, $\mathbf{p}(\{1,\dots,k/2\}) > \mathbf{p}(\{k/2+1,\dots,k\}) + 2\varepsilon$. Therefore, we only need $n = O(1/\varepsilon^2)$ players, where each player sends 1 if their sample was in $\{1,\dots,k/2\}$ and 0 otherwise. □

However, this counter-example is perhaps not satisfactory since the inference problem itself was compressible since the dimension of the parameter space was increased artificially.[7] In fact, it is natural to consider an extension of simulate-and-infer where we first compress the observation to capture the effective dimension of the underlying parameter space. To formally define such *compressed simulate-and-infer* schemes, we must first define a counterpart of sufficient statistic that will be relevant here.

Let $\mathcal{P}$ denote an inference problem (for instance, the distribution learning and the uniformity testing problem of the previous section) and $n(\mathcal{P})$ denote the minimum number of independent samples required to solve it, namely its sample complexity. Note that the description of the problem $\mathcal{P}$ includes the observation alphabet $\mathcal{X}$, the loss function used to evaluate the performance, and the required performance. For a (fixed, deterministic) mapping $f$ on $\mathcal{X}$, denote by $\mathcal{P}_f$ the problem where we replace each observed sample $X$ with $f(X)$.

**Definition 5.7** (The size of a problem)**.** A problem $\mathcal{P}$ is said to be *compressible to $\ell$ bits* if there exists a mapping $f\colon \mathcal{X} \to \{0,1\}^\ell$ such that $n(\mathcal{P}_f) = n(\mathcal{P})$. For such a function $f$ with $\ell \le \log|\mathcal{X}|$, we call $f(X)$ a *compressed statistic*.

The *size* $|\mathcal{P}|$ of a problem $\mathcal{P}$ is then defined as the least $\ell$ such that $\mathcal{P}$ is compressible to $\ell$. If a mapping $f$ attains $|\mathcal{P}|$, we call $f$ a *maximally compressed statistic* for $\mathcal{P}$.

*Example* 5.8. For the uniformity testing problem $\mathcal{T}^{\mathrm{u}}(k,\varepsilon)$ considered in the previous section, we must have

$$|\mathcal{T}^{\mathrm{u}}(k,\varepsilon)| \ge \log k - \log\frac{1}{1-\varepsilon}.$$

The proof follows by noting that for each mapping $f$ with range cardinality $k(1-\varepsilon)$, we can find a distribution $Q$ on $[k]$ that is $\varepsilon$-far from uniform, yet $f(X)$ is uniform under $Q$.

A compressed simulate-and-infer scheme then proceeds by replacing the original observation $X_j$ by its maximally compressed sufficient statistic $f(X_j)$ at player $j$ and then applying simulate-and-infer for $f(X)$. Note that this new scheme, too, has the appealing feature that we can use our distributed simulation protocol as a black-box communication step to enable distributed inference. But are such compressed simulate-and-infer schemes optimal? Formally,

[7] That is, the inference task was only "superficially" parameterized by $k$, but was actually a task on $\{0,1\}$ and entails only estimating the bias of a coin in disguise.

**Question 5.9** (The Flying Pony Question)**.** *To solve $\mathcal{P}$ in the distributed setting, must the number of parties $n$ satisfy $n = \Omega\left(2^{|\mathcal{P}|-\ell} \cdot n(\mathcal{P})\right)$?*

This essentially asks whether the most economical communication scheme to solve $\mathcal{P}$ is indeed to simulate $n(\mathcal{P})$ samples from a maximally compressed statistic. Observe that we noted the optimality of such a scheme for distribution learning, even when public-coin protocols are allowed. Further, to the best of our knowledge, previous results on this topic, in essence, establish lower bounds to show the optimality of such simple schemes (cf. [BGM+16, DGL+17, HÖW18a].

In spite of this evidence, we are able to refute this conjecture.[8] Specifically, we exhibit an inference task $\mathcal{P}$ over $k$-ary distributions which admits a 1-bit private-coin protocol with $n = o(2^{|\mathcal{P}|}n(\mathcal{P}))$ players.

**Theorem 5.10.** *There is an inference task $\mathcal{P}$ over $k$-ary distributions with $2^{|\mathcal{P}|} \cdot n(\mathcal{P}) = \Omega(k^{3/2})$, yet for which there exists a 1-bit private-coin protocol with $n = O(k)$ players.*

*Proof.* We start by describing the inference task in question. For every even $k \geq 2$, $\mathcal{P}$ consists in distinguishing between the following two cases: either $\mathbf{p} = \mathbf{u}_k$, the uniform distribution over $[k]$; or $\mathbf{p}$ is any of the $2^{k/2}$ possible uniform distributions over a subset of size $k/2$ defined as follows. For a parameter $\theta \in \{-1,1\}^{k/2}$, $\mathbf{p}_\theta$ is the distribution such that, for every $i \in [k/2]$,

$$\mathbf{p}_\theta(2i-1) = \frac{1+\theta_i}{k}, \qquad \mathbf{p}_\theta(2i) = \frac{1-\theta_i}{k}$$

and in particular $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_\theta, \mathbf{u}_k) = 1/2$ for every $\theta \in \{-1,1\}^{k/2}$.

By an easy birthday-paradox type argument, we have that $n(\mathcal{P}) = \Omega(\sqrt{k})$ (and this is tight), so to prove the first part of the statement it is enough to show that $|\mathcal{P}| = \Omega(k)$. To see why this is the case, set $L := |\mathcal{P}|$, and consider any maximally compressed statistic $f\colon [k] \to \{0,1\}^L$ for $\mathcal{P}$. This $f$ immediately implies a (private-coin) $L$-bit $(k,1/2)$-uniformity testing protocol: namely, a protocol where each player first applies $f$ to their sample, then sends the resulting $L$ bits to the referee. Further, by definition of a maximally compressed statistic, we have $n(\mathcal{P}_f) = n(\mathcal{P}) = \Theta(\sqrt{k})$; as in the aforementioned $L$-bit protocol the referee only needs $n(\mathcal{P}_f)$ samples from the distribution on $2^L$, this therefore gives an $O(\sqrt{k} \cdot 2^{L-L}) = O(\sqrt{k})$ upper bound on the number of players required.

However, peeking ahead, Theorem 1.7 shows that any $L$-bit protocol for $\mathcal{P}$ (even allowing for public coins) must have $\Omega(k/2^{L/2})$ players.[9] Combining this lower bound with the $O(\sqrt{k})$ upper bound we have just established yields

$$\frac{k}{2^{L/2}} \lesssim \sqrt{k}\,, \tag{5}$$

i.e., $k \lesssim 2^L$. This, along with the lower bound on $n(\mathcal{P})$, implies that $2^{|\mathcal{P}|} \cdot n(\mathcal{P}) = \Omega(k \cdot \sqrt{k}) = \Omega(k^{3/2})$, as claimed.

To obtain a contradiction, it remains to prove the second part of the statement, i.e., to describe a 1-bit private-coin protocol with $n = O(k)$ players. Consider the protocol where every of

[8] Thus implying that, even if wishes *were* horses, there would be no flying ponies.

[9] Indeed, this is because the inference task $\mathcal{P}$ described here is a specific case of the lower bound construction underlying the proof of Theorem 1.7, obtained by taking $\varepsilon = 1/2$.

the $n$ players simply sends $1$ if their sample is equal to $1$, and $0$ otherwise. If $\mathbf{p} = \mathbf{u}_k$, then each bit is independently $1$ with probability $1/k$. However, if $\mathbf{p}$ is one of the distributions uniform over $k/2$ elements, then $\mathbf{p}_1 \in \{0, 2/k\}$, and therefore either each player's bit is independently $1$ with probability $0$, or each player's bit is independently $1$ with probability $2/k$. In either case, the problem then amounts to distinguish a coin with bias $1/k$ to one with bias either $0$ or $2/k$; for which $n = O(k)$ players suffice, concluding the proof. □

While we have refuted the optimality compressed simulate-and-infer, the strategy used in the counter-example above still entails simulating samples from a fixed distribution at the referee. This statistic, while compressed form of the original problem, is not a compressed sufficient statistic as it mandates a higher number of samples in the centralized setting. We call such inference protocols that entail simulate-and-infer for some compressed statistic of the problem[10] *generalized simulate-and-infer*; the optimality of generalized simulate-and-infer is unclear, in general. For our foregoing example of uniformity testing, it is not even clear whether there is a private-coin protocol that requires fewer players than the vanilla simulate-and-infer scheme. Interestingly, we can provide a public-coin protocol that outperforms simulate-and-infer for uniformity testing and show that it is optimal. This is the content of the next section.

# 6 Public-Coin Uniformity Testing

In this section, we consider public-coin protocols for $(k, \varepsilon)$-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 $\ell$-bit public-coin $(k, \varepsilon)$-uniformity testing protocol for $n = O\left(\frac{k}{2^{\ell/2}\varepsilon^2}\right)$ players.*

Note that this is much fewer than the $O(k^{3/2}/(2^\ell \varepsilon^2))$ players required using private-coin protocols in Corollary 5.3. In fact, this is optimal, being the least number of players (up to constant factors) needed for any public-coin protocol:

**Theorem 6.2.** *For $1 \le \ell \le \log k$, any $\ell$-bit public-coin $(k, \varepsilon)$-uniformity testing protocol must have $n = \Omega\left(\frac{k}{2^{\ell/2}\varepsilon^2}\right)$ players.*

We establish Theorem 6.1 and Theorem 6.2 in Sections 6.1 and 6.2, respectively. Before delving into the proofs, 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 $\varepsilon$-far from) a reference distribution $\mathbf{q}$ known to all the players.

**Corollary 6.3.** *For $1 \le \ell \le \log k$, and for any fixed $\mathbf{q} \in \Delta([k])$, there exists an $\ell$-bit public-coin $(k, \varepsilon, \mathbf{q})$-identity testing protocol for $n = O\left(\frac{k}{2^{\ell/2}\varepsilon^2}\right)$ players. Further, any $\ell$-bit public-coin $(k, \varepsilon, \mathbf{q})$-identity testing protocol must have $\Omega\left(\frac{k}{2^{\ell/2}\varepsilon^2}\right)$ 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 Valiant and Valiant [VV17].

[10] This need not be a compressed sufficient statistic.

## 6.1 Upper bound: public-coin protocols

This section is dedicated to the proof of Theorem 6.1. We actually provide and analyze two different protocols achieving the stated upper bound: the first, in Section 6.1.1, is remarkably simple, and, moreover, is "smooth" – that is, no player's output depends too much on any particular symbol from $[k]$. However, this first protocol has the inconvenience of requiring a significant amount of public randomness, $\Theta(k \cdot \ell) = \Omega(k)$ bits.

To address this, we provide in Section 6.1.2 a different protocol requiring the optimal number of players, too, but necessitating much less randomness, only $\Theta_\varepsilon(2^\ell \log k) = O_{\varepsilon,\ell}(\log k)$ bits.[11] On the other hand, this second protocol is slightly more complex and highly "non-smooth" (specifically, the output of each player entirely depends on only $\ell$ symbols).

### 6.1.1 A simple "smooth" protocol

The protocol will rely on a generalization of the following observation: *if $\mathbf{p}$ is $\varepsilon$-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 $\varepsilon$-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.[12]

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 [ACFT18, Corollary 15] 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 $k$. Further, with $Y_1, \dots, Y_k$ independent and uniform random variables on $[L]$, let random variables $X_1, \dots, X_k$ have the same distribution as $Y_1, \dots, Y_k$ 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 $X_i$s are not independent. For $\mathbf{p} \in \Delta([k])$, define random variables $Z_1, \dots, Z_L$ as follows:

$$Z_r := \sum_{i=1}^{k} \mathbf{p}_i \mathbb{1}_{\{X_i = r\}} \,. \tag{6}$$

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

[11] Note that $2^\ell \log k \le k\ell$ for every $1 \le \ell \le k$.

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

**Theorem 6.4.** *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*

$$\Pr\left[\|\mathbf{q} - \mathbf{u}_L\|_2^2 > \frac{\varepsilon^2}{k}\right] \geq c\,.$$

*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 $\ell_2$ distance at least $\varepsilon'$ from uniform, where $\varepsilon' := \sqrt{\varepsilon^2/k}$. However, testing uniformity vs. $(\gamma/\sqrt{L})$-farness from uniformity in $\ell_2$ distance, over $[L]$, has sample complexity $O(\sqrt{L}/\gamma^2)$ (see e.g. [CDVV14, Proposition 3.1] or [CDGR17, Theorem 2.10]), and for our choice of $\gamma := \sqrt{L}\varepsilon' \in (0, 1)$, we have

$$\frac{\sqrt{L}}{\gamma^2} = \frac{\sqrt{L}}{L\varepsilon'^2} = \frac{k}{\sqrt{L}\varepsilon^2} = \frac{k}{2^{\ell/2}\varepsilon^2}, \tag{7}$$

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

**Corollary 6.5.** *For* $1 \leq \ell \leq \log k$*, there exists an* $\ell$*-bit public-coin* $(k, \varepsilon)$*-uniformity testing protocol for* $n = O\left(\frac{k}{2^{\ell/2}\varepsilon^2}\right)$ *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 $\ell_2$ distance from the uniform distribution over $[L]$, with failure probability $\delta$ (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 $\Pi$ of $[k]$ into subsets of size $k/2^\ell$.
- Next, they send to the referee the $\ell$ bits indicating which part of the partition their observed sample fell in.

The referee, receiving these $N$ messages (which correspond to $N$ independent samples of the distribution $\mathbf{q} \in \Delta\left([2^\ell]\right)$ induced by $\mathbf{p}$ on $\Pi$) runs the $\ell_2$ uniformity test, with failure probability $\delta$ 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 $\varepsilon$-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 $\ell_2$ 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. □

Note that the protocol underlying Corollary 6.5 is "smooth," in the sense that each player's output is the indicator of a set of $k/2^\ell$ elements, which for constant values of $\ell$ is $\Omega(k)$.

### 6.1.2 A randomness-efficient optimal protocol

We now provide our second optimal public-coin protocol, which albeit less simple than that of the previous section is much more randomness-efficient. We start with the case $\ell = 1$ (addressed in Proposition 6.8 below), before generalizing to an arbitrary $\ell \geq 1$ – the generalization is nontrivial and uses a more involved protocol. Before we present our actual scheme, to help the reader build heuristics, we present a simple, albeit non-optimal, scheme.

**Proposition 6.6** (Warmup)**.** *There exists a* $1$*-bit public-coin* $(k,\varepsilon)$*-uniformity testing protocol for* $n = O(k\log(1/\varepsilon)/\varepsilon^3)$*, which uses* $O((\log k)/\varepsilon)$ *bits of randomness.*

*Proof.* The starting point of the protocol is the straightforward observation that if $\mathbf{p}$ is $\varepsilon$-far from uniform then at least an $\Omega(\varepsilon)$ fraction of the domain must have an $\Omega(\varepsilon)/k$ deviation from uniform. Indeed, consider a $\mathbf{p} \in \Delta([k])$ such that $\mathrm{d}_{\mathrm{TV}}\left(\mathbf{p}, \mathbf{u}_{[k]}\right) \geq \varepsilon$. By contradiction, suppose that there are only $k' < \frac{\varepsilon}{2}\cdot k$ elements such that $\mathbf{p}_i < (1-\frac{\varepsilon}{2})\cdot\frac{1}{k}$. Then,

$$\mathrm{d}_{\mathrm{TV}}\left(\mathbf{p}, \mathbf{u}_{[k]}\right) = \sum_{i:\mathbf{p}_i<1/k}\left(\frac{1}{k}-\mathbf{p}_i\right) \leq k'\cdot\frac{1}{k} + (k-k')\cdot\frac{\varepsilon}{2k} < \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon$$

contradicting the assumption that $\mathbf{p}$ was $\varepsilon$-far from uniform. Therefore,

$$|\{i\in[k] : \mathbf{p}_i < (1-\frac{\varepsilon}{2})\cdot\frac{1}{k}\}| \geq \frac{\varepsilon}{2}\cdot k. \tag{8}$$

Next, we recall the well-known fact that a coin with bias $1/k$ can be distinguished from another with bias $(1-\varepsilon/2)/k$ with probability[13] $1-\delta$ using $ck\log(1/\delta)/\varepsilon^2$ independent coin tosses for some constant $c$. Therefore, any $i\in[k]$ with $\mathbf{p}_i < (1-\varepsilon/2)/k$ were known to the players, then testing if its probability is $1/k$ or $(1-\varepsilon/2)/k$ will require $ck\log 1/\delta)/\varepsilon^2$ players simply by using $\mathbb{1}_{\{X_j=i\}}$ as communication for player $j$.

We use this observation to build our protocol. Specifically, we divide the players into $m$ disjoint batches of size $n' = ck\log(1/\delta)/\varepsilon^2$ players; we will specify $m$ and $\delta$ later. We assign a random element $i$ to each batch, generated uniformly from $[k]$ using public randomness. Then, the parties in the batch apply the aforementioned test to distinguish if the probability of the selected $i$ is $1/k$ or $(1-\varepsilon/2)/k$. We accept $\mathbf{u}$ if all the batches accepted $1/k$ as the probability of their respectively assigned $i$s; else we reject $\mathbf{u}$. Note that since each batch's selected $i$ lies in the desired set in (8) with probability at least $\varepsilon/2$, with probability greater than $9/10$ at least one batch will be assigned an $i$ in the desired set if the number of batches satisfies

$$m \geq \frac{5}{\varepsilon}.$$

When the underlying distribution is uniform, the protocol will make an error only if one of the test for one of the batches fails, which can happen with probability less than $m\delta \leq 1/10$ if

$$\delta \leq 1/(10m).$$

[13] That is, denoting the two distributions by $P$ and $Q$, we can find a subset $A$ of sequences in $\{0,1\}^n$ such that $P^n(A)\geq 1-\delta$ and $Q^n(A)\leq\delta$.

On the other hand, if the underlying distribution is $\varepsilon$-far from uniform, then the test will fail only if either no $i$ in the desired set was selected or if the protocol failed for an $i$ in the desired set; the former happens with probability less than $1/10$ and the latter with probability less than $1/10m$ when we choose $\delta = 1/10m$. Thus, the overall probability of error in this case is less than $2/10$, whereby we have an $(k,\varepsilon)$-uniformity testing protocol using $mn' = 5ck\log(50/\varepsilon)/\varepsilon^3$ players. Moreover, the total number of random bits required is $O(m\log k)$, as claimed. □

**Improving on the warmup protocol using Levin's work investment strategy.** The main issue with the proof of Proposition 6.6 is the use of a "reverse" Markov style argument to identify the $i$ to focus on. This approach is inherently wasteful, as can be seen by considering the two extremes cases of distribution $\mathbf{p}$ such that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{u}) > \varepsilon$: First, when a constant fraction of the elements have probability $(1-\Omega(\varepsilon))/k$ and the rest have probability more than $1/k$, in which case we only need $m = O(1)$ batches to find such an element $i$ and $O(k/\varepsilon^2)$ players per batch to detect the bias. Second, when a fraction $O(\varepsilon)$ of the elements have probability $0$ and the rest have probability more than $1/k$, in which case we need $m = O(1/\varepsilon)$ batches to find such an element, but now only $O(k/\varepsilon)$ players per batch to detect the bias. In both cases, the total number of players should be $O(k/\varepsilon^2)$, in contrast to the $O(k\log 1/\varepsilon/\varepsilon^3)$ of the scheme described above. To circumvent this difficulty, we take recourse to a technique known as *Levin's work investment strategy*; see, for instance, [Gol14, Appendix A.2] for a review (cf. [Gol17, Section 8.2.4]). Heuristically, this technique allows us to identify an appropriate "scale" and invests matching "work" effort to it. Formally, we have following lemma:

**Lemma 6.7** ([Gol14, Fact A.2])**.** *Consider a random variable $X$ taking values in $\mathcal{X}$, a mapping $q\colon \mathcal{X} \to [0,1]$, and $\varepsilon \in (0,1]$. Suppose that $\mathbb{E}[q(X)] > \varepsilon$, and let $L := \lceil \log(2/\varepsilon) \rceil$. Then, there exists $j^* \in [L]$ such that $\Pr\left[ q(X) > 2^{-j^*} \right] > 2^{j^*}\varepsilon/(L+5-j^*)^2$.*

The next result follows upon modifying the warmup protocol by using this lemma to decide on the size and the number of batches.

**Proposition 6.8.** *There exists a $1$-bit public-coin $(k,\varepsilon)$-uniformity testing protocol for $n = O(k/\varepsilon^2)$, which uses $\tilde{O}((\log k)/\varepsilon)$ bits of randomness.*

*Proof.* Consider a $\mathbf{p}$ that is $\varepsilon$-far from uniform and set $L := \lceil \log(2/\varepsilon) \rceil$. We apply Lemma 6.7 to the function $q\colon [k] \to [0,1]$ given by

$$q(i) := k\left(\frac{1}{k} - \mathbf{p}_i\right)\mathbb{1}_{\{\mathbf{p}_i < 1/k\}}.$$

Note that for a uniformly distributed $X$, we have

$$\mathbb{E}[q(X)] = \sum_{i=1}^{k} \frac{1}{k} q(i) = \mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{u}) > \varepsilon.$$

Therefore, by Lemma 6.7, there exists $j^* \in [L]$ such that

$$\Pr_{i\sim\mathbf{u}}\left[ \mathbf{p}_i < \frac{1}{k} - \frac{2^{-j^*}}{k} \right] > 2^{j^*}\frac{\varepsilon}{(L+5-j^*)^2}\,. \tag{9}$$

We now proceed in a similar manner as the warmup protocol, with one batch invested for each $j \in [L]$. Specifically, consider $L$ batches of players with the $j$-th batch comprising $n_j := m_j \cdot O(2^{2j}k \cdot \log(1/\delta_j))$ players; both quantities $\delta_j$ and $m_j$ will be specified later. The $j$th batch assumes that $j \in [L]$ will satisfy (9) and further divides its assigned set of players into $m_j$ mini-batches. Each mini-batch selects an $i$ uniformly from $[k]$ and applies the previously mentioned test to distinguish if the probability of $i$ is $1/k$ or $(1-2^{-j})/k$ with probability $(1-\delta_j)$. This, as before, requires the assigned $O(2^{2j}k\log(1/\delta_j))$ players. Note that if

$$m_j \geq 5(L+5-j)^2/(2^j\varepsilon),$$

whenever $j$ satisfies (9), with probability more than $9/10$ at least one of the mini-batches assigned to $j$ will select an $i$ for which $\mathbf{p}_i < (1-2^{-j})/k$. Our uniformity testing protocol is as before:

Accept uniform if every mini-batch of every batch accepts $1/k$ as the probability of their respectively assigned elements; else declare $\varepsilon$-far from uniform.

If the underlying distribution is indeed uniform, the protocol will reject it when at least one of the mini-batches erroneously rejects $1/k$, an event which occurs with probability at most

$$\sum_{j=1}^{L} m_j\delta_j \leq \frac{1}{10}\sum_{j=1}^{L}\frac{1}{(L+5-j)^2} < \frac{1}{10}\sum_{j\geq 5}\frac{1}{j^2} < \frac{1}{40},$$

when we select

$$\delta_j \leq \frac{1}{10(L+5-j)^2 m_j}.$$

Note that this choice of $\delta_j$ depending on $j$ is important for omitting the extra $\log(1/\varepsilon)$ cost that appeared in the warmup protocol.

If the underlying distribution is $\varepsilon$-far from uniform, an error occurs if, for every $j$, no mini-batch of batch $j$ selects an element $i$ that satisfies (9) or if the mini-batch test fails. By construction, there exists a $j$ that satisfies (9), and by our choice of $m_j$, all the mini-batches assigned to it fails to select an $i$ in the set of (9) with probability less than $1/10$. On the other hand, the mini-batch makes an error with probability less than $\delta_j < 1/40$. Thus, the overall probability of error in this case is less than $1/8$. To conclude, the overall protocol makes an error with probability at most $1/8$ in both cases.

Finally, we can bound the total number of players $n$ required by the protocol as

$$\begin{aligned}
n &= c\sum_{j=1}^{L} m_j \cdot 2^{2j} k \cdot \log \frac{1}{\delta_j} \\
&\le \frac{ck}{\varepsilon}\sum_{j=1}^{L}(L+5-j)^2 2^j \log \frac{10(L+5-j)^4}{2^j\varepsilon} \\
&\le O\left(\frac{k}{\varepsilon}\right)\sum_{j=1}^{L}(L+5-j)^2 2^j \log \frac{2^{j-L-5}}{L+5-j} \\
&\le O\left(\frac{k}{\varepsilon}\right)2^L \sum_{j'=5}^{L}(j')^2 2^{-j'} \log(j'2^{j'}) \\
&\le O\left(\frac{k}{\varepsilon}\right)2^L \sum_{j'=5}^{L}(j')^2 2^{-j'} \log j' + O\left(\frac{k}{\varepsilon}\right)2^L \sum_{j'=5}^{L}(j')^3 2^{-j'} \\
&\le O\left(\frac{k}{\varepsilon^2}\right),
\end{aligned}$$

where the final bound uses $\sum_{j\ge 5} 2^{-j} j^\alpha \log j = O(1)$ for every $\alpha$. To conclude, note that the total number of random bits required is $O(\log k)\cdot\sum_{j=1}^{L} m_j = O(\log k \cdot \log^2(1/\varepsilon)/\varepsilon)$. □

Next, we move to the more general case when $\ell \ge 1$ bits of communication per player are allowed and establish Theorem 6.1. While we build on the heuristics developed thus far, the form of the protocol deviates. Instead of assigning one symbol to each mini-batch, we now assign a subset of size $s = 2^\ell - 1$ to each mini-batch; one $\ell$-bit sequence is reserved to indicate that none of the symbol in the subset occured. The referee uses the symbols occuring in the subset to distinguish uniform and $\varepsilon$-far from uniform, which can be done when conditional distributions (i.e., given that the symbols lie in the subset) are separated in total variation distance. We use Levin's work investment strategy again to decide how many players must be assigned to each subset. But now there is a new constraint: If a subset has small probability, then we need to assign a large number of mini-batches to get symbols from it. However, we can circumvent this difficulty by noting that if the subset has probability smaller than $(1-\varepsilon)s/k$, we can anyway distinguish the underlying distribution from uniform using $O(k/(s\varepsilon^2))$ players. Thus, we can condition on the complementary event by adding extra $O(k/(s\alpha^2))$ players per batch and take $O(k/s)$ mini-batches to get at least one mini-batch assigned to a good subset. Note that under the uniform distribution the conditional distribution given a subset of size $s$ is uniform on $[s]$. Then, we can distinguish the conditional distributions using roughly[14] $O(\sqrt{s}/\varepsilon^2)$ players. The overall number of players is dominated by the players assigned for the conditional test and is given by $O\Big(k/(2^{\ell/2}\varepsilon^2)\Big)$. We provide the formal proof next.

**Theorem 6.9.** *Let $\ell \in \{1,\dots,\log k\}$. Then, there exists an $\ell$-bit public-coin $(k,\varepsilon)$-uniformity testing protocol for $n = O\Big(\frac{k}{2^{\ell/2}\varepsilon^2}\Big)$ players, which uses $O_\varepsilon(2^\ell \log k)$ bits of randomness.*

[14] The good subset need not have the conditional distributions separated by exactly $\varepsilon$, but this is where we use Levin's work investment strategy to get an effective separation of $\varepsilon$.

*Proof of Theorem 6.9.* Set $L := \lceil \log(2/\varepsilon) \rceil$ and define $q$ as in the proof of Proposition 6.8. For a subset $S \subseteq [k]$, denote by $\mathbf{p}^S$ the conditional distribution

$$\mathbf{p}_i^S = \frac{\mathbf{p}_i \mathbb{1}_{\{i \in S\}}}{\mathbf{p}(S)},$$

where $\mathbf{p}(S) = \sum_{i \in S} \mathbf{p}_i$. Observe that if $\mathbf{p} = \mathbf{u}$, then for every subset $S \subseteq [k]$ we have $\mathbf{p}_i^S = 1/|S|$ for every $i \in S$. On the other hand, when $\mathbf{p}$ is $\varepsilon$-far from uniform, we have the following result:

**Claim 6.10.** *Suppose* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}) > \varepsilon$. *For any* $1 \le s \le k$ *and* $S \subseteq [k]$ *of size* $s$ *chosen uniformly at random, we have*

$$\mathbb{E}_S \sum_{i \in S} \mathbb{1}_{\{\mathbf{p}_i \le \frac{1}{k}\}} \left( \frac{1}{k} - \mathbf{p}_i \right) > \varepsilon \cdot \frac{s}{k}.$$

*Proof.* On expanding the expectation, we obtain

$$\begin{aligned}
&\mathbb{E}_S \sum_{i \in S} \mathbb{1}_{\{\mathbf{p}_i \le \frac{1}{k}\}} \left( \frac{1}{k} - \mathbf{p}_i \right) \\
&\quad = \frac{1}{\binom{k}{s}} \sum_{S \subseteq [k] : |S| = s} \sum_{i \in S} \mathbb{1}_{\{\mathbf{p}_i \le \frac{1}{k}\}} \left( \frac{1}{k} - \mathbf{p}_i \right) = \frac{1}{\binom{k}{s}} \sum_{i=1}^{k} \sum_{\substack{S \subseteq [k] : |S| = s \\ S \ni i}} \mathbb{1}_{\{\mathbf{p}_i \le \frac{1}{k}\}} \left( \frac{1}{k} - \mathbf{p}_i \right) \\
&\quad = \frac{1}{\binom{k}{s}} \sum_{i=1}^{k} \mathbb{1}_{\{\mathbf{p}_i \le \frac{1}{k}\}} \left( \frac{1}{k} - \mathbf{p}_i \right) \sum_{\substack{S' \subseteq [k] \setminus i \\ |S'| = s-1}} 1 = \frac{1}{\binom{k}{s}} \sum_{i=1}^{k} \mathbb{1}_{\{\mathbf{p}_i \le \frac{1}{k}\}} \left( \frac{1}{k} - \mathbf{p}_i \right) \cdot \binom{k-1}{s-1} \\
&\quad = \frac{s}{k} \sum_{i=1}^{k} \mathbb{1}_{\{\mathbf{p}_i \le \frac{1}{k}\}} \left( \frac{1}{k} - \mathbf{p}_i \right) \\
&\quad = \frac{s}{k} \cdot \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}) > \varepsilon \cdot \frac{s}{k}.
\end{aligned}$$

□

For brevity, set $s := 2^\ell - 1$. Using Claim 6.10 together with Lemma 6.7 we get that there exists $j^* \in [L]$ such that

$$\Pr_S \left[ \sum_{i \in S} \mathbb{1}_{\left\{ \frac{k \mathbf{p}_i}{s} \le \frac{1}{s} \right\}} \left( \frac{1}{s} - \mathbf{p}_i \cdot \frac{k}{s} \right) > 2^{-j^*} \right] > 2^{j^*} \cdot \frac{\varepsilon}{(L + 5 - j^*)^2}. \tag{10}$$

Note that the event on the left-side of the inequality above essentially bounds the total variation distance between $\mathbf{u}_S$ and $\mathbf{p}^S$ when $\mathbf{p}(S) \approx s/k$. Therefore, in case players have access to such a subset $S$, they can accomplish uniformity testing by applying a standard uniform test for a domain of size $s$. Thus, as in the $\ell = 1$ protocol, we can use a public randomness to select a subset $S$ randomly and assign it to an appropriate number of players; this constitutes one mini-batch, and we need one mini-batch per $j \in [L]$. However, this will only work if the selected $S$ has $\mathbf{p}(S) \approx s/k$. To circumvent this difficulty, we use a separate test for checking closeness of $\mathbf{p}(S)$ to $s/k$. Specifically, we once again use the fact that a coin with bias $s/k$ can be distinguished from another with bias outside the interval $[(1-\alpha)s/k, (1+\alpha)s/k$ with probability of error less than $\delta$ using $O(k \log(1/\delta)/(s\alpha^2))$ independent coin tosses.

Once we have verified that the set $S$ has probability close to $s/k$, we can apply a standard uniformity test. Indeed, we set $\alpha = 2^{-j^*}/8$, and once the test above has verified that $\mathbf{p}(S) \in [1-2^{-j^*}/8, 1+2^{-j^*}/8] \cdot \frac{s}{k}$, the total variation distance of $\mathbf{p}^S$ to uniformity can be bounded as follows: Let vector $\tilde{\mathbf{p}}^S$ be given by $\tilde{\mathbf{p}}^S_i = \mathbf{p}_i k/s$. Then, by the triangle inequality we get

$$\mathrm{d}_{\mathrm{TV}}\left(\mathbf{p}^S, \mathbf{u}_S\right) = \frac{1}{2}\|\mathbf{p}^S - \mathbf{u}_S\|_1 \geq \frac{1}{2}\left(\|\tilde{\mathbf{p}}^S - \mathbf{u}_S\|_1 - \|\tilde{\mathbf{p}}^S - \mathbf{p}^S\|_1\right) > \frac{1}{2}\left(2^{-j^*} - \|\tilde{\mathbf{p}}^S - \mathbf{p}^S\|_1\right).$$

Further,

$$\|\tilde{\mathbf{p}}^S - \mathbf{p}^S\|_1 = \sum_{i\in S} \mathbf{p}_i \left|\frac{k}{s} - \frac{1}{\mathbf{p}(S)}\right| \leq \frac{2^{-j}k}{4s} \cdot \mathbf{p}(S) \leq \frac{2^{-j}}{3},$$

which gives

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_S, \mathbf{u}_S) \geq \frac{2^{-j}}{3}. \tag{11}$$

Therefore, upon the first test verifying that $\mathbf{p}(S)$ is sufficiently close to $s/k$, we can proceed to testing $\mathbf{p}^S$ versus $\mathbf{u}_S$. To that end, we need sufficiently many samples from the $\mathbf{p}^S$, which we generate using rejection sampling.

We have now collected all the components needed for our scheme. As in the $\ell = 1$ case, set parameters $\varepsilon_j = 2^{-j}/8$, $m_j = (L+5-j)^2/(2^j\varepsilon)$, and $\delta_j = 1/(10(L+5-j)^2 m_j)$. Consider $L$ batches of players, with the $j$th batch comprising $m_j$ mini-batches of

$$n_j = c_1\left(\frac{k}{s\varepsilon_j^2}\log\frac{1}{\delta_j}\right) + c_2\left(\frac{k}{s}\log\frac{1}{\delta_j}\right)\cdot c_3\left(\frac{\sqrt{s}}{\varepsilon_j^2}\log\frac{1}{\delta_j}\right)$$

players each; the constants $c_1, c_2, c_3$ will be set to get appropriate probability of errors. Each mini-batch of the $j$th batch generates a random subset $S$ of $[k]$. Each player in the mini-batch communicates as follows: It sends the all-zero sequence of length $\ell$ to indicate if its observed element is not in $S$ and, otherwise, uses the remaining sequences of length $\ell$ to indicate which of the $s = 2^\ell - 1$ elements it has observed. The referee uses the communication from the first $(c_1 k \log 1/\delta_j/(s\varepsilon_j^2))$ players of the mini-batch to check if the $|\mathbf{p}(S) - s/k| < \varepsilon_j s/k$ or not. If it is not, the mini-batch fails. Else, the referee considers the communication from players that did not send the all-zero sequence (i.e., those players that saw elements in $S$) and tests if the conditional distribution $\mathbf{p}^S$ is uniform on $S$ or not. If it is not, the mini-batch fails; the referee declares uniformity if none of the mini-batches declared failure.

The analysis of this protocol is completed in a similar manner to that of the $\ell = 1$ protocol. If the underlying distribution is uniform, the output is erroneous if at least one of the mini-batches declared failure. This can happen in two ways: First, if the test based on communication from the first set of $c_1 k\log(1/\delta_j)/s\varepsilon_j^2$ players erroneously declared fail, an event that can happen with probability $\delta_j/3$ for an appropriately chosen constant $c_1$. Second, if the first test passes, but the uniformity test for $\mathbf{p}^S$ based on the remaining remaining set of players fails, which can happen either when there are less than $c_2\sqrt{s}/\varepsilon_j^2\log(1/\delta_j)$ players that see samples from $S$, which given that the first set has passed will fail with probability less than $\delta_j/3$ for an appropriate $c_2$, or when the second test fails which for an appropriate $c_3$ will happen with probability less than $\delta_j/3$ by Eq. (11). Thus, the overall probability of error is less than $\sum_{j=1}^{L} m_j\delta_j < 1/40$.

If the underlying distribution is $\varepsilon$-far from uniform, the test will erroneously select the uniform if the referee makes an error for the mini-batches corresponding to $j^*$ guaranteed by (10).

But this can only happen if either none of these mini-batches select an $S$ satisfying the condition on the left-side of (10), which happens with probability less than $1/10$ or a mini-batch that selected an appropriate $S$ failed the second test, which can happen with probability $\delta_j/3 \leq 1/10$. Thus, the overall probability of error is less than $2/10$.

We complete the proof by evaluating the number of players used by the protocol. As in the proof for the $\ell = 1$ case, we have that the total number of players $n$ satisfies

$$\begin{aligned} n &\leq \sum_{j=1}^{L} m_j O\left(\frac{k}{s\varepsilon_j^2}\log\frac{1}{\delta_j} + \frac{k}{\sqrt{s}\varepsilon_j^2}\left(\log\frac{1}{\delta_j}\right)^2\right) \\ &\leq \sum_{j=1}^{L} m_j O\left(\frac{k}{\sqrt{s}\varepsilon_j^2}\left(\log\frac{1}{\delta_j}\right)^2\right) \\ &\leq O\left(\frac{k}{\sqrt{s}\varepsilon}\right) 2^L \sum_{j=5}^{L} 2^{-j}(j)^5(\log j)^2 \\ &\leq O\left(\frac{k}{2^{\frac{\ell}{2}}\varepsilon^2}\right), \end{aligned}$$

where we followed the same steps as the bound for $\ell = 1$ case. To conclude, note that since each subset $S$ of size $s$ requires $\log\binom{k}{s} = O(2^\ell \log\frac{k}{2^\ell})$ bits to specify, the total number of random bits required is $O(2^\ell \log\frac{k}{2^\ell}) \cdot \sum_{j=1}^{L} m_j = O(2^\ell \log\frac{k}{2^\ell} \cdot \log^2(1/\varepsilon)/\varepsilon)$. □

## 6.2 Lower bound for public-coin protocols

We now establish a lower bound on the number of players $n$ required for any $\ell$-bit public-coin $(k,\varepsilon)$-uniformity testing protocol. We begin with the simpler setting of $\ell = 1$ and establish Theorem 6.2 for this special case, before moving to the more general case. The same construction is used for both the restricted and the general case, but the simpler proof we present for the special case does not yield the more general result.

**Proposition 6.11.** *Any $1$-bit public-coin $(k,\varepsilon)$-uniformity testing protocol must have $n = \Omega(k/\varepsilon^2)$ players.*

*Proof.* Without loss of generality, we assume that $k$ is even. We consider the standard "Paninski construction":[15] For every $\theta \in \{-1,1\}^{k/2}$, let

$$\mathbf{p}_\theta(2i-1) = \frac{1+2\varepsilon\theta_i}{k}, \qquad \mathbf{p}_\theta(2i-1) = \frac{1-2\varepsilon\theta_i}{k}, \quad \forall\, i \in [k/2].$$

Note that each distribution $\mathbf{p}_\theta$ is at a total variation distance $\varepsilon$ from the uniform distribution $\mathbf{u}$ on $[k]$. Thus, any $(k,\varepsilon)$-uniformity testing protocol should be able to distinguish between any distribution $\mathbf{p}_\theta$ and $\mathbf{u}$. We will establish an upper bound on the average total variation distance between $\mathbf{p}_\theta$ and $\mathbf{u}$, whereby there must be at least one $\mathbf{p}_\theta$ satisfying the bound. The desired lower bound for the number of players will then follow from the standard Le Cam's two-point method argument.

[15] This construction was given in [Pan08] to prove the lower bound for the sample complexity of uniformity testing in the standard centralized setting.

We derive the aforementioned upper bound on average total variation distance for private-coin protocols first. Specifically, consider a private-coin protocol for uniformity testing where, as before, the 1-bit communication of player $j$ is described by the channel $W_j\colon [k] \to \{0,1\}$ such that $W_j(1|x) \in [0,1]$ is the probability that player $j$ sends 1 to the referee upon observing $x$. For any $1 \le j \le n$, it is immediate to see that, if $\mathbf{u}$ is the underlying distribution, the probability that player $j$ sends 1 to the referee is

$$\rho_j^{\mathbf{u}} := \frac{2}{k}\sum_{i=1}^{k/2}\left(\frac{W_j(1|2i-1)+W_j(1|2i)}{2}\right),$$

while under $\mathbf{p}_\theta$ it is

$$\begin{aligned}\rho_j^{\theta} &:= \frac{2}{k}\sum_{i=1}^{k/2}\left(\frac{W_j(1|2i-1)+W_j(1|2i)+2\varepsilon\theta_i\left(W_j(1|2i-1)-W_j(1|2i)\right)}{2}\right)\\ &= \rho_j^{\mathbf{u}} + \frac{\varepsilon}{k}\sum_{i=1}^{k/2}\theta_i\left(W_j(1|2i-1)-W_j(1|2i)\right).\end{aligned}$$

Moreover, since each player gets an independent sample from the same distribution $\mathbf{p} \in \{\mathbf{u}\} \cup \{\mathbf{p}_\theta\}_{\theta\in\{-1,1\}^{k/2}}$, the observation of the referee $r \in \{0,1\}^n$ is generated from a product distribution. Specifically, the bits communicated by the players are independent with the $j$th bit distributed as $\mathrm{Bern}(\mathbf{p}_j^{\mathbf{u}})$ or $\mathrm{Bern}(\mathbf{p}_j^{\theta})$, respectively, when the underlying distribution of the sample are $\mathbf{u}$ or $\mathbf{p}_\theta$. Denoting by $\mathbf{R}^{\mathbf{u}}$ and $\mathbf{R}^{\theta}$ the distributions of the transmitted bits under $\mathbf{u}$ and $\mathbf{p}_\theta$, repectively, we have

$$\begin{aligned}\mathrm{d_{TV}}\left(\mathbf{R}^{\mathbf{u}},\mathbf{R}^{\theta}\right)^2 &\le \frac{1}{2}D\left(\mathbf{R}^{\theta}\|\mathbf{R}^{\mathbf{u}}\right)\\ &= \frac{1}{2}\sum_{j=1}^{n}D\left(\mathrm{Bern}(\mathbf{p}_j^{\theta})\|\,\mathrm{Bern}(\mathbf{p}_j^{\mathbf{u}})\right)\\ &\le \frac{1}{2}\sum_{j=1}^{n}\chi^2(\mathrm{Bern}(\mathbf{p}_j^{\theta}),\mathrm{Bern}(\mathbf{p}_j^{\mathbf{u}}))\\ &= \frac{1}{2}\sum_{j=1}^{n}\frac{(\rho_j^{\mathbf{u}}-\rho_j^{\theta})^2}{\rho_j^{\mathbf{u}}(1-\rho_j^{\mathbf{u}})},\end{aligned}$$

where the first inequality is Pinsker's inequality, the second inequality uses $\ln x \le (x-1)$. Further, abbreviating for convenience $\alpha_{i,j} := W_j(1|2i-1)$ and $\beta_{i,j} := W_j(1|2i)$ for $i \in [k/2]$ and $j \in [n]$, to bound the right-side we note that

$$\begin{aligned}\frac{(\rho_j^{\mathbf{u}}-\rho_j^{\theta})^2}{\rho_j^{\mathbf{u}}(1-\rho_j^{\mathbf{u}})} &= \frac{4\varepsilon^2}{k^2}\frac{1}{\rho_j^{\mathbf{u}}(1-\rho_j^{\mathbf{u}})}\left(\sum_{i=1}^{k/2}\theta_i\left(\alpha_{i,j}-\beta_{i,j}\right)\right)^2\\ &= \frac{4\varepsilon^2}{k^2}\frac{1}{\rho_j^{\mathbf{u}}(1-\rho_j^{\mathbf{u}})}\sum_{i,i'=1}^{k/2}\theta_i\theta_i'\left(\alpha_{i,j}-\beta_{i,j}\right)\left(\alpha_{i',j}-\beta_{i',j}\right).\end{aligned}$$

On taking expectation over $\theta$, we obtain

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{(\rho_j^{\mathbf{u}}-\rho_j^\theta)^2}{\rho_j^{\mathbf{u}}(1-\rho_j^{\mathbf{u}})}\right] &= \frac{4\varepsilon^2}{k^2}\frac{1}{\rho_j^{\mathbf{u}}(1-\rho_j^{\mathbf{u}})}\sum_{i,i'=1}^{k/2}\mathbb{E}_\theta[\theta_i\theta_i']\left(\alpha_{i,j}-\beta_{i,j}\right)\left(\alpha_{i',j}-\beta_{i',j}\right)\\ &= \frac{4\varepsilon^2}{k^2}\frac{1}{\rho_j^{\mathbf{u}}(1-\rho_j^{\mathbf{u}})}\sum_{i=1}^{k/2}\left(\alpha_{i,j}-\beta_{i,j}\right)^2\\ &= \frac{4\varepsilon^2}{k^2}\frac{\sum_{i=1}^{k/2}\left(\alpha_{i,j}-\beta_{i,j}\right)^2}{\frac{2}{k}\sum_{i=1}^{k/2}\frac{\alpha_{i,j}+\beta_{i,j}}{2}\left(1-\frac{2}{k}\sum_{i=1}^{k/2}\frac{\alpha_{i,j}+\beta_{i,j}}{2}\right)}.\end{aligned}$$

To bound the expression on the right-side further, we consider the case when $\frac{2}{k}\sum_{i=1}^{k/2}\frac{\alpha_{i,j}+\beta_{i,j}}{2} \le 1/2$; the other case can be handled similarly by symmetry. We have

$$\begin{aligned}\frac{\sum_{i=1}^{k/2}\left(\alpha_{i,j}-\beta_{i,j}\right)^2}{\frac{2}{k}\sum_{i=1}^{k/2}\frac{\alpha_{i,j}+\beta_{i,j}}{2}\left(1-\frac{2}{k}\sum_{i=1}^{k/2}\frac{\alpha_{i,j}+\beta_{i,j}}{2}\right)} &\le 2\frac{\sum_{i=1}^{k/2}\left(\alpha_{i,j}-\beta_{i,j}\right)^2}{\frac{2}{k}\sum_{i=1}^{k/2}\frac{\alpha_{i,j}+\beta_{i,j}}{2}}\\ &\le 2k\frac{\sum_{i=1}^{k/2}\left|\alpha_{i,j}-\beta_{i,j}\right|\left(\alpha_{i,j}+\beta_{i,j}\right)}{\sum_{i=1}^{k/2}\left(\alpha_{i,j}+\beta_{i,j}\right)}\\ &\le 2k\max_{1\le i\le k/2}\left|\alpha_{i,j}-\beta_{i,j}\right| \le 2k,\end{aligned}$$

where the previous inequality holds since $\alpha_{i,j}-\beta_{i,j}\in[-1,1]$ for all $i,j$. Combining the foregoing bounds yields

$$\mathbb{E}_\theta\left[\mathrm{d_{TV}}\left(\mathbf{R}^{\mathbf{u}},\mathbf{R}^\theta\right)^2\right] \le \frac{4\varepsilon^2}{k}\cdot n.$$

In particular, there exists a fixed $\theta$ for which $\mathrm{d_{TV}}\left(\mathbf{R}^{\mathbf{u}},\mathbf{R}^\theta\right)^2 \le 4\varepsilon^2 n/k$. By the two-point method argument, the uniformity testing protocol can only distinguish $\mathbf{u}$ and $\mathbf{p}_\theta$ if $\mathrm{d_{TV}}\left(\mathbf{R}^{\mathbf{u}},\mathbf{R}^\theta\right)=\Omega(1)$, which yields $n=\Omega(k/\varepsilon^2)$ as claimed.

Finally, to extend the result to public-coin protocols, note that the observation of the referee now includes $U$ in addition to the communication. Denote by $\mathbf{R}_U^{\mathbf{u}}$ and $\mathbf{R}_U^\theta$ the distribution of the communicated bits under $\mathbf{u}$ and $\mathbf{p}_\theta$, respectively. Then the total variational distance between the distributions of the observation of the adversary under the two distributions is given by $\mathbb{E}_U\left[\mathrm{d_{TV}}\left(\mathbf{R}_U^{\mathbf{u}},\mathbf{R}_U^\theta\right)^2\right]$. Therefore, it suffices to find a uniform upper bound for the expected value of the total variation with respect to $\theta$ for different fixed values of the public randomness $U$. This uniform bound can be shown to be $4\varepsilon^2 n/k$ by repeating the proof above for every fixed $U=u$. □

Moving now to the case of a general $\ell$, we can follow the argument above to obtain an $O(\varepsilon^2 n 2^\ell/k)$ upper bound for the expected total variation distance. However, this only yields an $\Omega(k/(2^\ell\varepsilon^2))$ lower bound for $n$, which is off by a factor of $2^{\ell/2}$ from the desired bound of Theorem 6.1. The slackness in the bound stems from the gap between the average total variation distance and the total variation distance between the average $\mathbf{p}_\theta$ and $\mathbf{u}$. Indeed, since the uniformity testing protocol can distinguish $p_\theta$ and $\mathbf{u}$ for every $\theta$, it can also distinguish $\mathbb{E}_\theta[p_\theta]$ and $\mathbf{u}$. Note that the

KL-divergence-based bound for total variation distance used above is not amenable to handling the distance between a mixture of product distribution and a fixed product distribution. Instead, we take recourse to an argument of Pollard [Pol03] which established, in essence, the following result.

**Lemma 6.12.** *For any two product distributions $P^n = P_1 \times \cdots \times P_n$ and $Q^n = Q_1 \times \cdots \times Q_n$ on the alphabet $\mathcal{X}$,*

$$\chi^2(Q^n, P^n) = \prod_{i=1}^{n}(1 + \chi^2(Q_i, P_i)) - 1.$$

For our application, we need to extend this to the case when the product distribution $P^n$ is replaced by a mixture of product distributions. To that end, we use the following result which is a slight but crucial extension of this result, also described in [Pol03] (a similar observation was used in [Pan08]); we include a proof for completeness.

**Lemma 6.13.** *Consider a random variable $Z$ such that for each $Z = z$ the distribution $Q_z^n$ is defined as $Q_{1,z} \times \cdots \times Q_{n,z}$. Further, let $P^n = P_1 \times \cdots \times P_n$ be a fixed product distribution. Then,*

$$\chi^2(\mathbb{E}_Z[Q_Z^n], P^n) = \mathbb{E}_{ZZ'}\left[\prod_{i=1}^{n}(1 + H_i(Z, Z'))\right] - 1,$$

*where $Z'$ is an independent copy of $Z$ and, with $\Delta_i^z$ denoting $(Q_{i,z}(X_i) - P_i(X_i))/P_i(X_i)$,*

$$H_i(z, z') = \mathbb{E}\left[\Delta_i^z \Delta_i^{z'}\right],$$

*where the expectation is over $X_i$ distributed according to $P_i$.*

*Proof.* Using the definition of $\chi^2$-distance, we have

$$\begin{aligned}\chi^2(\mathbb{E}_Z[Q_Z^n], P^n) &= \mathbb{E}_{P^n}\left[\left(\mathbb{E}_Z\left[\frac{Q_Z^n(X^n)}{P^n(X^n)}\right]\right)^2\right] - 1 \\ &= \mathbb{E}_{P^n}\left[\left(\mathbb{E}_Z\left[\prod_{i=1}^{n}(1 + \Delta_i^Z)\right]\right)^2\right] - 1,\end{aligned}$$

where the outer expectation is for $X^n$ using the distribution $P^n$. The product in the expression above can be expanded as

$$\prod_{i=1}^{n}(1 + \Delta_i^Z) = 1 + \sum_{i \in [n]} \Delta_i^Z + \sum_{i_1 > i_2} \Delta_{i_1}^Z \Delta_{i_2}^Z + \ldots,$$

whereby we get

$$\begin{aligned}\chi^2(\mathbb{E}_Z[Q_Z^n], P^n) &= \mathbb{E}_{P^n}\left[\left(1 + \sum_i \mathbb{E}_Z\left[\Delta_i^Z\right] + \sum_{i_1 > i_2} \mathbb{E}_Z\left[\Delta_{i_1}^Z \Delta_{i_2}^Z\right] + \ldots\right)^2\right] - 1 \\ &= \mathbb{E}_{P^n}\left[\sum_i \mathbb{E}_Z\left[\Delta_i^Z\right] + \sum_j \mathbb{E}_{Z'}\left[\Delta_j^{Z'}\right] + \sum_{i,j} \mathbb{E}_{Z,Z'}\left[\Delta_i^Z \Delta_j^{Z'}\right] \ldots\right].\end{aligned}$$

Observe now that $\mathbb{E}_{P^n}[\Delta_i^z] = 0$ for every $z$. Furthermore, $Z$ is an independent copy of $Z'$ and $\Delta_i^Z$ and $\delta_j^Z$ are independent for $i \neq j$. Therefore, the expectation on the right-side above equals

$$\mathbb{E}\left[\sum_i H_i(Z,Z') + \sum_{i_1>i_2} H_{i_1}(Z,Z')H_{i_2}(Z,Z') + \ldots\right] = \mathbb{E}\left[\prod_{i=1}^n (1+H_i(Z,Z'))\right] - 1,$$

which completes the proof. □

We are now in a position to establish Theorem 6.2.

*Proof of Theorem 6.2.* As before, it suffices to derive a uniform upper bound for the total variation distance between the message distributions for private-coin protocols. In fact, it suffices to consider deterministic protocols since for a fixed public randomness $U = u$, the protocol is deterministic. We apply Lemma 6.13 to the distribution of the messages for a deterministic protocol; we retain the channel $W_j$ notation from the $\ell = 1$ proof with the understanding that it denotes a deterministic map. Note that the messages are independent under uniform and under $\mathbf{p}_\theta$ (for a fixed public randomness $U$). For our setting, $\theta$ plays the role of $Z$ in Lemma 6.13. Note that under uniform observations, player $j$ sends the message $m_j \in \{0,1\}^\ell$ with probability

$$\rho_{j,m}^{\mathbf{u}} = \frac{2}{k}\sum_{i=1}^{k/2}\left(\frac{W_j(m|2i-1) + W_j(m|2i)}{2}\right),$$

and under $\mathbf{p}_\theta$ with probability

$$\rho_{j,m}^{\theta} = \rho_{j,m}^{\mathbf{u}} + \frac{\varepsilon}{k}\sum_{i=1}^{k/2}\theta_i\left(W_j(m|2i-1) - W_j(m|2i)\right).$$

Therefore, the quantity $\Delta_i^\theta$ required in Lemma 6.13 is given by

$$\Delta_j^\theta = \frac{\varepsilon\sum_{i=1}^{k/2}\theta_i(W_j(M_j|2i) - W_j(M_j|2i-1))}{\sum_{i=1}^{k/2}(W_j(M_j|2i) + W_j(M_j|2i-1))},$$

where $M_j$ is the random message sent under the uniform distribution. Consequently, we can express $H_j(\theta,\theta')$ of Lemma 6.13, $1 \le j \le n$, as

$$\begin{aligned} H_j(\theta,\theta') &= \frac{\varepsilon^2}{k}\cdot\sum_{m\in\{0,1\}^\ell}\sum_{i_1,i_2\in[k/2]}\theta_{i_1}\theta'_{i_2}\frac{\left(W_j(m|2i_1-1) - W_j(m|2i_1)\right)\left(W_j(m|2i_2-1) - W_j(m|2i_2)\right)}{\sum_{i=1}^{k/2}\left(W_j(m|2i-1) + W_j(m|2i)\right)} \\ &= \frac{\varepsilon^2}{k}\cdot\theta^T H_j\theta', \end{aligned}$$

where $\theta^T$ denotes the transpose of the vector $\theta \in \{-1,+1\}^{k/2}$ and $H_j$ is an $[k/2]\times[k/2]$ matrix with the $(i_1,i_2)$th entry given by

$$\sum_{m\in\{0,1\}^\ell}\frac{\left(W_j(m|2i_1-1) - W_j(m|2i_1)\right)\left(W_j(m|2i_2-1) - W_j(m|2i_2)\right)}{\sum_{i=1}^{k/2}\left(W_j(m|2i-1) + W_j(m|2i)\right)}.$$

Note that the matrix $H_j$ is symmetric (in fact, it has the outer product form $AA^T$ for an $(k/2) \times 2^\ell$ matrix $A$). Therefore, we obtain from Lemma 6.13 that

$$\begin{aligned}\mathbb{E}\Big[\mathrm{d_{TV}}\Big(\mathbb{E}_\theta\Big[\mathbf{R}^\theta\Big], \mathbf{R}^{\mathbf{u}}\Big)^2\Big] &\le \frac{1}{4}\mathbb{E}\Big[\chi^2(\mathbb{E}_\theta\Big[\mathbf{R}^\theta\Big], \mathbf{R}^{\mathbf{u}})\Big] \\ &= \frac{1}{4}\Big(\mathbb{E}_{\theta\theta'}\left[\prod_{j=1}^{n}\left(1 + \frac{\varepsilon^2}{k}\theta^T H_j \theta'\right)\right] - 1\Big) \\ &\le \frac{1}{4}\Big(\mathbb{E}_{\theta\theta'}\left[e^{\frac{n\varepsilon^2}{k}\theta^T \bar{H}\theta'}\right] - 1\Big),\end{aligned}$$

where we have used $1 + x \le e^x$ and

$$\bar{H} = \frac{1}{n}\sum_{j=1}^{n} H_j.$$

Thus, we need to bound the moment generating function of the random variable $\theta^T \bar{H}\theta'$. We will establish a sub-Gaussian bound using a by-now-standard bound that follows from transportation method. Specifically, we show the following:

**Claim 6.14.** *Consider random vectors $\theta, \theta' \in \{-1,1\}^{k/2}$ with each $\theta_i$ and $\theta'_i$ distributed uniformly over $\{-1,1\}$, independent of each other and independent for different $i$s. Then, for any symmetric matrix $H$*

$$\ln \mathbb{E}_{\theta\theta'}\left[e^{\lambda\theta^T H\theta'}\right] \le \lambda^2 \|H\|_F^2, \quad \forall\, \lambda > 0,$$

*where $\|\cdot\|_F$ denotes the Frobenius norm.*

Before we prove the claim, we use it to complete our proof. Combining this claim with our foregoing bound, we obtain

$$\mathbb{E}\Big[\mathrm{d_{TV}}\Big(\mathbb{E}_\theta\Big[\mathbf{R}^\theta\Big], \mathbf{R}^{\mathbf{u}}\Big)^2\Big] \le \frac{1}{4}\left(e^{\frac{n^2\varepsilon^4}{k^2}\|\bar{H}\|_F^2} - 1\right) \le \frac{1}{4}\left(e^{\frac{n^2\varepsilon^4}{k^2}\frac{1}{n}\sum_{j=1}^{n}\|H_j\|_F^2} - 1\right),$$

where in the previous inequality we used the convexity of squared-norm. To complete the proof, we show now that for every $j \in [n]$, $\|H_j\|_F^2 \le 2^\ell$. This is where we need to use the assumption that the protocol is deterministic. Specifically, for every pair $m, m' \in \{0,1\}^\ell$, let

$$S_{m,m'} := \{\ i \in [k/2] :\ W_j(m|2i-1) = W_j(m'|2i) = 1\ \} \cup \{\ i \in [k/2] :\ W_j(m|2i) = W_j(m'|2i-1) = 1\ \}$$

and

$$S_m := \{\ i \in [k/2] :\ W_j(m|2i-1) = 1\ \} \cup \{\ i \in [k/2] :\ W_j(m|2i) = 1\ \} = \bigcup_{m' \in \{0,1\}^\ell} S_{m,m'}\,.$$

It is then immediate to see that the $S_m$'s are disjoint and that

$$\|H_j\|_F^2 \le \sum_{m,m'} \frac{|S_{m,m'}|^2}{|S_m|\,|S_{m'}|} \le \sum_{m,m'} \frac{|S_{m,m'}|}{|S_m|} = \sum_m \frac{\sum_{m'}|S_{m,m'}|}{|S_m|} = 2^\ell\,,$$

whereby

$$\mathbb{E}\Big[\mathrm{d_{TV}}\Big(\mathbb{E}_\theta\Big[\mathbf{R}^\theta\Big], \mathbf{R}^{\mathbf{u}}\Big)^2\Big] \le \frac{1}{4}\left(e^{\frac{n^2\varepsilon^4 2^\ell}{k^2}} - 1\right).$$

The proof of the theorem can now be completed as the proof for $\ell = 1$ by first noting that the same bound holds for the total variation distance even with public randomness, since we have a uniform bound for each fixed realization of public randomness, and taking recourse to the standard two-point argument.

It only remains to establish Claim 6.14. To that end, we use the following bound which can be obtained by combining the transportation lemma with Marton's transportation-cost inequality (cf. [BLM13, Chapter 8]).

**Lemma 6.15.** *Consider independent random variables $X = (X_1, \ldots, X_n)$ and a function $f$ such that for every $x, y$*

$$f(x) - f(y) \leq \sum_{i=1}^{n} c_i(x) \mathbb{1}_{\{x_i \neq y_i\}} .$$

*Then, setting $v := \sum_{i=1}^{n} \mathbb{E}\big[c_i^2(X)\big]$, we have, for every $\lambda > 0$, $\ln \mathbb{E}\Big[e^{\lambda f(X)}\Big] \leq \frac{\lambda^2 v}{2}$.*

We apply this lemma to $f(Z, Z') = Z^T H Z'$, where $H$ is a symmetric matrix and $Z, Z'$ are independent copies of $\{-1, +1\}^n$-valued i.i.d. Rademacher vectors. In this case,

$$v = 2 \sum_{i=1}^{n} \mathbb{E}\left[ \left( Z_i \sum_{j=1}^{n} H_{ij} \right)^2 \right] = 2\|H\|_F^2,$$

which completes the proof of the claim and thereby that of Theorem 6.2. □

**Acknowledgments.** 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.

## References


[AC86] Rudolf Ahlswede and Imre Csiszár. Hypothesis testing with communication constraints. *IEEE Transactions on Information Theory*, 32(4):533–542, July 1986. 1, 1.3

[ACFT18] Jayadev Acharya, Clément L. Canonne, Cody Freitag, and Himanshu Tyagi. Test without Trust: Optimal Locally Private Distribution Testing. *CoRR*, abs/1808.02174, 2018. 6.1.1, C, C, C

[ADOS17] Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 11–21. PMLR, 2017. 1.3

[AOST17] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating Renyi Entropy of Discrete Distributions. *IEEE Trans. Information Theory*, 63(1):38–56, 2017. 1.3

[ASZ18] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Communication efficient, sample optimal, linear time locally private discrete distribution estimation. *CoRR*, abs/1802.04705, 2018. 1.3

[BBFM12] Maria-Florina Balcan, Avrim Blum, Shai Fine, and Yishay Mansour. Distributed learning, communication complexity and privacy. In *Proceedings of COLT*, volume 23 of *JMLR Proceedings*, pages 26.1–26.22. JMLR.org, 2012. 1, 1.3

[BCG17] Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. In *Computational Complexity Conference*, volume 79 of *LIPIcs*, pages 28:1–28:40. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017. 1.3, A, A, A

[BFR+00] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of FOCS*, pages 189–197, 2000. 1.3

[BGH+16] Mohammad Bavarian, Badih Ghazi, Elad Haramaty, Pritish Kamath, Ronald L. Rivest, and Madhu Sudan. The optimality of correlated sampling. *arXiv preprint arXiv:1612.01041*, 2016. 1.3

[BGM+16] Mark Braverman, Ankit Garg, Tengyu Ma, Huy L. Nguyen, and David P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of STOC*, pages 1011–1020. ACM, 2016. 1.3, 5.2

[BKR04] Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of STOC*, pages 381–390, New York, NY, USA, 2004. ACM. 1.3

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. 6.2

[BPC+11] Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 1, 1.3

[Bro97] Andrei Z. Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings*, pages 21–29. IEEE, 1997. 1.3

[BW18] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *Ann. Appl. Stat.*, 12(2):727–749, 2018. 1.3

[Can15] Clément L. Canonne. A Survey on Distribution Testing: your data is Big. But is it Blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, April 2015. 1, 1.3

[CDGR17] Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, pages 1–59, 2017. 1.3, 6.1.1

[CDVV14] Siu-on Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of SODA*, pages 1193–1203, 2014. 6.1.1

[DGL+17] Ilias Diakonikolas, Elena Grigorescu, Jerry Li, Abhiram Natarajan, Krzysztof Onak, and Ludwig Schmidt. Communication-efficient distributed learning of discrete distributions. In *Proceedings of NIPS*, pages 6394–6404, 2017. 1, 1.1, 1.3, 5.2

[DGPP17] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. *Electronic Colloquium on Computational Complexity (ECCC)*, 24:133, 2017. 1.3, 5.1

[Dia16] Ilias Diakonikolas. Learning structured distributions. In *Handbook of Big Data*. CRC Press, 2016. 1.3

[DJW13] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *Proceedings of FOCS*, pages 429–438. IEEE Computer Society, 2013. 1.3

[DKN15] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of SODA*, 2015. 1.3

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer New York, 2001. 1.3

[DMN18] Anindya De, Elchanan Mossel, and Joe Neeman. Non interactive simulation of correlated distributions is decidable. In *Proceedings of SODA*, pages 2728–2746. SIAM, 2018. 1.3

[GGR98] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, July 1998. 1.3

[GK73] Peter Gács and János Körner. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2(2):149–162, 1973. 1.3

[GKS16] Badih Ghazi, Pritish Kamath, and Madhu Sudan. Decidability of non-interactive simulation of joint distributions. In *Proceedings of FOCS*, pages 545–554. IEEE Computer Society, 2016. 1.3

[GMN14] Ankit Garg, Tengyu Ma, and Huy L. Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *NIPS*, pages 2726–2734, 2014. 1.3

[Gol14] Oded Goldreich. On Multiple Input Problems in Property Testing. In Klaus Jansen, José D. P. Rolim, Nikhil R. Devanur, and Cristopher Moore, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, volume 28 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 704–720, Dagstuhl, Germany, 2014. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. 1.2, 6.1.2, 6.7

[Gol16] Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:15, 2016. 1.3, A, A, 17

[Gol17] Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017. 1, 6.1.2

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity (ECCC), 2000. 1.3

[HA98] Te Sun Han and Shun-ichi Amari. Statistical inference under multiterminal data compression. *IEEE Transactions on Information Theory*, 44(6):2300–2324, October 1998. 1, 1.3

[Han87] Te Sun Han. Hypothesis testing with multiterminal data compression. *IEEE Transactions on Information Theory*, 33(6):759–772, November 1987. 1, 1.3

[HMÖW18] Yanjun Han, Pritam Mukherjee, Ayfer Özgür, and Tsachy Weissman. Distributed statistical estimation of high-dimensional and nonparametric distributions with communication constraints, February 2018. Talk given at ITA 2018. 1.3, 4, 5.5

[Hol07] Thomas Holenstein. Parallel repetition: simplifications and the no-signaling case. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 411–419. ACM, 2007. 1.3

[HÖW18a] Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Proceedings of COLT*, volume 75 of *Proceedings of Machine Learning Research*, pages 3163–3188. PMLR, 2018. 1, 1.1, 1.3, 5.5, 6, 5.2

[HÖW18b] Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric Lower Bounds for Distributed Parameter Estimation under Communication Constraints. *ArXiv e-prints*, February 2018. First version (https://arxiv.org/abs/1802.08417v1). B, B.2, B

[JVHW17] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Trans. Information Theory*, 63(10):6774–6798, 2017. 1.3

[JVYW15] Jiantao Jiao, Kartik Venkat, Han Yanjun, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, May 2015. 1.3

[KA12] Sudeep Kamath and Venkat Anantharam. Non-interactive simulation of joint distributions: The Hirschfeld-Gebelein-Rényi maximal correlation and the hypercontractivity ribbon. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1057–1064. IEEE, 2012. 1.3

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, New York, NY, USA, 1997. 3

[KT02] Jon Kleinberg and Eva Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. *Journal of the ACM (JACM)*, 49(5):616–639, 2002. 1.3

[Pan04] Liam Paninski. Estimating entropy on $m$ bins given fewer than $m$ samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004. 1.3

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. 1.2, 1.3, 5.1, 15, 6.2

[Pol03] David Pollard. Asymptopia, 2003. Manuscript. 1.2, 6.2, 6.2

[RRSS09] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distributions support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009. 1.3

[RS96] Ronitt Rubinfeld and Madhu Sudan. Robust characterization of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996. 1.3

[Rub12] Ronitt Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24, sep 2012. 1, 1.3

[Sha14] Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems*, pages 163–171, 2014. 1, 1.3

[ST18] K R Sahasranand and Himanshu Tyagi. Extra samples can reduce communication for independence testing. In *ISIT*. IEEE, 2018. 1.3

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. 1.3

[VV10a] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:179, 2010. 6.2

[VV10b] Gregory Valiant and Paul Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:180, 2010. 6.2

[VV11] Gregory Valiant and Paul Valiant. The power of linear estimators. In *Proceedings of FOCS*, pages 403–412, October 2011. See also [VV10a] and [VV10b]. 1.3

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017. 5.1, 6, A, A

[Wat17] Shun Watanabe. Neyman-Pearson test for zero-rate multiterminal hypothesis testing. *Proc. IEEE International Symposium on Information Theory*, pages 2157–8117, 2017. 1.3

[Wat18] Thomas Watson. Communication complexity of statistical distance. *TOCT*, 10(1):2:1–2:11, 2018. 1.3

[WT16] Michele Wigger and Roy Timo. Testing against independence with multiple decision centers. *IEEE International Conference on Signal Processing and Communications, IISc, Bangalore*, June 2016. 1.3

[WY16] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, June 2016. 1.3

[Wyn75] Aaron Wyner. The common information of two dependent random variables. *IEEE Transactions on Information Theory*, 21(2):163–179, 1975. 1.3

[XK13] Yu Xiang and Young Han Kim. Interactive hypothesis testing against independence. *Proc. IEEE International Symposium on Information Theory*, pages 1782–1786, 2013. 1.3

[XR17] Aolin Xu and Maxim Raginsky. Information-theoretic lower bounds on Bayes risk in decentralized estimation. *IEEE Transactions on Information Theory*, 63(3):1580–1600, 2017. 1.3

[ZDJW13] Yuchen Zhang, John Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013. 1.3

# 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[16] ("instance-optimal" in the vocabulary of Valiant and Valiant, who introduced it [VV17]). These two results will be seen as a straightforward consequence of [Gol16], which establishes the former reduction in the standard non-distributed setting; and of [BCG17], 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 $\ell$-bit protocol $\pi$ for testing uniformity of $k$-ary distributions, with number of players $n(k,\ell,\varepsilon)$ and failure probability $1/3$. Then there exists an $\ell$-bit protocol $\pi'$ 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 [Gol16], 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 $\mathbf{p} \in \Delta([k])$ $\varepsilon$-far from $\mathbf{q}$.[17] 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, as outlined in Section 5.1.

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 [VV17] 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

[16] 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 $\mathbf{q}$.

[17] In [Gol16], Goldreich exhibits a randomized mapping that converts the problem from testing identity over domain of size $k$ with proximity parameter $\varepsilon$ 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 $\alpha = 1/5$.

in [VV17] 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 [BCG17] 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 [BCG17, Section 6] 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 [VV17]. 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

$$\Phi(\mathbf{p}, \gamma) := 2\kappa_{\mathbf{p}}^{-1}(1-\gamma)^2 \,. \tag{12}$$

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

*Further, this reduction preserves the setting of randomness (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 [BCG17], which at a high-level reduces testing identity to $\mathbf{q}$ to three tasks: (i) computing the $(\varepsilon/3)$-effective support[18] 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 $O(n(|S_{\mathbf{q}}(\varepsilon)|, \ell, \frac{\varepsilon}{18}))$ 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 [BCG17, Section 7.2], we have $|S_{\mathbf{q}}(\varepsilon)| \leq \Phi(\mathbf{q}, \frac{\varepsilon}{9})$, 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 $n(\Phi(\mathbf{q}, \frac{\varepsilon}{9}), \ell, \frac{\varepsilon}{18})) + O(1/\varepsilon) = O(n(\Phi(\mathbf{q}, \frac{\varepsilon}{9}), \ell, \frac{\varepsilon}{18}))$. □

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

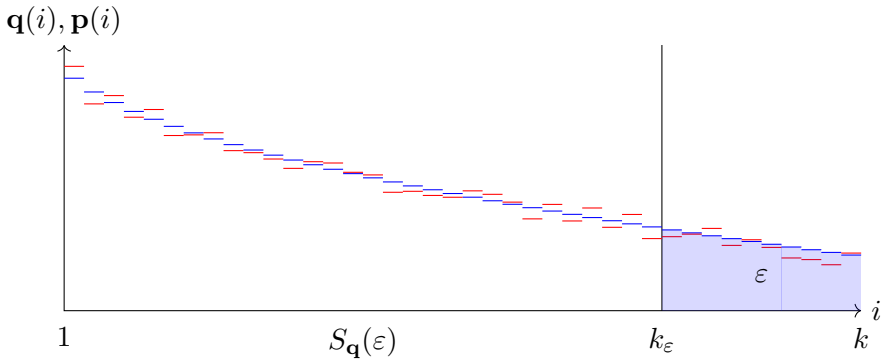


Figure 1: The reference distribution $\mathbf{q}$ (in blue; assumed non-increasing without loss of generality) and the unknown distribution $\mathbf{p}$ (in red). By the reduction above, testing equality of $\mathbf{p}$ to $\mathbf{q}$ is tantamount to (i) determining $S_{\mathbf{q}}(\varepsilon)$, which 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 Distributed learning lower bound (for public-randomness adaptive protocols)

**Theorem B.1.** *For $1 \le \ell \le \log k$, any $\ell$-bit public-coin (possibly adaptive) $(k, \varepsilon, 1/3)$-learning protocol must have $n = \Omega\left(\frac{k^2}{2^\ell \varepsilon^2}\right)$ players.*

*Proof.* We will show that the $\ell_1$ minimax rate is

$$\inf_{(W,\delta)} \inf_{\hat{\mathbf{p}}} \sup_{\mathbf{p} \in \Delta([k])} \mathbb{E}_{\mathbf{p}} \|\hat{\mathbf{p}} - \mathbf{p}\|_1 \ge C \cdot \left( \frac{k}{\sqrt{n(2^\ell \wedge k)}} \wedge 1 \right) = C \cdot \left( \sqrt{\frac{k}{n}} \vee \frac{k}{\sqrt{n 2^\ell}} \wedge 1 \right)$$

for some absolute constant $C > 0$, which implies the result. Note that since the collocated model can simulate the distributed one, the term $\Omega(\sqrt{k/n})$ in the lower bound, which dominates when $2^\ell \ge k$, is an immediate consequence of the standard lower bound in the collocated case (and holds without restriction on the range of $n$). Thus, it suffices to focus on the remaining case $2^\ell < k$.

In the argument below, we fix the realization shared randomness and restrict to deterministic protocols. Our proof of lower bound uses standard argument to relate minimax risk to probability of error in multiple hypothesis testing problem with uniform prior on the hypotheses. Since for every public-coin protocol we must have a deterministic protocol with same probability of error for the multiple hypothesis testing problem, there is no loss in restricting to deterministic protocols.

First, we establish the rate $\Omega\left(k/\sqrt{n2^\ell}\right)$, assuming $n \ge k^2/2^\ell$. (Recall that for $n \le k^2/2^\ell$, the lower bound in the RHS above is 1.) We follow the proof of Han, Özgür, and Weissman [HÖW18b, Proposition 1], with the necessary modifications to adapt it to $\ell_1$ loss (instead of squared $\ell_2$) and remove the constraint that $n \ge k^2/2^\ell$. (As in their proof, assume without loss of generality that $k$ is even.)

To handle the dependences between the $2^\ell$ outputted by any given player, we consider the Poissonized observation model, where we instead of $n$ players sending a message $Y_j$ in $\{0,1\}^\ell$ we have $n$ players sending each a message $\tilde{Y}_j$ in $\mathbb{N}^\ell$, where each bit of the message is a (conditionally) independent Poisson random variable: $\tilde{Y}^n = (\tilde{Y}_1, \dots, \tilde{Y}_n) \in (\mathbb{N}^\ell)^n$, with

$$\forall j \in [n], \forall m \in [2^\ell], \quad \tilde{Y}_{j,i} \mid b^{j-1} \sim \mathrm{Poisson}\Big(\Pr\Big[\, Y_j = m \;\Big|\; X, b^{j-1} \,\Big]\Big)$$

where, for $j \in [n]$, $b^j = (b_1, \dots, b_j) \in \{0,1\}^j$ is the ("side information") tuple of bits with $b_j := \mathbb{1}_{\left\{\sum_{m=1}^{2^\ell} \tilde{Y}_{j,m}=1\right\}}$; and for each $j \in [n]$ $(\tilde{Y}_{j,1}, \dots, \tilde{Y}_{j,2^\ell})$ are independent conditioned on $b^{j-1}$. In other terms, we replace the $[2^\ell]$-valued message of player $j$ by $2^\ell$ different Poisson random variables, each with the right expectation (and, for technical reasons, with side information about the messages sent some other players). As established in Lemma 1 of [HÖW18b], for distribution estimation a lower bound on the Poissonized model implies the same lower bound (up to constant factors) for our original setting.

In order to prove the lower bound, we define the family of hard instances (which will be random small perturbation of the uniform distribution $\mathbf{u}_k$). Letting $U$ be uniformly distributed in the hypercube $\{-1,1\}^t$ (where $t := \frac{k}{2}$), we choose $\gamma \in [0,1]$ (suitably set later in the proof) and let $\mathbf{p}_U \in \Delta([k])$ be defined by its probability mass function

$$\mathbf{p}_U = \frac{1}{k}\left(1 + \gamma U_1, \dots, 1 + \gamma U_t, 1 - \gamma U_1, \dots, , 1 - \gamma U_t\right) .$$

This defines a class $\mathcal{C} \subseteq \Delta([k])$ of $2^t$ distributions. Since clearly the $\ell_1$ minimax risk over *all* $k$-ary distributions is no less than that over $\mathcal{C}$, it suffices to lower bound the later. We will rely on the following lemma to first bound the mutual information between the tuple of Poissonized messages $\tilde{Y}^n$ and the unknown parameter $U$ to estimate:

**Lemma B.2** ([HÖW18b, Lemma 3])**.** *The following upper bound holds:*

$$I\left(U; \tilde{Y}^n\right) \le 2 \sum_{j=1}^{n} \sum_{m=1}^{2^\ell} \mathbb{E}_{U,U'} \left[ \frac{(\Pr_{\mathbf{p}_U}[Y_j = m \mid X_j, b^{j-1}] - \Pr_{\mathbf{p}_{U'}}[Y_j = m \mid X_j, b^{j-1}])^2}{\mathbb{E}_U \Pr_{\mathbf{p}_U}[Y_j = m \mid X_j, b^{j-1}]} \right]$$

*where $U'$ is an independent copy of $U$.*

To handle the right-hand-side of the above bound, observe that any randomized strategy $W\colon [k] \to \{0,1\}$ can be identified with a vector $w \in [0,1]^k$. For every such $w$, we have

$$\begin{aligned} \mathbb{E}_{U,U'} \frac{(\mathbb{E}_{\mathbf{p}_U} W(Y \mid X) - \mathbb{E}_{\mathbf{p}'_U} W(Y \mid X))^2}{\mathbb{E}_U \mathbb{E}_{\mathbf{p}_U} W(Y \mid X)} &= k \frac{w^T \mathbb{E}_{U,U'}[(\mathbf{p}_U - \mathbf{p}_{U'})(\mathbf{p}_U - \mathbf{p}_{U'})^T] w}{w^T \mathbf{1}} \\ &\le k \frac{4\gamma^2}{k^2} \cdot \frac{w^T w}{w^T \mathbf{1}} \le \frac{4\gamma^2}{k} \end{aligned} \tag{13}$$

the last step since $\|w\|_\infty \le 1$. We will use this later on, after relating this mutual information $I\left(U; \tilde{Y}^n\right)$ to the quantity we are trying to analyze, the $\ell_1$ minimax risk over our class $\mathcal{C}$ – which we do next. It is not hard to show, via a standard "Assouad's Lemma"-type argument that this $\ell_1$ minimax risk can be lower bounded as

$$\inf_{\hat{\mathbf{p}}} \sup_{\mathbf{p} \in \mathcal{C}} \mathbb{E}_{\mathbf{p}} \|\hat{\mathbf{p}} - \mathbf{p}\|_1 \ge c \cdot \gamma \inf_{\hat{U}} \Pr\Big[\, \mathrm{dist}\Big(\hat{U}, U\Big) \ge t/5 \,\Big] \tag{14}$$

where $\mathrm{dist}(\cdot,\cdot)$ is the unnormalized Hamming distance and $U$ is a uniform random vector in $\{-1,1\}^t$ and $c>0$ is an absolute constant. (This is another part where we depart from the argument of Han, Özgür, and Weissman, concerned with the squared $\ell_2$ loss.) Invoking Lemma B.2, along with (13) and the same distance-based Fano's inequality as in [HÖW18b, Lemma 2], we can conclude that

$$\inf_{\hat{U}} \Pr\Big[\,\mathrm{dist}\Big(\hat{U},U\Big) \geq t/5\,\Big] \geq 1-\frac{I\left(U;Y^n\right)+\ln 2}{t/8} \geq 1-\frac{2\cdot n2^\ell\cdot\frac{4\gamma^2}{k}+\ln 2}{t/8} = 1-16\frac{8\gamma^2 n2^\ell + k\ln 2}{k^2}\,.$$

The RHS will be at least say $1/2$, for large enough $k$, by setting $\gamma^2 := c'\cdot\frac{k^2}{n2^\ell}$ for a constant $c'>0$ sufficiently small (but independent of $k,n,\ell$). For this choice of $\gamma$, (14) becomes

$$\inf_{\hat{\mathbf{p}}}\sup_{\mathbf{p}\in\mathcal{C}} \mathbb{E}_{\mathbf{p}}\|\hat{\mathbf{p}}-\mathbf{p}\|_1 \geq \frac{c}{2}\gamma = C\cdot\frac{k}{\sqrt{n2^\ell}}$$

(where $C := \frac{c\cdot\sqrt{c'}}{2} > 0$), concluding the proof. (Note that the constraint $n\geq k/2^\ell$ was used in the setting of $\gamma^2$, to ensure that $\gamma\in[0,1]$.)

Finally, we are left with the case $n\leq k^2/2^\ell$, where we must show that the rate is $\Omega(1)$. We can prove it by reducing it to the previous case: namely, divide the domain $[k]$ into $k' := \sqrt{2^\ell n} < k$ disjoint intervals of equal size (assuming for simplicity, and with little loss of generality, that $k'$ divides $k$). Apply now the previous construction to the induced domain over $k'$ elements, setting the distribution $\mathbf{p}_U$ to be uniform on each of the $k'$ intervals. This leads to the setting of $\gamma^2 = \frac{k'^2}{n2^\ell}\in[0,1]$, and a lower bound on the risk of $\Omega(\gamma)=\Omega(1)$. □

# C Proof of Theorem 6.4

In this appendix, we prove Theorem 6.4, stating that taking a random balanced partition of the domain in $L\geq 2$ parts preserves the $\ell_2$ distance between distributions with constant probability. Note that, as mentioned in Section 6.1.1, the special case of $L=2$ was proven in [ACFT18]. In fact, the proof for general $L$ is similar to the proof in [ACFT18], 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 C.1** (Paley–Zygmund)**.** *Suppose $U$ is a non-negative random variable with finite variance. Then, for every $\theta\in[0,1]$,*

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

We will prove a more general version of Theorem 6.4, showing that the $\ell_2$ distance to any fixed distribution $\mathbf{q}\in\Delta([k])$ is preserved with a constant probability.[19] 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 $r\in[L]$, $\sum_{i=1}^k \mathbb{1}_{\{X_i=r\}} = \frac{k}{L}$.

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

**Theorem C.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*

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

*Proof of Theorem C.2.* As in [ACFT18, Theorem 14], 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 [ACFT18]. For ease of presentation, we have divided the proof into smaller results.

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

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

*Proof.* By linearity of expectation,

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

□

**Lemma C.4** (The $\ell_2^2$ distance to uniform of the flattening has the right expectation)**.** *For every $r \in [L]$,*

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

*In particular, the expected squared $\ell_2$ norm of $Z$ is*

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

*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

$$\operatorname{Var}[Z_r] = \mathbb{E}\left[Z_r^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{k} \delta_i \mathbb{1}_{\{X_i = r\}}\right)^2\right] = \sum_{1 \le i,j \le k} \delta_i \delta_j \mathbb{E}\left[\mathbb{1}_{\{X_i = r\}} \mathbb{1}_{\{X_j = r\}}\right]$$

$$= \sum_{i=1}^{k} \delta_i^2 \mathbb{E}\left[\mathbb{1}_{\{X_i = r\}}\right] + 2 \sum_{1 \le i < j \le k} \delta_i \delta_j \mathbb{E}\left[\mathbb{1}_{\{X_i = r\}} \mathbb{1}_{\{X_j = r\}}\right]\,.$$

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

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\{X_{k-1}=r\}}\mathbb{1}_{\{X_k=r\}}\right] &= \mathbb{E}\left[\mathbb{E}\left[\,\mathbb{1}_{\{X_{k-1}=r\}}\mathbb{1}_{\{X_k=r\}} \;\middle|\; \mathbb{1}_{\{X_k=r\}}\,\right]\right] = \frac{1}{L}\Pr[\,X_{k-1}=r \mid X_k=r\,] \\ &= \frac{1}{L}\Pr\left[\,X_{k-1}=r \;\middle|\; \sum_{i=1}^{k-1}\mathbb{1}_{\{X_i=r\}} = \frac{k}{L}-1\right] = \frac{1}{L^2}\cdot\frac{k-L}{k-1}, \end{aligned} \tag{15}$$

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:

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

□

**Lemma C.5** (The $\ell_2^2$ distance to uniform of the flattening has the required second moment)**.** *There exists an absolute constant $C>0$ such that*

$$\mathbb{E}\left[\|Z\|_2^4\right] \le C\|\delta\|_2^4.$$

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

$$\mathbb{E}\left[\|Z\|_2^4\right] = \mathbb{E}\left[\left(\sum_{r=1}^{L}Z_r^2\right)^2\right] = \sum_{r=1}^{L}\mathbb{E}\left[Z_r^4\right] + 2\sum_{r<r'}\mathbb{E}\left[Z_r^2Z_{r'}^2\right] \tag{16}$$

We will bound both terms separately. For the first term, we note that using [ACFT18, Equation(21)] with $\mathbb{1}_{\{X_i=r\}}$ in the role of $X_i$ there, each term $\mathbb{E}[Z_r^4]$ is bounded above by $19\|\delta\|_2^4/L$ whereby

$$\sum_{r=1}^{L}\mathbb{E}\left[Z_r^4\right] \le 19\|\delta\|_2^4. \tag{17}$$

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

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

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

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

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 $1/L$.

$$\begin{aligned}\mathbb{E}\Big[Z_r^2 Z_{r'}^2\Big] = &\sum_{|\{a,b,c,d\}|=2} \delta_a^2\delta_b^2 \mathbb{E}\Big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r'\}}\Big] \\ &+ \sum_{|\{a,b,c,d\}|=3} \delta_a^2\delta_b\delta_c \mathbb{E}\Big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r'\}}\mathbb{1}_{\{X_c=r'\}}\Big] \\ &+ \sum_{|\{a,b,c,d\}|=3} \delta_a\delta_b^2\delta_c \mathbb{E}\Big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r\}}\mathbb{1}_{\{X_c=r'\}}\Big] \\ &+ \sum_{|\{a,b,c,d\}|=4} \delta_a\delta_b\delta_c\delta_d \mathbb{E}\Big[\mathbb{1}_{\{X_a=r\}}\mathbb{1}_{\{X_b=r\}}\mathbb{1}_{\{X_c=r'\}}\mathbb{1}_{\{X_d=r'\}}\Big] . \end{aligned} \tag{18}$$

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

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

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

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

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

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

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

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

□

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 $\mathbb{E}\Big[\|Z\|_2^2\Big] \geq \frac{1}{2}\|\delta\|_2^2$ and $\mathbb{E}\Big[\|Z\|_2^4\Big] \leq C\|\delta\|_2^4$, for some absolute constant $C > 0$. Therefore, by the Payley–Zygmund inequality (Theorem C.1) applied to $\|Z\|_2^2$ for $\theta = 1/2$,

$$\Pr\Big[\|Z\|_2^2 > \frac{1}{4}\|\delta\|_2^2\Big] \geq \Pr\Big[\|Z\|_2^2 > \frac{1}{2}\mathbb{E}\Big[\|Z\|_2^2\Big]\Big] \geq \frac{1}{4}\frac{\mathbb{E}\Big[\|Z\|_2^2\Big]^2}{\mathbb{E}\Big[\|Z\|_2^4\Big]} \geq \frac{1}{16C} .$$

This concludes the proof. □



ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il