Distribution Testing Lower Bounds via Reductions from Communication Complexity

Eric Blais[1], Clément L. Canonne[2], and Tom Gur[3]

1   University of Waterloo, Waterloo, ON, Canada
    eric.blais@uwaterloo.ca
2   Columbia University, New York, NY, USA
    ccanonne@cs.columbia.edu
3   UC Berkeley, Berkeley, CA, USA
    tom.gur@berkeley.edu

Abstract

We present a new methodology for proving distribution testing lower bounds, establishing a connection between distribution testing and the simultaneous message passing (SMP) communication model. Extending the framework of Blais, Brody, and Matulef [15], we show a simple way to reduce (private-coin) SMP problems to distribution testing problems. This method allows us to prove new distribution testing lower bounds, as well as to provide simple proofs of known lower bounds.

Our main result is concerned with testing identity to a specific distribution p, given as a parameter. In a recent and influential work, Valiant and Valiant [53] showed that the sample complexity of the aforementioned problem is closely related to the $\ell_{2/3}$-quasinorm of p. We obtain alternative bounds on the complexity of this problem in terms of an arguably more intuitive measure and using simpler proofs. More specifically, we prove that the sample complexity is essentially determined by a fundamental operator in the theory of interpolation of Banach spaces, known as Peetre's *K-functional*. We show that this quantity is closely related to the size of the effective support of p (loosely speaking, the number of supported elements that constitute the vast majority of the mass of p). This result, in turn, stems from an unexpected connection to functional analysis and refined concentration of measure inequalities, which arise naturally in our reduction.

1   Introduction

Distribution testing, as first explicitly introduced in [9], is a branch of property testing [48, 31] concerned with the study of sublinear algorithms for making approximate decisions regarding probability distributions over massive domains. These algorithms are granted access to independent samples from an unknown distribution and are required to *test* whether this distribution has a certain global property. That is, a tester for property Π of distributions over domain Ω receives a proximity parameter $\varepsilon > 0$ and is asked to determine whether a distribution p over Ω (denoted $p \in \Delta(\Omega)$) has the property Π or is ε-far (say, in ℓ_1-distance) from any distribution that has Π, using a small number of independent samples from p. The *sample complexity* of Π is then the minimal number of samples needed to test it. Throughout

COMPUTATIONAL COMPLEXITY CONFERENCE

the introduction, we fix ε to be small constant and refer to a tester with respect to proximity parameter ε as an ε-*tester*.

In recent years, distribution testing has been studied extensively. In a significant body of work spanning more than a decade [8, 41, 7, 42, 2, 14, 37, 53, 23, 56, 36, 13, 25], a myriad of properties has been investigated under this lens. Starting with [32, 10, 8], this includes the testing of symmetric properties [45, 54, 51, 52], of structured families [11, 35, 3, 17, 4, 20, 19], as well as testing under some assumption on the unknown instance [47, 24, 27, 26]. Tight upper and lower bounds on the sample complexity have been obtained for properties such as uniformity, identity to a specified distribution, monotonicity, and many more (see references above, or [46, 18] for surveys). However, while by now numerous techniques and approaches are available to design distribution testers, our arsenal of tools for proving lower bounds on the sample complexity of distribution testing is significantly more limited. There are only a handful of standard techniques to prove lower bounds; and indeed the vast majority of the lower bounds in the literature are shown via *Le Cam's two-point method* (also known as the "easy direction" of Yao's minimax principle) [57, 44]. In this method, one first defines two distributions $\mathcal{Y}$ and $\mathcal{N}$ over distributions that are respectively yes-instances (having the property) and no-instances (far from having the property). Then it remains to show that with high probability over the choice of the instance, every tester that can distinguish between $p^{\text{yes}} \sim \mathcal{Y}$ and $p^{\text{no}} \sim \mathcal{N}$ must use at least a certain number of samples. In view of this scarcity, there has been in recent years a trend towards trying to obtain more, or simpler to use, techniques [54, 25]; however, this state of affairs largely remains the same.

In this work, we reveal a connection between distribution testing and the simultaneous message passing (SMP) communication model, which in turn leads to a new methodology for proving distribution testing lower bounds. Recall that in a private-coin SMP protocol, Alice and Bob are given strings $x, y \in \{0, 1\}^k$ (respectively), and each of the players is allowed to send a message to a referee (which depends on the player's input and private randomness) who is then required to decide whether $f(x, y) = 1$ by only looking at the players' messages and flipping coins.

Extending the framework of Blais, Brody, and Matulef [15], we show a simple way of reducing (private-coin) SMP problems to distribution testing problems. This foregoing methodology allows us to prove new distribution testing lower bounds, as well as to provide simpler proofs of known lower bounds for problems such as testing uniformity, monotonicity, and k-modality (see Section 8).

Our main result is a characterization of the sample complexity of the distribution identity testing problem in terms of a key operator in the study of interpolation spaces, which arises naturally from our reduction and for which we are able to provide an intuitive interpretation. Recall that in this problem, the goal is to determine whether a distribution q over domain Ω (denoted $q \in \Delta(\Omega)$) is identical to a fixed distribution p; that is, given a full description of $p \in \Delta(\Omega)$, we ask how many independent samples from q are needed to decide whether $q = p$, or whether q is ε-far in ℓ_1-distance from p.[1]

In a recent and influential work, Valiant and Valiant [53] showed that the sample complexity of the foregoing question is closely related to the $\ell_{2/3}$-quasinorm of p, defined as $\|p\|_{2/3} = \left(\sum_{\omega \in \Omega} |p(\omega)|^{2/3} \right)^{3/2}$. That is, viewing a distribution $p \in \Delta(\Omega)$ as an $|\Omega|$-dimensional vector of probabilities, let $p_{-\varepsilon}^{-\max}$ be the vector obtained from p by zeroing

---

[1] Note that this is in fact a family of massively parameterized properties $\{\Pi_p\}_{p \in \Delta(\Omega)}$, where Π_p is the property of being identical to p. See [39] for an excellent survey concerning massively parameterized properties.

■ **Table 1** Summary of results. All the bounds are stated for constant proximity parameter ε.

Property	Our results	Previous bounds
Uniformity	$\tilde{\Omega}(\sqrt{n})$	$\Theta(\sqrt{n})$ [32, 42]
Identity to p	$\Omega(\kappa_p^{-1}(1-\varepsilon)), O(\kappa_p^{-1}(1-c\cdot\varepsilon))$	$\Omega\left(\|p_{\varepsilon}^{-\max}\|_{2/3}\right), O\left(\|p_{c'\cdot\varepsilon}^{-\max}\|_{2/3}\right)$ [53]
Monotonicity	$\tilde{\Omega}(\sqrt{n})$	$\Theta(\sqrt{n})$ [11, 4, 20]
k-modal	$\tilde{\Omega}(\sqrt{n})$	$\tilde{\Omega}(\max(\sqrt{n},k))$ [19]
Log-concavity, Monotone Hazard Rate	$\tilde{\Omega}(\sqrt{n})$	$\Theta(\sqrt{n})$ [4, 20]
Binomial, Poisson Binomial	$\tilde{\Omega}(n^{1/4})$	$\Theta(n^{1/4})$ ([3, 20]
Symmetric sparse support	$\tilde{\Omega}(\sqrt{n})$	
Junta distributions (PAIRCOND model)	$\Omega(k)$	

out the largest entry as well as the set of smallest entries summing to ε (note that $p_{-\varepsilon}^{\max}$ is no longer a probability distribution). Valiant and Valiant gave an ε-tester for testing identity to p with sample complexity $O\left(\|p_{-c\varepsilon}^{\max}\|_{2/3}\right)$, where $c > 0$ is a universal constant, and complemented this result with a lower bound of $\Omega\left(\|p_{-\varepsilon}^{\max}\|_{2/3}\right)$.[2]

In this work, using our new methodology, we show alternative and similarly tight bounds on the complexity of identity testing, in terms of a more intuitive measure (as we discuss below) and using simpler arguments. Specifically, we prove that the sample complexity is essentially determined by a fundamental quantity in the theory of interpolation of Banach spaces, known as Peetre's *K-functional*. Formally, for a distribution $p \in \Delta(\Omega)$, the K-functional between ℓ_1 and ℓ_2 spaces is the operator defined for $t > 0$ by

$$\kappa_p(t) = \inf_{p'+p''=p} \|p'\|_1 + t\|p''\|_2.$$

This operator can be thought of as an interpolation norm between the ℓ_1 and ℓ_2 norms of the distribution p (controlled by the parameter t), naturally inducing a partition of p into two sub-distributions: p', which consists of "heavy hitters" in ℓ_1-norm, and p'', which has a bounded ℓ_2-norm. Indeed, the approach of isolating elements with large mass and testing in ℓ_2-norm seems inherent to the problem of identity testing, and is the core component of both early works [32, 8] and more recent ones [27, 25, 30]. As a further connection to the identity testing question, we provide an easily interpretable proxy for this measure κ_p, showing that the K-functional between the ℓ_1 and ℓ_2 norms of the distribution p is closely related to the size of the effective support of p, which is the number of supported elements that constitute the vast majority of the mass of p; that is, we say that p has ε-*effective support* of size T if $1 - O(\varepsilon)$ of the mass of p is concentrated on T elements (see Section 2.4 for details).

Having defined the K-functional, we can proceed to state the lower bound we derive for the problem.[3]

---

[2] We remark that for certain p's, the asymptotic behavior of $O\left(\|p_{-c\varepsilon}^{\max}\|_{2/3}\right)$ strongly depends on the constant c, and so it cannot be omitted from the expression. We further remark that this result was referred to by Valiant and Valiant as "instance-optimal identity testing" as the resulting bounds are phrased as a function of the distribution p itself – instead of the standard parameter which is the domain size n.

[3] As stated, this result is a slight strengthening of our communication complexity reduction, which yields

▶ **Theorem 1** (Informally stated). *Any ε-tester of identity to $p \in \Delta(\Omega)$ must have sample complexity $\Omega\big(\kappa_p^{-1}(1 - 2\varepsilon)\big)$.*

In particular, straightforward calculations show that for the uniform distribution we obtain a tight lower bound of $\Omega(\sqrt{n})$, and for the Binomial distribution we obtain a tight lower bound of $\Omega\big(n^{1/4}\big)$.

　　To show that tightness of the lower bound above, we complement it with a nearly matching upper bound, also expressed in terms of the K-functional.

▶ **Theorem 2** (Informally stated). *There exist an absolute constant $c > 0$ and an ε-tester of identity to $p \in \Delta(\Omega)$ that uses $O\big(\kappa_p^{-1}(1 - c\varepsilon)\big)$ mples.*[4]

　　In the following section, we provide an overview of our new methodology as well as the proofs for the above theorems. We also further discuss the interpretability of the K-functional and show its close connection to the effective support size. We conclude this section by outlining a couple of extensions of our methodology.

Dealing with sub-constant values of the proximity parameter. Similarly to the communication complexity methodology for proving property testing lower bounds [15], our method inherently excels in the regime of *constant* values of the proximity parameter ε. Therefore, in this work we indeed focus on the constant proximity regime. However, in Section 5.1 we demonstrate how to obtain lower bounds that asymptotically increase as ε tends to zero, via an extension of our general reduction.

Extending the methodology to testing with conditional samples. Testers with sample access are by far the most commonly studied algorithms for distribution testing. However, many scenarios that arise both in theory and practice are not fully captured by this model. In a recent line of works [22, 21, 1, 28, 29], testers with access to *conditional* samples were considered, addressing situations in which one can control the samples that are obtained by requesting samples conditioned on membership on subsets of the domain. In Section 9, we give an example showing that it is possible to extend our methodology to obtain lower bounds in the conditional sampling model.

1.1 Organization

We first give a technical overview in Section 2, demonstrating the new methodology and presenting our bounds on identity testing. Section 3 then provides the required preliminaries for the main technical sections. In Section 4 we formally state and analyze the SMP reduction methodology for proving distribution testing lower bounds. In Section 5, we instantiate the basic reduction, obtaining a lower bound on uniformity testing, and in Section 5.1 show how to extend the methodology to deal with sub-constant values of the proximity parameter. (We stress that Section 5.1 is *not* a prerequisite for the rest of the sections, and can be skipped at the reader's convenience.) In Section 6 we provide an exposition to the K-functional and generalize inequalities that we shall need for the following sections. Section 7 then contains the proofs of both lower and upper bounds on the problem of identity testing, in terms of

---

a lower bound of $\Omega\big(\kappa_p^{-1}(1 - 2\varepsilon)/\log n\big)$. This strengthening is described in Section 7.3.

[4] Similarly to the [53] bound, for certain p's, the asymptotic behavior of $O\big(\kappa_p^{-1}(1 - 2\varepsilon)\big)$ depends on the constant c, and so it cannot be omitted from the expressi

the K-functional. In Section 8, we demonstrate how to easily obtain lower bounds for other distribution testing problems. Finally, in Section 9 we discuss extensions to our methodology; specifically, we explain how to obtain lower bounds in various metrics, and show a reduction from communication complexity to distribution testing in the conditional sampling model.

2   Technical Overview

In this section we provide an overview of the proof of our main result, which consists of new lower and upper bounds on the sample complexity of testing identity to a given distribution, expressed in terms of an intuitive, easily interpretable measure. To do so, we first introduce the key component of this proof, the methodology for proving lower bounds on distribution testing problems via reductions from SMP communication complexity. We then explain how the relation to the theory of interpolation spaces and the so-called K-functional naturally arises when applying this methodology to the identity testing problem.

For the sake of simplicity, throughout the overview we fix the domain $\Omega = [n]$ and fix the proximity parameter ε to be a small constant. We begin in Section 2.1 by describing a simple "vanilla" reduction for showing an $\tilde{\Omega}(\sqrt{n})$ lower bound on the complexity of testing that a distribution is uniform. Then, in Section 2.2 we extend the foregoing approach to obtain a new lower bound on the problem of testing identity to a fixed distribution. This lower bound depends on the best rate obtainable by a special type of error-correcting codes, which we call *p-weighted codes*. In Section 2.3, we show how to relate the construction of such codes to concentration of measure inequalities for weighted sums of Rademacher random variables; furthermore, we discuss how the use of the K-*functional*, an interpolation norm between ℓ_1 and ℓ_2 spaces, leads to stronger concentration inequalities than the ones derived by Chernoff bounds or the central limit theorem. Finally, in Section 2.4 we establish nearly matching upper bounds for testing distribution identity in terms of this K-functional, using a proxy known as the Q-norm. We then infer that the sample complexity of testing identity to a distribution p is roughly determined by the size of the *effective support* of p (which is, loosely speaking, the number of supported elements which together account for the vast majority of the mass of p).

2.1   Warmup: Uniformity Testing

Consider the problem of testing whether a distribution $q \in \Delta([n])$ is the *uniform distribution*; that is, how many (independent) samples from q are needed to decide whether q is the uniform distribution over $[n]$, or whether q is ε-far in ℓ_1-distance from it. We reduce the SMP communication complexity problem of *equality* to the distribution testing problem of uniformity testing.

Recall that in a private-coin SMP protocol for equality, Alice and Bob are given strings $x, y \in \{0, 1\}^k$ (respectively), and each of the players is allowed to send a message to a referee (which depends on the player's input and private randomness) who is then required to decide whether $x = y$ by only looking at the players' messages and flipping coins.

The reduction is as follows. Assume there exists a uniformity tester with sample complexity s. Each of the players encodes its input string via a balanced asymptotically good code C (that is, $C \colon \{0, 1\}^k \to \{0, 1\}^n$ is an error-correcting code with constant rate and relative distance $\delta = \Omega(1)$, which satisfies the property that each codeword of C contains the same number of 0's and 1's). Denote by $A \subset [n]$ the locations in which $C(x)$ takes the value 1 (i.e., $A = \{ i \in [n] : C(x)_i = 1 \}$), and denote by $B \subset [n]$ the locations in which $C(y)$ takes the value 0 (i.e., $B = \{ i \in [n] : C(y)_i = 0 \}$). Alice and Bob each send $O(s)$ uniformly
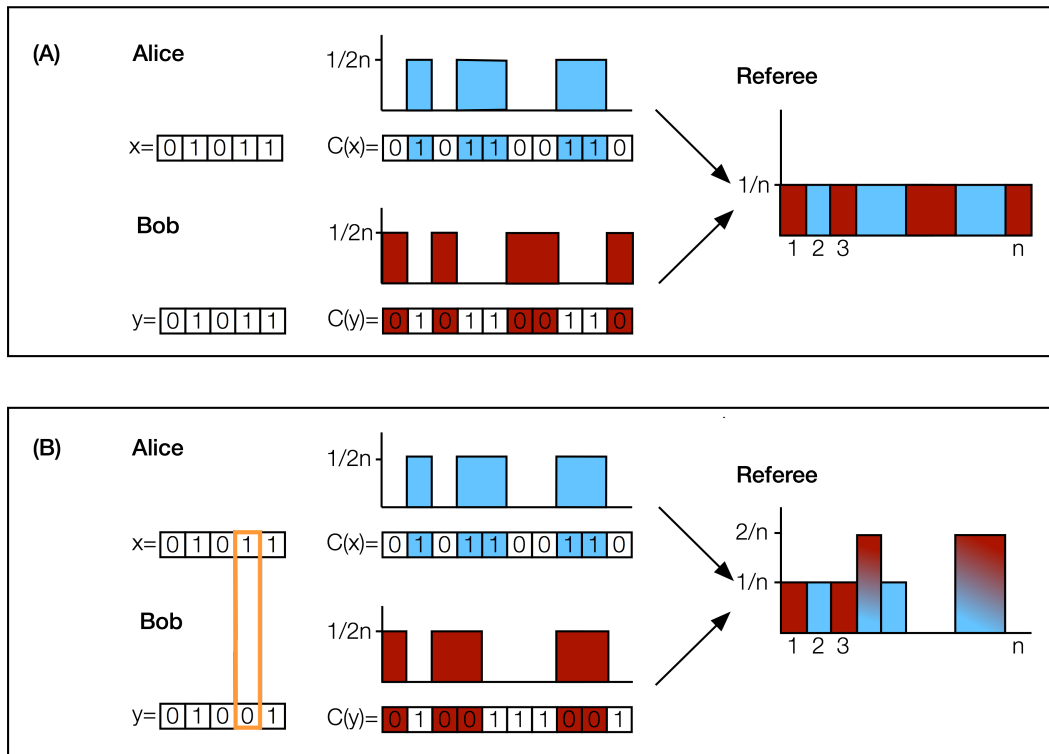
Figure 1 The reduction from equality in the SMP model to uniformity testing of distributions. In (A) we see that the uniform distribution is obtained when $x = y$, whereas in (B) we see that when $x \neq y$, we obtain a distribution that is "far" from uniform.

distributed samples from A and B, respectively. Finally, the referee invokes the uniformity tester with respect to the distribution $q = (\mathcal{U}_A + \mathcal{U}_B)/2$, emulating each draw from q by tossing a random coin and deciding accordingly whether to use a sample by Alice or Bob. See Figure 1.

The idea is that if $x = y$, then $C(x) = C(y)$, and so A and B are a *partition* of the set $[n]$. Furthermore, since $|C(x)| = |C(y)| = n/2$, this is a equipartition. Now, since Alice and Bob send uniform samples from an equipartition of $[n]$, the distribution q that the referee emulates is in fact the uniform distribution over $[n]$, and so the uniformity tester will accept. On the other hand, if $x \neq y$, then $C(x)$ and $C(y)$ disagree on a constant fraction of the domain. Thus, A and B intersect on $\delta/2$ elements, as well as do not cover $\delta/2$. Therefore q is uniform on a $(1 - \delta)$-fraction of the domain, unsupported on a $(\delta/2)$-fraction of the domain, and has "double" weight $2/n$ on the remaining $(\delta/2)$-fraction. In particular, since $\delta = \Omega(1)$, the emulated distribution q is $\Omega(1)$-far (in ℓ_1-distance) from uniform, and it will be rejected by the uniformity tester.

As each sample sent by either Alice or Bob was encoded with $O(\log n)$ bits, the above constitutes an SMP protocol for equality with communication complexity $O(s \log(n))$. Yet it is well known [40] that the players must communicate $\Omega(\sqrt{k})$ bits to solve this problem (see Section 4), and so we deduce that $s = \Omega(\sqrt{k}/\log(n)) = \tilde{\Omega}(\sqrt{n})$.

2.2   Revisiting Distribution Identity Testing: A New Lower Bound

Next, consider the problem of testing whether a distribution $q \in \Delta([n])$ is identical to a fixed distribution p, provided as a (massive) parameter; that is, given a full description of $p \in \Delta([n])$, we ask how many independent samples from q are needed to decide whether $q = p$, or whether q is ε-far in ℓ_1-distance from p. As mentioned earlier, Valiant and Valiant [53] established both upper and lower bounds on this problem, involving the $\ell_{2/3}$-quasinorm of p. We revisit this question, and show different – and more interpretable – upper and lower bounds. First, by applying our new communication complexity methodology to the distribution identity problem, we obtain a simple lower bound expressed in terms of a new parameter, which is closely related to the *effective support size* of p.

Consider any fixed $p \in \Delta([n])$. As a first idea, it is tempting to reduce equality in the SMP model to testing identity to p by following the uniformity reduction described in Section 2.1, only instead of having Alice and Bob send *uniform* samples from A and B, respectively, we have them send samples from p *conditioned* on membership in A and B respectively. That is, as before Alice and Bob encode their inputs x and y via a balanced, asymptotically good code C to obtain the sets $A = \{ i \in [n] : C(x)_i = 1 \}$ and $B = \{ i \in [n] : C(y)_i = 0 \}$, which partition $[n]$ if $x = y$, and intersect on $\Omega(n)$ elements (as well as fail to cover $\Omega(n)$ elements of $[n]$) if $x \neq y$. Only now, Alice sends samples independently drawn from $p|_A$, i.e., p conditioned on the samples belonging to A, and Bob sends samples independently drawn from $p|_B$, i.e., p conditioned on the samples belonging to B; and the referee emulates the distribution $q = (p|_A + p|_B)/2$.

However, two problems arise in the foregoing approach. The first is that while indeed when $x = y$ the reduction induces an equipartition A, B of the domain, the resulting weights $p(A)$ and $p(B)$ in the mixture may still be dramatically different, in which case the referee will need much more samples from one of the parties to emulate p. The second is a bit more subtle, and has to do with the fact that the properties of this partitioning are with respect to the *size* of the symmetric difference $A\Delta B$, while really we are concerned about its *mass* under the emulated distribution q (and although both are proportional to each other in the case of the uniform distribution, for general p we have no such guarantee). Namely, when $x \neq y$ the domain elements which are responsible for the distance from p (that is, the elements which are covered by both parties ($A \cap B$) and by neither of the parties ($[n] \setminus (A \cup B)$)) may only have a small mass according to p, and thus the emulated distribution q will not be sufficiently far from p. A natural attempt to address these two problems would be to preprocess p by discarding its light elements, focusing only on the part of the domain where p puts enough mass pointwise; yet this approach can also be shown to fail, as in this case the reduction may still not generate enough distance.[5]

Instead, we take a different route. The key idea is to consider a new type of codes, which we call *p-weighted codes*, which will allow us to circumvent the second obstacle. These are code whose distance guarantee is weighted according to the distribution p; that is, instead of requiring that every two codewords c, c' in a code C satisfy $\text{dist}(x, y) \stackrel{\text{def}}{=} \sum_{i=1}^{n} |x_i - y_i| \geq \delta$,

---

[5] In more detail, this approach would consider the distribution p' obtained by iteratively removing the lightest elements of p until a total of ε probability mass was removed. This way, every element i in the support of p' is guaranteed to have mass $p'_i \geq \varepsilon/n$: this implies that the weights $p'(A)$ and $p'(B)$ are proportional, and that each element that is either covered by both parties or not covered at all will contribute ε/n to the distance from p'. However, the total distance of q from p would only be $\Omega\big(|\operatorname{supp}(p')| \cdot \varepsilon/n\big)$; and this only suffices if p and p' have comparable support size, i.e. $|\operatorname{supp}(p)| = O\big(|\operatorname{supp}(p')|\big)$.

we consider a code $C_p \colon \{0,1\}^k \to \{0,1\}^n$ such that every $c, c' \in C_p$ satisfy

$$\operatorname{dist}_p(x, y) \overset{\text{def}}{=} \sum_{i=1}^{n} p(i) \cdot |x_i - y_i| \geq \delta.$$

Furthermore, to handle the first issue, we adapt the "balance" property accordingly, requiring that each codeword be balanced according to p, that is, every $c \in C_p$ satisfies $\sum_{i=1}^{n} p(i) \cdot c_i = 1/2$.

It is straightforward to see that if we invoke the above reduction while letting the parties encode their inputs via a balance p-weighted code C_p, then both of the aforementioned problems are resolved; that is, by the p-balance property the weights $p(A)$ and $p(B)$ are equal, and by the p-distance of C_p we obtain that for $x \neq y$ the distribution $q = (p_A + p_B)/2$ is $\Omega(1)$-far from p. Hence we obtain a lower bound of $\Omega\left(\sqrt{k}/\log(n)\right)$ on the query complexity of testing identity to p. To complete the argument, it remains to construct such codes, and determine what the best rate k/n that can be obtained by p-weighted codes is.

2.3 Detour: p-weighted Codes, Peetre's K-functional, and beating the CLT

The discussion of previous section left us with the task of constructing high-rate p-weighted codes. Note that unlike standard (uniformly weighted) codes, for which we can easily obtain constant rate, there exist some p's for which high rate is impossible (for example, if $p \in \Delta([n])$ is only supported on one element, we can only obtain rate $1/n$). In particular, by the sphere packing bound, every p-weighted code $C \colon \{0,1\}^k \to \{0,1\}^n$ with distance δ must satisfy

$$\underbrace{2^k}_{\#\text{codewords}} \leq \frac{2^n}{\operatorname{Vol}_{\mathbb{F}_2^n, \operatorname{dist}_p}(\delta/2)},$$

where $\operatorname{Vol}_{\mathbb{F}_2^n, \operatorname{dist}_p}(r)$ is the volume of the p-ball of radius r in the n-dimensional hypercube, given by

$$\operatorname{Vol}_{\mathbb{F}_2^n, \operatorname{dist}_p}(r) \overset{\text{def}}{=} \left| \left\{ w \in \mathbb{F}_2^n \ : \ \sum_{i=1}^{n} p_i \cdot w_i \leq r \right\} \right|.$$

Hence, we must have $k \leq n - \log \operatorname{Vol}_{\mathbb{F}_2^n, \operatorname{dist}_p}(\delta/2)$.

In Section 7.1 we show that there exist (roughly) balanced p-weighted codes with nearly-optimal rate,[6] and so it remains to determine the volume of the p-ball of radius ε in the n-dimensional hypercube, where recall that ε is the proximity parameter of the test. To this end, it will be convenient to represent this quantity as a concentration inequality of sums of weighted Rademacher random variables, as follows

$$\operatorname{Vol}_{\mathbb{F}_2^n, \operatorname{dist}_p}(\varepsilon) = 2^n \Pr_{Y \sim \{0,1\}^n} \left[\sum_{i=1}^{n} p_i Y_i \leq \varepsilon \right] = 2^n \Pr_{X \sim \{-1,1\}^n} \left[\sum_{i=1}^{n} p_i X_i \geq 1 - 2\varepsilon \right]. \quad (1)$$

Applying standard tail bounds derived from the central limit theorem (CLT), we have that

$$\Pr_{X \sim \{-1,1\}^n} \left[\sum_{i=1}^{n} p_i X_i \geq 1 - 2\varepsilon \right] \leq e^{\frac{-(1-2\varepsilon)^2}{2\|p\|_2^2}}, \quad (2)$$

---

[6] We remark that since these codes are not perfectly p-balanced, a minor modification to the reduction needs to be done. See Section 7.1 for details.

and so we can obtain a p-weighted code $C_p\colon \{0,1\}^k \to \{0,1\}^n$ with dimension $k = O(1/\|p\|_2^2)$, which in turn, by the reduction described in Section 2.2, implies a lower bound of $\Omega(1/(\|p\|_2 \cdot \log(n)))$ on the complexity of testing identity to p.

Unfortunately, the above lower bound is not as strong as hoped, and in particular, far weaker than the $\|p_{-\varepsilon}^{-\max}\|_{2/3}$ bound of [53].[7] Indeed, it turns out that the CLT-based bound in Equation 2 is only tight for distributions satisfying $\|p\|_\infty = O(\|p\|_2^2)$, and is in general too crude for our purposes. Instead, we look for stronger concentration of measure inequalities that "beat" the CLT. To this end, we shall use powerful tools from the theory of interpolation spaces. Specifically, we consider Peetre's *K-functional* between ℓ_1 and ℓ_2 spaces. Loosely speaking, this is the operator defined for $t > 0$ by

$$\kappa_p(t) = \inf_{p'+p''=p} \|p'\|_1 + t\|p''\|_2.^8$$

This K-functional can be thought of as an interpolation norm between the ℓ_1 and ℓ_2 norms of the distribution p (and accordingly, for any fixed t it defines a norm on the space $\ell_1 + \ell_2$). In particular, note that for large values of t the function $\kappa_p(t)$ is close to $\|p\|_1$, whereas for small values of t it will behave like $t\|p\|_2$.

The foregoing connection is due to Montgomery-Smith [38], who established the following concentration of measure inequality for weighted sums of Rademacher random variables,

$$\Pr\left[\sum_{i=1}^n p_i X_i \ge \kappa_p(t)\right] \le e^{-\frac{t^2}{2}}. \tag{3}$$

Furthermore, he proved that this concentration bound is essentially tight (see Section 6 for a precise statement). Plugging (3) into (1), we obtain a lower bound of $\Omega(\kappa_p^{-1}(1-2\varepsilon)/\log(n))$ on the complexity of testing identity to p.

To understand and complement this result, we describe in the next subsection a nearly tight upper bound for this problem, also expressed in terms of this K-functional; implying that this unexpected connection is in fact not a coincidence, but instead capturing an intrinsic aspect of the identity testing question. We also give a natural interpretation of this bound, showing that the size of the *effective support* of p (roughly, the number of supported elements that constitute the vast majority of the mass of p) is a good proxy for this parameter $\kappa_p^{-1}(1-2\varepsilon)$ – and thus for the complexity of testing identity to p.

2.4 Using the Q-norm Proxy to Obtain an Upper Bound

To the end of obtaining an upper bound on the sample complexity of testing identity to p, in terms of the K-functional, it will actually be convenient to look at a related quantity, known as the Q-*norm* [38]. At a high-level, the Q-norm of a distribution p, for a given parameter $T \in \mathbb{N}$, is the maximum one can reach by partitioning the domain of p into T sets and taking

---

[7] For example, fix $\alpha \in (0,1)$, and consider the distribution $p \in \Delta([n])$ in which $n/2$ elements are of mass $1/n$, and $n^\alpha/2$ elements are of mass $1/n^\alpha$. It is straightforward to verify that $\|p\|_2^{-1} = \Theta((\sqrt{n})^\alpha)$, whereas $\|p\|_{2/3} = \Theta(\sqrt{n})$. (Intuitively, this is because the ℓ_2-norm is mostly determined by the few heavy elements, whereas the $\ell_{2/3}$-quasinorm is mostly determined by the numerous light elements.)

[8] Interestingly, Holmstedt [34] showed that the infimum is *approximately* obtained by partitioning $p = (p', p'')$ such that p' consists of heaviest t^2 coordinates of p and p'' consists of the rest (for more detail, see Theorem 15).

the sum of the ℓ_2 norms of these T subvectors. That is

$$\|p\|_{Q(T)} \overset{\text{def}}{=} \sup \left\{ \sum_{j=1}^{T} \left(\sum_{i \in A_j} p_i^2 \right)^{1/2} \; : \; (A_j)_{1 \le j \le T} \text{ partition of } \mathbb{N} \right\}.$$

Astashkin [6], following up Montgomery-Smith [38], showed that the Q-norm constitutes a good approximation of K-functional, by proving that

$$\|p\|_{Q(t^2)} \le \kappa_p(t) \le \sqrt{2}\|p\|_{Q(t^2)}.$$

In Section 6 we further generalize this claim and show it is possible to get a tradeoff in the upper bound; specifically, we prove that $\kappa_p(t) \le \|p\|_{Q(2t^2)}$. Thus, it suffices to prove an upper bound on distribution identity testing in terms of the Q-norm.

From an algorithmic point of view, it is not immediately clear that switching to this Q-norm is of any help. However, we will argue that this value captures – in a very quantitative sense – the notion of the *sparsity* of p. As a first step, observe that if $\|p\|_{Q(T)} = 1$, then the distribution p is supported on at most T elements. To see this, denote by p_{A_j} the restriction of the sequence p to the indices in A_j, and note that if $\|p\|_{Q(T)} \overset{\text{def}}{=} \sum_{j=1}^{T} \|p_{A_j}\|_2 = 1$, then by the monotonicity of ℓ_p norms and since $\sum_{j=1}^{T} \|p_{A_j}\|_1 = \|p\|_1 = 1$ we have that

$$\sum_{j=1}^{T} (\underbrace{\|p_{A_j}\|_1 - \|p_{A_j}\|_2}_{\ge 0}) = 0,$$

which implies that $\|p_{A_j}\|_1 = \|p_{A_j}\|_2$ for all $j \in [T]$.

Now, it turns out that it is possible to obtain a *robust* version of the foregoing observation, yielding a sparsity lemma that, roughly speaking, shows thats if $\|p\|_{Q(T)} \ge 1 - \varepsilon$, then $1 - O(\varepsilon)$ of the mass of p is concentrated on T elements: in this case we say that p has $O(\varepsilon)$-*effective support* of size T. (See Lemma 31 for precise statement of the sparsity lemma.)

This property of the Q-norm suggests the following natural test for identity to a distribution p: Simply fix T such that $\|p\|_{Q(T)} = 1 - \varepsilon$, and apply one of the standard procedures for testing identity to a distribution with support size T, which require $O(\sqrt{T})$ samples. But by the previous discussion, we have $\|p\|_{Q(2t^2)} \ge \kappa_p(t)$, so that setting $T = 2t^2$ for the "right" choice of $t = \kappa_p^{-1}(1 - 2\varepsilon)$ will translate to an $O(t)$ upper bound – which is what we were aiming for.

3 Preliminaries

Notation.   We write $[n]$ for the (ordered) set of integers $\{1, \ldots, n\}$, and ln, log for respectively the natural and binary logarithms. We use the notation $\tilde{\Omega}(f)$ to hide polylogarithmic dependencies on the argument, i.e. for expressions of the form $\Omega(f \log^c f)$ (for some absolute constant c). All throughout the paper, we denote by $\Delta(\Omega)$ the set of discrete probability distributions over domain Ω. When the domain is a subset of the natural numbers $\mathbb{N}$, we shall identify a distribution $p \in \Delta(\Omega)$ with the sequence $(p_i)_{i \in \mathbb{N}} \in \ell_1$ corresponding to its probability mass function (pmf). For a subset $S \subseteq \Omega$, we denote by $p|_S$ the normalized projection of p to S (so $p|_S$ is a probability distribution).

For an alphabet Σ, we denote the projection of $x \in \Sigma^n$ to a subset of coordinates $I \subseteq [n]$ by $x|_I$. For $i \in [n]$, we write $x_i = x|_{\{i\}}$ to denote the projection to a singleton. We denote the *relative Hamming distance*, over alphabet Σ, between two strings $x \in \Sigma^n$ and $y \in \Sigma^n$

by $\mathrm{dist}(x,y) \stackrel{\mathrm{def}}{=} |\{\, x_i \neq y_i \,:\, i \in [n] \,\}|\,/n$. If $\mathrm{dist}(x,y) \leq \varepsilon$, we say that x is ε-*close* to y, and otherwise we say that x is ε-*far* from y. Similarly, we denote the *relative Hamming distance* of x from a non-empty set $S \subseteq \Sigma^n$ by $\mathrm{dist}(x,S) \stackrel{\mathrm{def}}{=} \min_{y \in S} \mathrm{dist}(x,y))$. If $\mathrm{dist}(x,S) \leq \varepsilon$, we say that x is ε-*close* to S, and otherwise we say that x is ε-*far* from S.

Distribution Testing. A *property* of distributions over Ω is a subset $\mathcal{P} \subseteq \Delta(\Omega)$, consisting of all distributions that have the property. Given two distributions $p, q \in \Delta(\Omega)$, the ℓ_1 distance between p and q is defined as the ℓ_1 distance between their pmf's, namely $\|p - q\|_1 = \sum_{i \in \Omega} |p_i - q_i|$.[9] Given a property $\mathcal{P} \subseteq \Delta(\Omega)$ and a distribution $p \in \Delta(\Omega)$, we then define the distance of p to $\mathcal{P}$ as $\ell_1(p, \mathcal{P}) = \inf_{q \in \mathcal{P}} \|p - q\|_1$.

A *testing algorithm* for a fixed property $\mathcal{P}$ is then a randomized algorithm $\mathcal{T}$ which takes as input n, $\varepsilon \in (0,1]$, and is granted access to independent samples from an unknown distribution p; and satisfies the following.
1. if $p \in \mathcal{P}$, the algorithm outputs accept with probability at least $2/3$;
2. if $\ell_1(p, \mathcal{P}) \geq \varepsilon$, it outputs reject with probability at least $2/3$.
In other words, $\mathcal{T}$ must accept with high probability if the unknown distribution has the property, and reject if it is ε-*far* from having it. The *sample complexity* of the algorithm is the number of samples it draws from the distribution in the worst case.

Inequalities. We now state a standard probabilistic result that some of our proofs will rely on, the Paley–Zygmund anticoncentration inequality:

▶ **Theorem 3** (Paley–Zygmund inequality). *Let X be a non-negative random variable with finite variance. Then, for any $\theta \in [0,1]$,*

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

We will also require the following version of the rearrangement inequality, due to Hardy and Littlewood (cf. for instance [12, Theorem 2.2]):

▶ **Theorem 4** (Hardy–Littlewood Inequality). *Fix any $f, g \colon \mathbb{R} \to [0, \infty)$ such that $\lim_{\pm\infty} f = \lim_{\pm\infty} g = 0$. Then,*

$$\int_{\mathbb{R}} fg \leq \int_{\mathbb{R}} f^* g^*$$

where f^*, g^* denote the symmetric decreasing rearrangements of f, g respectively.

Error-Correcting Codes. Let $k, n \in \mathbb{N}$, and let Σ be a finite alphabet. A *code* is a one-to-one function $C \colon \Sigma^k \to \Sigma^n$ that maps *messages* to *codewords*, where k and n are called the code's *dimension* and *block length*, respectively. The *rate* of the code, measuring the redundancy of the encoding, is defined to be $\rho \stackrel{\mathrm{def}}{=} k/n$. We will sometime identify the code C with its image $C(\Sigma^k)$. In particular, we shall write $c \in C$ to indicate that there exists $x \in \{0,1\}^k$ such that $c = C(x)$, and say that c is a codeword of C. The *relative distance* of a code is the minimal relative distance between two codewords of C, and is denoted by $\delta \stackrel{\mathrm{def}}{=} \min_{c \neq c' \in C}\{\mathrm{dist}(c,c')\}$.

We say that C is an *asymptotically good code* if it has constant rate and constant relative distance. We shall make an extensive use of asymptotically good codes that are *balanced*, that is, codes in which each codeword consists of the same number of 0's and 1's

---

[9] Note that this is equal, up to a factor 2, to the total variation distance between p and q.

▶ **Proposition 5** (Good Balanced Codes). *For any constant $\delta \in [0, 1)$, there exists a good balanced code $C \colon \{0, 1\}^k \to \{0, 1\}^n$ with relative distance δ and constant rate. Namely, there exists a constant $\rho > 0$ such that the following holds.*

1. *Balance: $|C(x)| = \frac{n}{2}$ for all $x \in \{0, 1\}^k$;*
2. *Relative distance: $\mathrm{dist}(C(x), C(y)) > \delta$ for all distinct $x, y \in \{0, 1\}^k$;*
3. *Constant rate: $\frac{k}{n} \geq \rho$.*

Proof. Fix any code C' with linear distance δ and constant rate (denoted ρ'). We transform $C' \colon \{0, 1\}^k \to \{0, 1\}^{n'}$ to a balanced code $C \colon \{0, 1\}^k \to \{0, 1\}^{2n'}$ by representing 0 and 1 as the balanced strings 01 and 10 (respectively). More accurately, we let $C(x) \stackrel{\text{def}}{=} C'(x) \odot \overline{C'(x)} \in \{0, 1\}^n$ for all $x \in \{0, 1\}^k$, where $\odot$ denotes the concatenation and $\bar{z}$ is the bitwise negation of z. It is immediate to check that this transformation preserves the distance, and that C is a balanced code with rate $\rho \stackrel{\text{def}}{=} 2\rho'$. ◀

On uniformity. For the sake of notation and clarity, throughout this work we define all algorithms and objects non-uniformly. Namely, we fix the relevant parameter (typically $n \in \mathbb{N}$), and restrict ourselves to inputs or domains of size n (for instance, probability distributions over domain $[n]$). However, we still view it as a generic parameter and allow ourselves to write asymptotic expressions such as $O(n)$. Moreover, although our results are stated in terms of non-uniform algorithms, they can be extended to the uniform setting in a straightforward manner.

4 The Methodology: From Communication Complexity to Distribution Testing

In this section we adapt the methodology for proving property testing lower bounds via reductions from communication complexity, due to Blais, Brody, and Matulef [15], to the setting of distribution testing. As observed in [15, 16], to prove lower bounds on the query complexity of *non-adaptive* testers it suffices to reduce from one-sided communication complexity. We show that for distribution testers (which are inherently non-adaptive), it suffices to reduce from the more restricted communication complexity model of *private-coin* simultaneous message passing (SMP).

Recall that a private-coin SMP protocol for a communication complexity predicate $f \colon \{0, 1\}^k \times \{0, 1\}^k \to \{0, 1\}$ consists of three computationally unbounded parties: Two players (commonly referred to as Alice and Bob), and a Referee. Alice and Bob receive inputs $x, y \in \{0, 1\}^k$. Each of the players simultaneously (and independently) sends a message to the referee, based on its input and (private) randomness. The referee is then required to successfully compute $f(x, y)$ with probability at least 2/3, using its private randomness and the messages received from Alice and Bob. The communication complexity of an SMP protocol is the total number of bits sent by Alice and Bob. The private-coin SMP complexity of f, denoted $\mathsf{SMP}(f)$, is the minimum communication complexity of all SMP protocols that solve f with probability at least 2/3.

Generally, to reduce an SMP problem f to ε-testing a distribution property Π, Alice and Bob can send messages $m_A(x, r_A, \varepsilon)$ and $m_B(y, r_B, \varepsilon)$ (respectively) to the Referee, where r_A and r_B are the private random strings of Alice and Bob. Subsequently, the Referee uses the messages $m_A(x, r_A, \varepsilon)$ and $m_A(y, r_B, \varepsilon)$, as well as its own private randomness, to feed the property tester samples from a distribution p that satisfies the following conditions: (1) *completeness*: if $f(x, y) = 1$, then $p \in \Pi$; and (2) *soundness*: if $f(x, y) = 0$, then p is ε-far from Π in ℓ_1-distance.

We shall focus on a special type of the foregoing reductions, which is particularly convenient to work with and suffices for all of the our lower bounds. Loosely speaking, in these reductions Alice and Bob both send the prover samples from sub-distributions that can be combined by the Referee to obtain samples from a distribution that satisfies the completeness and soundness conditions. The following lemma gives a framework for proving lower bounds based on such reductions.

▶ **Lemma 6.** *Let $\varepsilon > 0$, and let Ω be a finite domain of cardinality n. Fix a property $\Pi \subseteq \Delta(\Omega)$ and a communication complexity predicate $f\colon \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$. Suppose that there exists a mapping $p\colon \{0,1\}^k \times \{0,1\}^k \to \Delta(\Omega)$ that satisfies the following conditions.*

1. *Decomposability: For every $x, y \in \{0,1\}^k$, there exist constants $\alpha = \alpha(x), \beta = \beta(y) \in [0,1]$ and distributions $p_A(x), p_B(y)$ such that*

$$p(x,y) = \frac{\alpha}{\alpha + \beta} \cdot p_A(x) + \frac{\beta}{\alpha + \beta} \cdot p_B(y)$$

and α, β can each be encoded with $O(\log n)$ bits.
2. *Completeness: For every $(x,y) = f^{-1}(1)$, it holds that $p(x,y) \in \Pi$.*
3. *Soundness: For every $(x,y) = f^{-1}(0)$, it holds that $p(x,y)$ is ε-far from Π in ℓ_1 distance.*

Then, every ε-tester for Π needs $\Omega\left(\frac{\mathsf{SMP}(f)}{\log(n)}\right)$ samples.

Proof. Suppose there exists an ε-tester for Π with sample complexity s'; assume without loss of generality that the soundness of the foregoing tester is $5/6$, at the cost of increasing the query complexity to $s = O(s')$. Let $x, y \in \{0,1\}^k$ be the inputs of Alice and Bob (respectively) for the SMP problem. Alice computes the distribution $p_A(x)$ and the "decomposability parameter" $\alpha = \alpha(x)$ and sends $6s$ independent samples from $p_A(x)$, as well as the parameter α. Analogously, Bob computes $p_B(y)$ and its parameter $\beta = \beta(y)$, and sends $6s$ independent samples from $p_B(y)$ as well as the parameter β. Subsequently, the referee generates a sequence of q independent samples from $p(x,y)$, where each sample is drawn as follows: with probability $\frac{\alpha}{\alpha+\beta}$ use a (fresh) sample from Alice's samples, and with probability $1 - \frac{\alpha}{\alpha+\beta}$ use a (fresh) sample from Bob's samples. Finally the referee feeds the generated samples to the ε-tester for Π.

By Markov's inequality, the above procedure indeed allows the referee to retrieve, with probability at least $1 - \frac{\alpha s}{6s} \geq \frac{5}{6}$, at least s independent samples from the distribution $\frac{\alpha}{\alpha+\beta} \cdot p_A(x) + \frac{\beta}{\alpha+\beta} \cdot p_B(y)$, which equals to $p(x,y)$, by the decomposability condition. If $(x,y) = f^{-1}(1)$, then by the completeness condition $p(x,y) \in \Pi$, and so the ε-tester for Π is successful with probability at least $\frac{5}{6} \cdot \frac{5}{6}$. Similarly, if $(x,y) = f^{-1}(0)$, then by the soundness condition $p(x,y)$ is ε-far from Π, and so the ε-tester for Π is successful with probability at least $\frac{5}{6} \cdot \frac{5}{6}$. Finally, note that since each one of the samples provided by Alice and Bob requires sending $\log n$ bits, the total communication complexity of the protocol is $12s \log n + O(\log n)$ (the last term from the cost of sending α, β), hence $s' = \Omega\left(\frac{\mathsf{SMP}(f)}{\log(n)}\right)$. ◀

We conclude this section by stating a well-known SMP lower bound on the equality problem. Let $\mathrm{EQ}_k\colon \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$ be the equality predicate, i.e., $\mathrm{EQ}_k(x,y) = 1$ if and only if $x = y$. In this work, we shall frequently use the following (tight) lower bound on the EQ_k predicate:

▶ **Theorem 7** (Newman and Szegedy [40]). *For every $k \in \mathbb{N}$ it holds that $\mathsf{SMP}(\mathrm{EQ}_k) = \Omega\left(\sqrt{k}\right)$.*

5 The Basic Reduction: The Case of Uniformity

▶ **Theorem 8.** *For any $\varepsilon \in (0, 1/2)$ and finite domain Ω, testing that $p \in \Delta(\Omega)$ is uniform, with respect to proximity parameter ε, requires $\tilde{\Omega}(\sqrt{n})$ samples, where $n = |\Omega|$.*

Proof. Assume there exists a q-query ε-tester for the uniform distribution, with error probability $1/6$. For a sufficiently large $k \in \mathbb{N}$, let $C \colon \{0,1\}^k \to \{0,1\}^n$ be a balanced code as promised by Proposition 5 with distance ε. Namely, there exists an absolute constant $\rho > 0$ such that

1. $|C(x)| = \frac{n}{2}$ for all $x \in \{0,1\}^k$;
2. $\operatorname{dist}(C(x), C(y)) > \varepsilon$ for all distinct $x, y \in \{0,1\}^k$;
3. $\frac{k}{n} \geq \rho$.

Given their respective inputs $x, y \in \{0,1\}^k$ from EQ_k, Alice and Bob separately create inputs $(C(x), C(y)) \in \{0,1\}^n \times \{0,1\}^n$, and the corresponding sets $A \stackrel{\text{def}}{=} \{\, i \in [n] \,:\, C(x)_i = 1 \,\}$, $B \stackrel{\text{def}}{=} \{\, i \in [n] \,:\, C(y)_i = 0 \,\}$. We then invoke the general reduction of Lemma 6 as follows: we set $\alpha = \beta = \frac{1}{2}$, and $p_A(x) \in \Delta([n])$ (respectively $p_B(y) \in \Delta([n])$) to be the uniform distribution on the set A (respectively B). It is clear that the decomposability condition of the lemma is satisfied for $p(x, y) = \frac{\alpha}{\alpha+\beta} \cdot p_A(x) + \frac{\beta}{\alpha+\beta} \cdot p_B(y) = \frac{1}{2}(p_A(x) + p_B(y))$; we thus turn to the second and third conditions.

Completeness. If $(x, y) \in \mathrm{EQ}_k^{-1}(1)$, then $C(x) = C(y)$ and $A = [n] \setminus B$. This implies that $p(x, y)$ is indeed the uniform distribution on $[n]$, as desired.

Soundness. If $(x, y) \in \mathrm{EQ}_k^{-1}(0)$, then $\operatorname{dist}(C(x), C(y)) > \varepsilon$, and therefore $|A \triangle \bar{B}| > \varepsilon n$ by construction. Since $p(x, y)$ assigns mass $2/n$ to each element in $A \cap B = A \setminus \bar{B}$, and mass 0 to any element in $\bar{A} \cap \bar{B} = \bar{B} \setminus A$, we have $\|p(x, y) - u\|_1 = \frac{1}{n} \cdot |A \triangle \bar{B}| > \varepsilon$; that is, $p(x, y)$ is ε-far from uniform.

The desired $\Omega\left(\frac{\sqrt{n}}{\log n}\right)$ lower bound then immediately follows from Lemma 6 and Theorem 7. ◀

5.1 Obtaining ε-Dependency

In this section, we explain how to generalize the reduction from the previous section to obtain some dependence (albeit non optimal) on the distance parameter ε in the lower bound. This generalization will rely on an extension of the methodology of Lemma 6: instead of having the referee define the distribution $p(x, y)$ as a mixture of $p_A(x)$ and $p_B(y)$ (namely, $p(x, y) = \frac{\alpha(x)}{\alpha(x)+\beta(y)} p_A(x) + \frac{\beta(y)}{\alpha(x)+\beta(y)} p_B(y)$), he will instead use a (random) combination function F_ε, function of ε and its private coins only. Given this function, which maps a larger domain of size $m = \Theta(n/\varepsilon^2)$ to $[n]$, $p(x, y)$ will be defined as the mixture

$$p(x, y) = \frac{\alpha(x)}{\alpha(x) + \beta(y)} p_A(x) \circ F_\varepsilon^{-1} + \frac{\beta(y)}{\alpha(x) + \beta(y)} p_B(y) \circ F_\varepsilon^{-1}.$$

More simply, this allows Alice and Bob to send to the referee samples from their respective distributions on a much larger domain $m \gg n$; the referee, who has on its side chosen how to randomly partition this large domain into only n different "buckets," converts these draws from Alice and Bob into samples from the induced distributions on the n buckets, and takes a mixture of these two distributions instead. By choosing each bucket to chave size roughly $1/\varepsilon^2$, we expect this random "coarsening" of Alice and Bob's distributions to yield a distribution at distance only $\Omega(\varepsilon)$ from uniformity (instead of constant distance) in the no-case; but now letting us get a lower bound on the *original* support size m, i.e. $\tilde{\Omega}\left(\sqrt{n/\varepsilon^2}\right)$, instead of $\tilde{\Omega}(\sqrt{n})$ as before.

▶ **Theorem 9.** *For any $\varepsilon \in (0, 1/2)$ and finite domain Ω, testing that $p \in \Delta(\Omega)$ is uniform, with respect to proximity parameter ε, requires $\tilde{\Omega}(\sqrt{n}/\varepsilon)$ samples, where $n = |\Omega|$.*

Proof of Theorem 9. We will reduce from EQ_k, where $k \in \mathbb{N}$ is again assumed big enough (in particular, with regard to $1/\varepsilon^2$). Alice and Bob act as in Section 5, separately creating $(a, b) = (C(x), C(y)) \in \{0, 1\}^m \times \{0, 1\}^m$ from their respective inputs $x, y \in \{0, 1\}^k$ (where $C \colon \{0, 1\}^k \to \{0, 1\}^m$ is a balanced code with linear rate and distance $\delta \overset{\text{def}}{=} 1/3$). As before, they consider the sets $A \overset{\text{def}}{=} \{ i \in [m] : C(x)_i = 1 \}$, $B \overset{\text{def}}{=} \{ i \in [m] : C(y)_i = 0 \}$, set $\alpha = \beta = \frac{1}{2}$, and consider the distributions $p_A(x), p_B(y) \in \Delta([m])$ which are uniform respectively on A and B.

This is where we deviate from the proof of Theorem 8: indeed, setting $n \overset{\text{def}}{=} c\varepsilon^2 m$ (where $c > 0$ is an absolute constant determined later), the referee will combine the samples from $p_A(x)$ and $p_B(y)$ in a different way to emulate a distribution $p(x, y) \in \Delta([n])$ – that is, with a much smaller support than that of $p_A(x), p_B(y)$ (instead of setting $p(x, y)$ to be, as before, a mixture of the two).

To do so, the referee randomly partitions $[m]$ into n sets $B_1, \ldots, B_n$ of equal size $r \overset{\text{def}}{=} |B_j| = \frac{m}{n} = \frac{1}{c\varepsilon^2}$, $j \in [n]$, by choosing a uniformly random equipartition of $[m]$. He then defines the distribution $p = p(x, y) \in \Delta([n])$ by $p(j) = \Pr[i \in B_j]$ (where $i \in [m]$ is received from either Alice or Bob). Viewed differently, the random equipartition chosen by the referee induces a mapping $F_\varepsilon \colon [m] \to [n]$ such that $|F^{-1}(j)| = r$ for all $j \in [n]$; and, setting $p'(x, y) = \frac{1}{2}(p_A(x) + p_B(y)) \in \Delta([m])$, we obtain $p(x, y)$ as the *coarsening* of $p'(x, y)$ defined as

$$p(x, y)(j) = \sum_{i \in F_\varepsilon^{-1}(j)} p'(x, y)(i) = p'(x, y)(F_\varepsilon^{-1}(j))$$

$$= \frac{1}{2}\left(p_A(x)(F_\varepsilon^{-1}(j)) + p_B(y)(F_\varepsilon^{-1}(j))\right), \qquad j \in [n].$$

Note furthermore that each sample sent by Alice and Bob (who have no knowledge of the randomly chosen F_ε) can be encoded with $O(\log m) = O(\log \frac{n}{\varepsilon})$ bits.

We then turn to establish the analogue in this generalized reduction of the last two conditions of Lemma 6, i.e. the completeness and soundness. The former, formally stated below, will be an easy consequence of the previous section.

▶ **Claim 10.** *If $x = y$, then $p(x, y)$ is uniform on $[n]$.*

Proof. As in the proof of Theorem 8, in this case the distribution $p'(x, y) = \frac{1}{2}(p_A(x) + p_B(y)) \in \Delta([m])$ is uniform; since each "bucket" $B_j = F_\varepsilon^{-1}(j)$ has the same size, this implies that $p(x, y)(j) = p'(x, y)(B_j) = \frac{1}{n}$ for all $j \in [n]$. ◀

Establishing the soundness, however, is not as straightforward:

▶ **Claim 11.** *If $x \neq y$, then with probability at least $1/100$ (over the choice of the equipartition $(B_1, \ldots, B_n)$), $p(x, y)$ is ε-far from uniform.*

Proof. Before delving into the proof, we provide a high-level idea of why this holds. Since the partition was chosen uniformly at random, on expectation each element $j \in [n]$ will have probability $\mathbb{E}[p(x, y)(j)] = \mathbb{E}[p'(x, y)(B_j)] = \frac{1}{n}$. However, since a constant fraction of elements $i \in [m]$ (before the random partition) has probability mass either 0 or $2/m$ (as in the proof of Theorem 8), and each bucket B_j contains $r = 1/(c\varepsilon^2)$ many elements chosen uniformly at random, we expect the fluctuations of $p'(x, y)(B_j)$ around its expectation to be

of the order of $\Omega(\sqrt{r}/m) = \Omega(\varepsilon/n)$ with constant probability, and summing over all j's this will give us the distance $\Omega(\varepsilon)$ we want.

To make this argument precise, we assume $x \neq y$, so that $A \triangle \bar{B} > \delta m$; and define $H \overset{\text{def}}{=} A \cap B, L \overset{\text{def}}{=} \bar{A} \cap \bar{B}$ (so that $|H| = |L| > \frac{\delta}{2}m$). For any $j \in [n]$, we then let the random variables $H^{(j)}, L^{(j)}$ be the number of "high" and "low" elements of $[m]$ in the bucket B_j, respectively:

$$H^{(j)} \overset{\text{def}}{=} |B_j \cap H|, \qquad L^{(j)} \overset{\text{def}}{=} |B_j \cap L|.$$

From the definition, we get that $p = p(x,y)$ satisfies $p(j) = \frac{1}{m}\left(2H^{(j)} + (r - H^{(j)} - L^{(j)})\right) = \frac{r}{m} + \frac{H^{(j)} - L^{(j)}}{m}$ for $j \in [n]$. Furthermore, it is easy to see that $\mathbb{E}[p(j)] = \frac{r}{m} = \frac{1}{n}$ for all $j \in [n]$, where the expectation is over the choice of the equipartition by the referee.

As previously discussed, we will analyze the deviation from this expectation; more precisely, we want to show that with good probability, a constant fraction of the j's will be such that $p(j)$ deviates from $1/n$ by at least an additive $\Omega(\sqrt{r}/m) = \varepsilon/n$. This anticoncentration guarantee will be a consequence of the Paley–Zygmund inequality (Theorem 3) to $Z^{(j)} \overset{\text{def}}{=} (H^{(j)} - L^{(j)})^2 \geq 0$; in view of applying it, we need to analyze the first two moments of this random variable.

▶ **Lemma 12.** *For any $j \in [n]$, we have the following. (i) $\mathbb{E}\left[(H^{(j)} - L^{(j)})^2\right] = \delta r \frac{m-r}{m-1}$, and (ii) $\mathbb{E}\left[(H^{(j)} - L^{(j)})^4\right] = 3(1 + o(1))\delta^2 r^2$.*

Proof. Fix any $j \in [n]$. We write for convenience X and Y for respectively $H^{(j)}$ and $L^{(j)}$. The distribution of $(X, Y, r - (X - Y))$ is then a *multivariate hypergeometric distribution* [55] with 3 classes:

$$(X, Y, r - (X + Y)) \sim \text{MultivHypergeom}_3(\underbrace{(\tfrac{1}{2}\delta m, \tfrac{1}{2}\delta m, (1 - \delta)m)}_{(K_1, K_2, K_3)}, m, r).$$

Conditioning on $U \overset{\text{def}}{=} X + Y$, we have that $\mathbb{E}[X \mid U]$ follows a hypergeometric distribution, specifically $\mathbb{E}[X \mid U] \sim \text{Hypergeom}(U, \frac{1}{2}\delta m, \delta m)$. Moreover, U itself is hypergeometrically distributed, with $U \sim \text{Hypergeom}(r, \delta m, m)$. We can thus write

$$\mathbb{E}\left[(X - Y)^2\right] = \mathbb{E}\left[\mathbb{E}\left[(X - Y)^2 \mid U\right]\right] = \mathbb{E}\left[\mathbb{E}\left[(2X - U)^2 \mid U\right]\right]$$

and

$$\mathbb{E}\left[(X - Y)^4\right] = \mathbb{E}\left[\mathbb{E}\left[(X - Y)^4 \mid U\right]\right] = \mathbb{E}\left[\mathbb{E}\left[(2X - U)^4 \mid U\right]\right].$$

By straightforward, yet tedious, calculations involving the computation of $\mathbb{E}\left[(2X - U)^2 \mid U\right]$ and $\mathbb{E}\left[(2X - U)^4 \mid U\right]$ (after expanding and using the known moments of the hypergeometric

---

[10] One can also use a formal computation system, e.g. Mathematica:

```
Expectation[ Expectation[(2 X - U)^2, {X \[Distributed] HypergeometricDistribution[U, a*m, 2 a*m]},
  {U \[Distributed] HypergeometricDistribution[r, 2*a*m, m]}]
Expectation[ Expectation[(2 X - U)^4, {X \[Distributed] HypergeometricDistribution[U, a*m, 2 a*m]},
  {U \[Distributed] HypergeometricDistribution[r, 2*a*m, m]}]
```

distribution),[10] we obtain

$$\mathbb{E}\big[(X-Y)^2\big] = \delta r \frac{m-r}{m-1} \underset{m\to\infty}{=} (1+o(1))\delta r$$

$$\mathbb{E}\big[(X-Y)^4\big] = \frac{(\delta r(r-m)((-1+3\delta(m-1)-m)m+6r^2(\frac{1}{2}\delta m-1)-6rm(\frac{1}{2}\delta m-1)))}{(m-3)(m-2)(m-1)}$$

$$\xrightarrow[m\to\infty]{} 3\delta^2 r^2 + (1-3\delta)\delta r = 3\delta^2 r^2$$

the last equality as $\delta = 1/3$. ◀

We can now apply the Paley–Zygmund inequality to $Z^{(j)}$. Doing so, we obtain that for $r \leq \frac{m}{4}$ (with some slack), and any $\theta \in [0,1]$,

$$\Pr\left[\left|H^{(j)}-L^{(j)}\right| \geq \theta\sqrt{\frac{1}{2}\delta r}\right] \geq \Pr\left[\left|H^{(j)}-L^{(j)}\right| \geq \theta\sqrt{\delta r\frac{m-r}{m-1}}\right]$$

$$\geq (1-\theta^2)^2 \frac{\mathbb{E}\big[(H^{(j)}-L^{(j)})^2\big]^2}{\mathbb{E}\big[(H^{(j)}-L^{(j)})^4\big]}.$$

By the lemma above, the RHS converges to $\frac{(1-\theta^2)^2}{3}$ when $m\to\infty$, and therefore is at least $\frac{(1-\theta^2)^2}{4}$ for m big enough. We set $\theta \stackrel{\text{def}}{=} 1/\sqrt{2}$ to obtain the following: there exists $M \geq 0$ such that

$$\Pr\left[\left|H^{(j)}-L^{(j)}\right| \geq \sqrt{\frac{\delta r}{4}}\right] \geq \frac{1}{16} \tag{4}$$

for every $m \geq M$.

Eq. (4) implies that the number K of *good* indices $j \in [n]$ satisfying $\left|H^{(j)}-L^{(j)}\right| \geq \sqrt{\frac{\delta r}{4}}$ is on expectation at least $\frac{n}{16}$, and by an averaging argument[11] we get that $K \geq \frac{n}{20}$ with probability at least $\frac{1}{76} > \frac{1}{100}$.

Whenever this happens, the distance from p to uniform is at least

$$\sum_{j\ \text{good}} \left|p(j)-\frac{1}{n}\right| = \sum_{j\ \text{good}} \frac{\left|H^{(j)}-L^{(j)}\right|}{m} \geq \frac{n}{20} \cdot \frac{\sqrt{\frac{\delta r}{4}}}{m} = \frac{\sqrt{\delta r}}{40}\frac{n}{m} = \frac{\sqrt{c}}{40\sqrt{3}}\varepsilon$$

and choosing $c \geq 4800$ so that $\frac{\sqrt{c}}{40\sqrt{3}} \geq 1$ yields the claim. ◀

From this lemma, we can complete the reduction: given a tester $\mathcal{T}$ for uniformity with query complexity q, we first convert it by standard amplification into a tester $\mathcal{T}'$ with failure probability $\delta \stackrel{\text{def}}{=} 1/1000$ and sample complexity $O(q)$. The referee can provide samples from the distribution $p(x,t)$, and on input ε:

- If $x = y$, then $\mathcal{T}'$ will return reject with probability at most $1/200$;
- If $x \neq y$, then $\mathcal{T}'$ will return reject with probability at least $199/200 \cdot 1/100 > 1/200$;

so repeating independently the protocol a constant (fixed in advance) number of times and taking a majority vote enables the referee to solve EQ_k with probability at least $2/3$. Since $\Omega\big(\sqrt{k}\big) = \Omega\big(\sqrt{n/\varepsilon^2}\big)$ bits of communication are required for this, and each sample sent by

---

[11] Applying Markov's inequality: $\Pr\left[K < \frac{n}{20}\right] = \Pr\left[n-K > \frac{19n}{20}\right] \leq \frac{n-\mathbb{E}[K]}{19n/20} \leq \frac{15/16}{19/20} = \frac{75}{76}$.

Alice or Bob to the referee only requires $\Theta\left(\log\frac{n}{\varepsilon}\right)$ bits, we get a lower bound of

$$\Omega\left(\frac{\sqrt{n}}{\varepsilon\log\frac{n}{\varepsilon}}\right) = \tilde{\Omega}\left(\frac{\sqrt{n}}{\varepsilon}\right)$$

on the sample complexity of $\mathcal{T}'$, and therefore of $\mathcal{T}$. ◀

6 The K-Functional: An Unexpected Journey

A quantity that will play a major role in our results is the K-*functional between ℓ_1 and ℓ_2*, a specific case of the key operator in interpolation theory introduced by Peetre [43]. We start by recalling below the definition and some of its properties, before establishing (for our particular setting) results that will be crucial to us. (For more on the K-functional and its use in functional analysis, the reader is referred to [12] and [6].)

▶ **Definition 13** (K-functional). Fix any two Banach spaces $(X_0, \|\cdot\|_0), (X_1, \|\cdot\|_1)$. The K-*functional* between X_0 and X_1 is the function $K_{X_0, X_1} \colon (X_0 + X_1) \times (0, \infty) \to [0, \infty)$ defined by

$$K_{X_0, X_1}(x, t) \overset{\text{def}}{=} \inf_{\substack{(x_0, x_1) \in X_0 \times X_1 \\ x_0 + x_1 = x}} \|x_0\|_0 + t\|x_1\|_1.$$

For $a \in \ell_1 + \ell_2$, we denote by κ_a the function $t \mapsto K_{\ell_1, \ell_2}(a, t)$.

In other terms, as t varies the quantity $\kappa_a(t)$ interpolates between the ℓ_1 and ℓ_2 norms of the sequence a (and accordingly, for any fixed t it defines a norm on $\ell_1 + \ell_2$). In particular, note that for large values of t the function $\kappa_a(t)$ is close to $\|x\|_1$, whereas for small values of t the function $\kappa_a(t)$ is close to $t\|x\|_2$ (see Corollary 17). We henceforth focus on the case of K_{ℓ_1, ℓ_2}, although some of the results mentioned hold for the general setting of arbitrary Banach X_0, X_1.

▶ **Proposition 14** ([12, Proposition 1.2]). *For any $a \in \ell_1 + \ell_2$, κ_a is continuous, increasing, and concave. Moreover, the function $t \in (0, 1) \mapsto \frac{\kappa_a}{t}$ is decreasing.*

Although no closed-form expression is known for κ_a, it will be necessary for us to understand its behavior, and therefore seek good upper and lower bounds on its value. We start with the following inequality, due to Holmstedt [34], which, loosely speaking, shows that the infimum in the definition of $\kappa_a(t)$ is *roughly* obtained by partitioning $a = (a_1, a_2)$ such that a_1 consists of heaviest t^2 coordinates of a, and a_2 consists of the rest.

▶ **Proposition 15** ([6, Proposition 2.2], after [34, Theorem 4.2]). *For any $a \in \ell_2$ and $t > 0$,*

$$\frac{1}{4}\left(\sum_{i=1}^{\lfloor t^2 \rfloor} a_i^* + t\left(\sum_{i=\lfloor t^2 \rfloor + 1}^{\infty} a_i^{*2}\right)^{\frac{1}{2}}\right) \leq \kappa_a(t) \leq \sum_{i=1}^{\lfloor t^2 \rfloor} a_i^* + t\left(\sum_{i=\lfloor t^2 \rfloor + 1}^{\infty} a_i^{*2}\right)^{\frac{1}{2}} \tag{5}$$

where a^* is a non-increasing permutation of the sequence $(|a_i|)_{i \in \mathbb{N}}$.

(We remark that for our purposes, this constant factor gap between left-hand and right-hand side is not innocuous, as we will later need to study te behavior of the *inverse* of the function κ_a.)

Incomparable bounds on κ_a were obtained [38], relating it to a different quantity, the "Q-norm," which we discuss and generalize next.

6.1 Approximating the K-Functional by the Q-norm

Loosely speaking, the Q-norm of a vector a (for a given parameter T) is a *mixed ℓ_1/ℓ_2 norm*: it is the maximum one can reach by partitioning the components of a into T sets, and taking the sum of the ℓ_2 norms of these T subvectors. Although not straightforward to interpret, this intuitively captures the notion of *sparsity* of a: indeed, if a is supported on k elements then its Q-norm becomes equal to the ℓ_1 norm for parameter $T \geq k$.

▶ **Proposition 16** ([6, Lemma 2.2], after [38, Lemma 2]). *For arbitrary $a \in \ell_2$ and $t \in \mathbb{N}$, define the norm*

$$\|a\|_{Q(t)} \stackrel{\text{def}}{=} \sup \left\{ \sum_{j=1}^{t} \left(\sum_{i \in A_j} a_i^2 \right)^{1/2} \ : \ (A_j)_{1 \leq j \leq t} \text{ partition of } \mathbb{N} \right\}.$$

Then, for any $a \in \ell_2$, and $t > 0$ such that $t^2 \in \mathbb{N}$, we have

$$\|a\|_{Q(t^2)} \leq \kappa_a(t) \leq \sqrt{2}\|a\|_{Q(t^2)}. \tag{6}$$

As we shall see shortly, one can generalize this result further, obtaining a tradeoff in the upper bound. Before turning to this extension in Lemma 18 and Lemma 21, we first state several other properties of the K-functional implied by the above:

▶ **Corollary 17.** *For any $a \in \ell_2$,*
1. $\kappa_a(t) = t\|a\|_2$ *for all $t \in (0, 1)$*
2. $\lim_{t \to 0^+} \kappa_a(t) = 0$
3. $\frac{1}{4}\|a\|_1 \leq \lim_{t \to \infty} \kappa_a(t) \leq \|a\|_1$.
Moreover, for a supported on finitely many elements, it is the case that $\lim_{t \to \infty} \kappa_a(t) = \|a\|_1$.

Proof. The first two points follow by definition; turning to Corollary 3, we first note the upper bound is a direct consequence of the definition of κ_a as an infimum (as, for all $t > 0$, $\kappa_a(t) \leq \|a\|_1$). (This itself ensures the limit as $t \to \infty$ exists by monotone convergence, as κ_a is a non-decreasing bounded function.) The lower bound follows from that of Theorem 15, which guarantees that for all $t > 0$ $\kappa_a(t) \geq \frac{1}{4} \sum_{i=1}^{\lfloor t^2 \rfloor} a_i^* \xrightarrow{t \to \infty} \frac{1}{4}\|a\|_1$. Finally, the last point can be obtained immediately from, e.g., the lower bound side of Proposition 16 and the upper bound given on Corollary 3 above. ◀

▶ **Lemma 18.** *For any $a \in \ell_2$ and t such that $t^2 \in \mathbb{N}$, we have*

$$\|a\|_{Q(t^2)} \leq \kappa_a(t) \leq \|a\|_{Q(2t^2)}. \tag{7}$$

Proof of Lemma 18. We follow and adapt the proof of [6, Lemma 2.2] (itself similar to that of [38, Lemma 2]). The first inequality is immediate: indeed, for any sequence $c \in \ell_2$, by the definition of $\|a\|_{Q(t^2)}$ and the monotonicity of the p-norms, we have $\|c\|_{Q(t^2)} \leq \|c\|_1$; and by Cauchy–Schwarz, for any partition $(A_j)_{1 \leq j \leq t^2}$ of $\mathbb{N}$,

$$\sum_{j=1}^{t^2} \left(\sum_{i \in A_j} c_i^2 \right)^{1/2} \leq t \left(\sum_{j=1}^{t^2} \sum_{i \in A_j} c_i^2 \right)^{1/2} = t\|c\|_2$$

and thus $\|c\|_{Q(t^2)} \leq t\|c\|_2$. This yields the lower bound, as

$$\kappa_a(t) = \inf_{\substack{a'+a''=a \\ a' \in \ell_1, a'' \in \ell_2}} \|a'\|_1 + t\|a''\|_2 \geq \inf_{\substack{a'+a''=a \\ a' \in \ell_1, a'' \in \ell_2}} \|a'\|_{Q(t^2)} + \|a''\|_{Q(t^2)} \geq \|a\|_{Q(t^2)}$$

by the triangle inequality.

We turn to the upper bound. As $\ell_2(\mathbb{R})$ is a symmetric space and $\kappa_a = \kappa_{|a|}$, without loss of generality, we can assume that $(a_k)_{k \in \mathbb{N}}$ is non-negative and monotone non-increasing, i.e. $a_1 \geq a_2 \geq \cdots \geq a_k \geq \ldots$. We will rely on the characterization of κ_a as

$$\kappa_a(t) = \sup \left\{ \sum_{k=1}^{\infty} a_k b_k \ : \ b \in \ell_2, \max(\|b\|_\infty, t^{-1}\|b\|_2) \leq 1 \right\}, \qquad t > 0$$

(see e.g. [6, Lemma 2.2] for a proof). The first step is to establish the existence of a "nice" sequence $b \in \ell_2$ arbitrarily close to this supremum:

▶ **Claim 19.** *For any $\delta > 0$, there exists a* non-increasing, *non-negative sequence $b^* \in \ell_2$ with* $\max(\|b^*\|_\infty, t^{-1}\|b^*\|_2) \leq 1$ *such that*

$$(1-\delta)\kappa_a \leq \sum_{k=1}^{\infty} a_k b_k^*.$$

Proof. By the above characterization, there exists a sequence $b \in \ell_2$ with $\max(\|b\|_\infty, t^{-1}\|b\|_2) \leq 1$ such that $(1-\delta)\kappa_a \leq \sum_{k=1}^{\infty} a_k b_k$. We now claim that we can further take b to be non-negative and monotone non-increasing as well. The first part is immediate, as replacing negative terms by their absolute values can only increase the sum (since a is itself non-negative). For the second part, we will invoke the Hardy–Littlewood rearrangement inequality (Theorem 4), which states that for any two non-negative functions f, g vanishing at infinity, the integral $\int_{\mathbb{R}} fg$ is maximized when f and g are non-increasing. We now apply this inequality to a, b, letting a^*, b^* be the non-increasing rearrangements of a, b (in particular, we have $a = a^*$) and introducing the functions f_a, f_b:

$$f_a = \sum_{j=1}^{\infty} a_j \mathbb{1}_{(j-1,j]}, \quad f_b = \sum_{j=1}^{\infty} b_j \mathbb{1}_{(j-1,j]}$$

which satisfy the hypotheses of Theorem 4. Thus, we get $\int_{\mathbb{R}} f_a f_b \leq \int_{\mathbb{R}} f_a^* f_b^*$; as it is easily seen that $f_a^* = f_{a^*}$ and $f_b^* = f_{b^*}$, this yields

$$\sum_{k=1}^{\infty} a_k b_k = \int_{\mathbb{R}} f_a f_b \leq \int_{\mathbb{R}} f_a^* f_b^* = \sum_{k=1}^{\infty} a_k^* b_k^* = \sum_{k=1}^{\infty} a_k b_k^*.$$

Moreover, it is immediate to check that $\max(\|b^*\|_\infty, t^{-1}\|b^*\|_2) \leq 1$. ◀

The next step is to relate the inner product $\sum_{k=1}^{\infty} a_k b_k^*$ to the Q-norm of a:

▶ **Claim 20.** *Fix $t > 0$ such that $t^2 \in \mathbb{N}$, and let $b^* \in \ell_2$ be any non-increasing, non-negative sequence with* $\max(\|b^*\|_\infty, t^{-1}\|b^*\|_2) \leq 1$. *Then*

$$\sum_{k=1}^{\infty} a_k b_k^* \leq \|a\|_{Q(2t^2)}.$$

Proof. We proceed constructively, by exhibiting a partition of $\mathbb{N}$ into $2t^2$ sets $A_1, \ldots, A_{2t^2}$ satisfying $\sum_{k=1}^{\infty} a_k b_k^* \leq \sum_{j=1}^{2t^2} \left(\sum_{i \in A_j} b_i^{*2} \right)^{1/2}$. This will prove the claim, by definition of $\|a\|_{Q(2t^2)}$ as the supremum over all such partitions.

Specifically, we inductively choose $n_0, n_1, \ldots, n_T \in \{0, \ldots, \infty\}$ as follows, where $T \stackrel{\text{def}}{=} \frac{t^2}{c}$ for some $c > 0$ to be chosen later (satisfying $T \in \mathbb{N}$). If $0 = n_0 < n_1 < \cdots < n_m$ are already

set, then

$$n_{m+1} \stackrel{\text{def}}{=} 1 + \sup \left\{ \ell \geq n_m : \sum_{i=n_m+1}^{\ell} b_i^{*\,2} \leq c \right\}.$$

From $\|b^*\|_2 \leq t$, it follows that $n_T = \infty$. Let m^* be the first index such that $n_{m^*+1} > n_{m^*}+1$. Note that this implies (by monotonicity of b^*) that $b_i^{*\,2} > c$ for all $i \leq n_{m^*}$, and $b_i^{*\,2} \leq c$ for all $i \geq n_{m^*}+1$. We can write

$$\sum_{i=1}^{\infty} a_i b_i^* = \sum_{m=1}^{T} \sum_{i=n_{m-1}+1}^{n_m} a_i b_i^* = \sum_{i=1}^{n_{m^*}} a_i b_i^* + \sum_{m=m^*+1}^{T} \sum_{i=n_{m-1}+1}^{n_m} a_i b_i^*$$

Since $\|b^*\|_\infty \leq 1$ and $n_{m-1}+1 = n_m$ for all $m \leq m^*$, the first term can be bounded as

$$\sum_{i=1}^{n_{m^*}} a_i b_i^* \leq \sum_{i=1}^{n_{m^*}} \sqrt{a_i^2} = \sum_{m=1}^{m^*} \left(\sum_{i=n_{m-1}+1}^{n_m} a_i^2 \right)^{1/2}.$$

Turning to the second term, we recall that $b_i^{*\,2} \leq c$ for all $i \geq n_{m^*}+1$, so that $\sum_{i=n_{m-1}+1}^{n_m} b_i^{*\,2} \leq 2c$ for all $m \geq m^*+1$. This allows us to bound the second term as

$$\sum_{m=m^*+1}^{T} \sum_{i=n_{m-1}+1}^{n_m} a_i b_i^* \leq \sum_{m=m^*+1}^{T} \left(\sum_{i=n_{m-1}+1}^{n_m} b_i^{*\,2} \right)^{1/2} \left(\sum_{i=n_{m-1}+1}^{n_m} a_i^2 \right)^{1/2}$$

$$\leq \sqrt{2c} \sum_{m=m^*+1}^{T} \left(\sum_{i=n_{m-1}+1}^{n_m} a_i^2 \right)^{1/2}$$

Therefore, by combining the two we get that

$$(1-\delta)\kappa_a(t) \leq \sum_{m=1}^{m^*} \left(\sum_{i=n_{m-1}+1}^{n_m} a_i^2 \right)^{1/2} + \sqrt{2c} \sum_{m=m^*+1}^{T} \left(\sum_{i=n_{m-1}+1}^{n_m} a_i^2 \right)^{1/2}$$

$$\leq \max(1, \sqrt{2c}) \sum_{m=1}^{T} \left(\sum_{i=n_{m-1}+1}^{n_m} a_i^2 \right)^{1/2} \leq \max(1, \sqrt{2c})\|a\|_{Q(T)} = \|a\|_{Q(2t^2)}$$

the last equality by choosing $c \stackrel{\text{def}}{=} \frac{1}{2}$. ◀

We now fix an arbitrary $\delta > 0$, and let b^* be as promised by Claim 19. As this sequence satisfies the assumptions of Claim 20, putting the two results together leads to

$$(1-\delta)\kappa_a(t) \leq \sum_{k=1}^{\infty} a_k b_k^* \leq \|a\|_{Q(2t^2)}.$$

Since this holds for all $\delta > 0$, taking the limit as $\delta \searrow 0$ gives the (upper bound of the) lemma. ◀

We observe that, with similar techniques, one can also establish the following generalization of Proposition 16:

▶ **Lemma 21** (Generalization of Proposition 16). *For any $a \in \ell_2$, t, and $\alpha \in [1, \infty)$ such that $t^2, \alpha t^2 \in \mathbb{N}$, we have*

$$\|a\|_{Q(t^2)} \leq \kappa_a(t) \leq \sqrt{1 + \alpha^{-1}} \|a\|_{Q(\alpha t^2)}. \tag{8}$$

Proof of Lemma 21 (Sketch). We again follow the proof of [6, Lemma 2.2], up to the inductive definition of $n_1, \ldots, n_j$, which we change as

$$n_{m+1} = 1 + \sup\left\{ \ell \geq n_m : \sum_{i=n_m+1}^{\ell} b_i^2 \leq \frac{1}{\alpha} \right\}.$$

Since $\|b\|_\infty \leq 1$, we have $\sum_{i=n_m+1}^{n_{m+1}} b_i^2 \leq 1 + \frac{1}{\alpha}$. From $\|b\|_2 \leq t$, it follows that $n_{\alpha t^2} = \infty$. Therefore, for any $\delta > 0$,

$$(1-\delta)\kappa_a(t) \leq \sum_{i=1}^{\infty} a_i b_i \leq \sum_{m=1}^{T} \left(\sum_{i=n_{m-1}+1}^{n_m} b_i^2 \right)^{1/2} \left(\sum_{i=n_{m-1}+1}^{n_m} a_i^2 \right)^{1/2} \leq \sqrt{1 + \frac{1}{\alpha}} \|a\|_{Q(\alpha t^2)}.$$

Since this holds for all $\delta > 0$, taking the limit gives the (upper bound of the) lemma. ◀

We note that further inequalities relating κ_a to other functionals of a were obtained in [33].

6.2 Concentration Inequalities for Weighted Rademacher Sums

The connection between the K-functional and tail bounds on weighted sums of Rademacher random variables was first made by Montgomery-Smith [38], to which the following result is due (we here state a version with slightly improved constants):

▶ **Theorem 22.** *Let $(X_i)_{i \in \mathbb{N}}$ be a sequence of independent Rademacher random variables, i.e. uniform on $\{-1, 1\}$. Then, for any $a \in \ell_2$ and $t > 0$,*

$$\Pr\left[\sum_{i=1}^{\infty} a_i X_i \geq \kappa_a(t) \right] \leq e^{-\frac{t^2}{2}}. \tag{9}$$

and, for any fixed $c > 0$ and all $t \geq 1$,

$$\Pr\left[\sum_{i=1}^{\infty} a_i X_i \geq \frac{1}{1+c} \kappa_a(t) \right] \geq e^{-\left(\frac{2}{c} \ln \frac{\sqrt{6}(1+c)}{c}\right)(t^2+c)}. \tag{10}$$

In particular,

$$\Pr\left[\sum_{i=1}^{\infty} a_i X_i \geq \frac{1}{2} \kappa_a(t) \right] \geq e^{-(\ln 24)(t^2+1)} \geq e^{-(2\ln 24)t^2}.$$

One can intepret the above theorem as stating that the (inverse of the) K-functional κ_a is the "right" parameter to consider in these tail bounds; while standard Chernoff or Hoeffding bounds will depend instead on the quantity $\|a\|_2$. Before giving the proof of this theorem, we remark that similar statements or improvements can be found in [33] and [6]; below, we closely follow the argument of the latter.

Proof of Theorem 22. The upper bound can be found in e.g. [38], or [6, Theorem 2.2]. For the lower bound, we mimic the proof due to Astashkin, improving the parameters of some of the lemmas it relies on.

▶ **Lemma 23** (Small improvement of (2.14) in [6, Lemma 2.3]). *If $a = (a_k)_{k \geq 1} \in \ell_2$, then, for any $\lambda \in (0, 1)$,*

$$\Pr\left[\left| \sum_{k=1}^{\infty} a_k X_k \right|^2 \geq \lambda \sum_{k=1}^{\infty} a_k^2 \right] \geq \frac{1}{3}(1 - \lambda)^2. \tag{11}$$

Proof of Lemma 23. The proof is exactly the same, but when invoking (1.10) for $p = 4$ we use the actual tight version proven there for $p = 2m$ (instead of the more general version that also applies to odd values of p): since $m = 2$, we get $\frac{(2m)!}{2^m m!} = 3$, and $\mathbb{E}[f]^2 \geq \frac{1}{3}\mathbb{E}[f^2]$ in the proof (instead of $(\frac{p}{2}+1)^{-\frac{p}{2}} = \frac{1}{9}$). ◀

Using the lemma above along with Lemma 18 in the proof of [6, Theorem 2.2], we can strenghten it as follows: letting $T \stackrel{\text{def}}{=} \frac{t^2}{c}$, for arbitrary $\delta > 0$ we fix a partition $A_1, \ldots, A_T$ of $\mathbb{N}$ such that $\|a\|_{Q(T)} \leq (1 + \delta) \sum_{j=1}^{T} \left(\sum_{k \in A_j} a_k^2 \right)^{1/2}$,

$$\Pr\left[\sum_{k=1}^{\infty} a_k X_k > \frac{1}{1+c} \kappa_a(t) \right] \geq \Pr\left[\sum_{k=1}^{\infty} a_k X_k > \frac{1}{\sqrt{1+c}} \|a\|_{Q(T)} \right] \quad \text{(by (7))}$$

$$\geq \Pr\left[\sum_{j=1}^{T} \sum_{k \in A_j} a_k X_k > \frac{1+\delta}{\sqrt{1+c}} \sum_{j=1}^{T} \left(\sum_{k \in A_j} a_k^2 \right)^{1/2} \right]$$

$$\geq \prod_{j=1}^{T} \Pr\left[\sum_{k \in A_j} a_k X_k > \frac{1+\delta}{\sqrt{1+c}} \left(\sum_{k \in A_j} a_k^2 \right)^{1/2} \right]$$

$$= \prod_{j=1}^{T} \frac{1}{2} \Pr\left[\left| \sum_{k \in A_j} a_k X_k \right|^2 > \left(\frac{1+\delta}{\sqrt{1+c}} \right)^2 \left(\sum_{k \in A_j} a_k^2 \right) \right]$$
$$\text{(symmetry)}$$

$$\geq \prod_{j=1}^{T} \frac{1}{6} \left(1 - \frac{(1+\delta)^2}{1+c} \right)^2. \quad \text{(Lemma 23)}$$

By taking the limit as $\delta \to 0^+$, we then obtain

$$\Pr\left[\sum_{k=1}^{\infty} a_k X_k > \frac{1}{1+c} \kappa_a(t) \right] \geq \left(\frac{1}{6} \left(1 - \frac{1}{1+c} \right)^2 \right)^T$$

$$= \left(\frac{c}{\sqrt{6}(1+c)} \right)^{\frac{2t^2}{c}} = e^{-\left(\frac{2}{c} \ln \frac{\sqrt{6}(1+c)}{c} \right) t^2}. \tag{12}$$

This takes care of the case where $\frac{t^2}{c}$ is an integer. If this is not the case, we consider $s \stackrel{\text{def}}{=} \sqrt{c\left(\lfloor \frac{t^2}{c} \rfloor + 1 \right)}$, so that $t^2 \leq s^2 \leq t^2 + c$. The monotonicity of κ_a then ensures that

$$\Pr\left[\sum_{k=1}^{\infty} a_k X_k > \frac{1}{1+c} \kappa_a(t) \right] \geq \Pr\left[\sum_{k=1}^{\infty} a_k X_k > \frac{1}{1+c} \kappa_a(s) \right]$$

$$\underset{(12)}{\geq} e^{-\left(\frac{2}{c} \ln \frac{\sqrt{6}(1+c)}{c} \right) s^2} \geq e^{-\left(\frac{2}{c} \ln \frac{\sqrt{6}(1+c)}{c} \right)(t^2+c)}$$

which concludes the proof. ◀

6.3   Some Examples

To gain intuition about the behavior of κ_a, we now compute tight asymptotic expressions for it in several instructive cases, specifically for some natural examples of probability distributions in $\Delta([n])$.

From the lower bound of Proposition 16 and the fact that $\kappa_p \leq \|p\|_1$ for any $p \in \ell_1$, it is clear that as soon as $t \geq \sqrt{n}$, $\kappa_p(t) = 1$ for any $p \in \Delta([n])$. It suffices then to consider the case $0 \leq t \leq \sqrt{n}$.

The uniform distribution.   We let p be the uniform distribution on $[n]$: $p_k = \frac{1}{n}$ for all $i \in [n]$. By considering a partition of $[n]$ into t^2 sets of size $\frac{n}{t^2}$, the lower bound of Proposition 16 yields $\kappa_p(t) \geq \|p\|_{Q(t^2)} \geq \frac{t}{\sqrt{n}}$. On the other hand, by definition $\kappa_p(t) = \inf_{p'+p''=p} \|p'\|_1 + t\|p''\|_2 \leq t\|p\|_2 = \frac{t}{\sqrt{n}}$, and thus

$$\kappa_p(t) = \begin{cases} \frac{t}{\sqrt{n}} & \text{if } t \leq \sqrt{n} \\ 1 & \text{if } t \geq \sqrt{n}. \end{cases}$$

We remark that in this case, the upper bound of Holmstedt from Proposition 15 only results in

$$\kappa_p(t) \leq \frac{t^2}{n} + t\sqrt{\frac{n-t^2}{n^2}} = f\left(\frac{t}{\sqrt{n}}\right)$$

where $f \colon x \in [0,1] \mapsto x^2 + x\sqrt{1-x^2}$. It is instructive to note this shows that this could not possibly have been the right upper bound (and therefore that Theorem 15 cannot be tight in general), as f is neither concave nor non-decreasing, and not even bounded by 1:
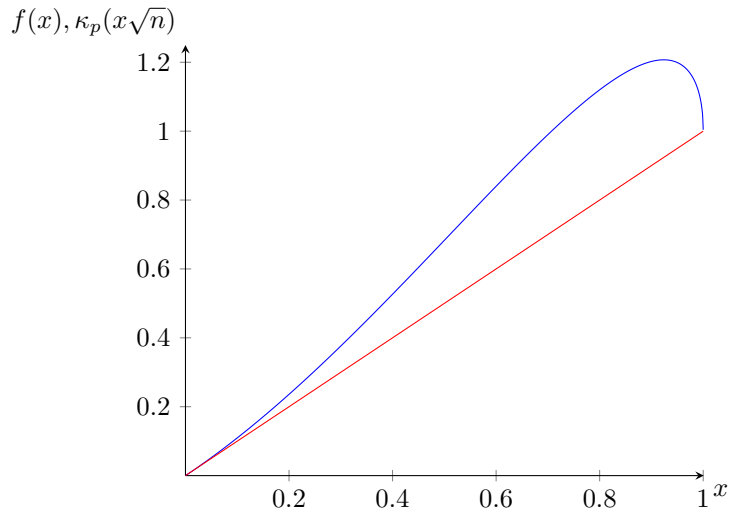
Figure 2 Holmstedt's upper bound (in blue) vs. true behavior of κ_p (in red).

From the above, we can now compare the behavior of $\kappa_p^{-1}(1 - 2\varepsilon)$ to the "2/3-norm functional" introduced by Valiant and Valiant [53]: for $\varepsilon \in (0, 1/2)$,

$$\kappa_p^{-1}(1 - 2\varepsilon) = (1 - 2\varepsilon)\sqrt{n}, \qquad \|p_{-\varepsilon}^{-\max}\|_{2/3} = (1 - \varepsilon)^{3/2}\sqrt{n} + o(1). \tag{13}$$

The Harmonic distribution. We now consider the case of the (truncated) Harmonic distribution, letting $p \in \Delta([n])$ be defined as $p_k = \frac{1}{kH_n}$ for all $i \in [n]$ (H_n being the n-th Harmonic number). By considering a partition of $[n]$ into $t^2 - 1$ sets of size 1 and one of size $n - t^2$, the lower bound of Proposition 16 yields

$$H_n \kappa_p(t) \geq \|p\|_{Q(t^2)} \geq \sum_{k=1}^{t^2-1} \frac{1}{k} + \sqrt{\sum_{k=t^2}^{n} \frac{1}{k^2}}$$

while Holmstedt's upper bound gives

$$H_n \kappa_p(t) \leq \sum_{k=1}^{t^2-1} \frac{1}{k} + t\sqrt{\sum_{k=t^2}^{n} \frac{1}{k^2}}.$$

For $t = O(1)$, this implies that $\kappa_p(t) = o(1)$; however, for $t = \omega(1)$ (but still less than $\sqrt{n}$), an asymptotic development of both upper and lower bounds shows that

$$\kappa_p(t) = \frac{2\ln t + O(1)}{\ln n}.$$

Using this expression, we can again compare the behavior of $\kappa_p^{-1}(1-2\varepsilon)$ to the 2/3-norm functional of [53]: for $\varepsilon \in (0, 1/2)$,

$$\kappa_p^{-1}(1-2\varepsilon) = \Theta\left(n^{\frac{1}{2}-\varepsilon}\right), \qquad \|p_{-\varepsilon}^{-\max}\|_{2/3} = \Theta\left(\frac{n^{\frac{1-\varepsilon}{2}}}{\log n}\right) = \Theta\left(n^{\frac{1-\varepsilon}{2}-o(1)}\right). \tag{14}$$

7 Identity Testing, revisited

For any $x \in (0, 1/2)$ and sequence $a \in \ell_1$, we let $t_x \stackrel{\text{def}}{=} \kappa_a^{-1}(1-2x)$, where κ_a is the K-functional of a as previously defined. Armed with the results and characterizations from the previous section, we will first in Section 7.1 describe an elegant reduction from communication complexity leading to a lower bound on instance-optimal identity testing parameterized by the quantity t_ε. Guided by this lower bound, we then will in Section 7.2 consider this result from the *upper bound* viewpoint, and in Theorem 30 establish that indeed this parameter captures the sample complexity of this problem. Finally, Section 7.3 is concerned with tightening our lower bound by using different arguments: specifically, showing that the bound that appeared naturally as a consequence of our communication complexity approach can, in hindsight, be established and slightly strenghtened with standard distribution testing arguments.

7.1 The Communication Complexity Lower Bound

In this subsection we prove the following lower bound on identity testing, via reduction from SMP communication complexity.

▶ **Theorem 24.** *Let Ω be a finite domain, and let $p = (p_1, \ldots, p_n) \in \Delta(\Omega)$ be a distribution, given as a parameter. Let $\varepsilon \in (0, 1/5)$, and set $t_\varepsilon \stackrel{\text{def}}{=} \kappa_p^{-1}(1-2\varepsilon)$. Then, given sample access to a distribution $q = (q_1, \ldots, q_n) \in \Delta(\Omega)$, testing $p = q$ versus $\|p - q\|_1 > \varepsilon$ requires $\Omega(t_\varepsilon/\log(n))$ samples from q.*

We will follow the argument outlined in Section 2.2: namely, applying the same overall idea as in the reduction for uniformity testing, but with an error-correcting code specifically designed for the distribution p instead of a standard Hamming one. To prove Theorem 24 we thus first need to define and obtain codes with properties that are tailored for our reduction; which we do next.

7.1.1 Balanced p-weighted codes

Recall that in our reductions so far, the first step is for Alice and Bob to apply a code to their inputs; typically, we chose that code to be a balanced code with constant rate, and linear distance *with respect to the uniform distribution* (i.e., with good Hamming distance). In order to obtain better bounds on a case-by-case basis, it will be useful to consider a generalization of these codes, under a different distribution:

▶ **Definition 25** (p-distance). For any $n \in \mathbb{N}$, given a probability distribution $p \in \Delta([n])$ we define the *p-distance* on $\{0,1\}^n$, denoted dist_p, as the weighted Hamming distance

$$\mathrm{dist}_p(x,y) \overset{\mathrm{def}}{=} \sum_{i=1}^{n} p(i) \cdot |x_i - y_i|$$

for $x, y \in \{0,1\}^n$. (In particular, this is a pseudometric on $\{0,1\}^n$.) The *p-weight* of $x \in \{0,1\}^n$ is given by $\mathrm{weight}_p(x) \overset{\mathrm{def}}{=} \mathrm{dist}_p(x, 0^n)$.

A p-weighted code is a code whose distance guarantee is with respect to the p-distance.

▶ **Definition 26** (p-weighted codes). Fix a probability distribution $p \in \Delta([n])$. We say that $C \colon \{0,1\}^k \to \{0,1\}^n$ is a (binary) *p-weighted code with relative distance $\gamma = \gamma(n)$ and rate $\rho = k/n$* if

$$\mathrm{dist}_p(C(x), C(y)) > \gamma$$

for all distinct $x, y \in \{0,1\}^k$.

Recall that the "vanilla" reduction in Section 5 relies on *balanced* codes. We generalize the balance property to the p-distance and allow the following relaxation.

▶ **Definition 27** (p-weighted τ-balance). A p-weighted code $C \colon \{0,1\}^k \to \{0,1\}^n$ is *τ-balanced* if there exists $\tau \in (0,1)$ such that $\mathrm{weight}_p(C(x)) \in \left(\frac{1}{2} - \tau, \frac{1}{2} + \tau\right)$ for all $x \in \{0,1\}^k$.

Now, for a distribution p, the volume of the p-ball in $\{0,1\}^n$ is given by

$$\mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\varepsilon) \overset{\mathrm{def}}{=} \left| \left\{ w \in \mathbb{F}_2^n : \mathrm{weight}_p(w) \leq \varepsilon \right\} \right|.$$

Next, we show that there exist nearly balanced p-weighted codes with constant relative distance nearly optimal rate.

▶ **Proposition 28** (Existence of nearly balanced p-weighted codes). *Fix a probability distribution $p \in \Delta([n])$, constants $\gamma, \tau \in \left(0, \frac{1}{3}\right)$, and $\varepsilon = \max\{\gamma, \frac{1}{2} - \tau\}$. There exists a p-weighted τ-balanced code $C \colon \{0,1\}^k \to \{0,1\}^n$ with relative distance γ such that $k = \Omega(n - \log \mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\varepsilon))$.*

In contrast, by the sphere packing bound, every p-weighted code $C \colon \{0,1\}^k \to \{0,1\}^n$ with distance γ satisfies

$$\underbrace{2^k}_{\#\text{codewords}} \leq \frac{2^n}{\mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\gamma/2)}.$$

Hence, we have $k \leq n - \log \mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\gamma/2)$.

Proof of Theorem 28. Note that

$$\mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\varepsilon) = \left| \left\{ w \in \mathbb{F}_2^n : \mathrm{weight}_p(w) \leq \varepsilon \right\} \right| = 2^n \cdot \Pr_{w \sim \{0,1\}^n} \left[\sum_{i=1}^n p_i w_i \leq \varepsilon \right].$$

The probability that a randomly chosen code $C : \{0,1\}^k \to \{0,1\}^n$ does *not* have distance γ is

$$\Pr_C \left[\exists x, y \in \{0,1\}^k \text{ such that } \mathrm{dist}_p(C(x), C(y)) \leq \gamma \right] \leq 2^{2k} \cdot \Pr_{w,w' \sim \{0,1\}^n} \left[\mathrm{dist}_p(w, w') \leq \gamma \right]$$

$$\leq 2^{2k} \cdot \Pr_{w \sim \{0,1\}^n} \left[\sum_{i=1}^n p_i w_i \leq \varepsilon \right]$$

$$= \frac{\mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\varepsilon)}{2^{n-2k}}.$$

Hence, for sufficiently small $k = \Omega(n - \log \mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\varepsilon))$, the probability that a random code is a p-weighted code with relative distance γ is at least $2/3$; fix such k. Similarly, the probability that a random code $C : \{0,1\}^k \to \{0,1\}^n$ is not τ-balanced (under the p-distance) is

$$\Pr_C \left[\exists x \in \{0,1\}^k \text{ such that } \mathrm{weight}_p(C(x)) \notin \left(\frac{1}{2} - \tau, \frac{1}{2} + \tau \right) \right]$$

$$\leq 2^k \cdot \Pr_{w \in \{0,1\}^n} \left[\left| \mathrm{weight}_p(w) - \frac{1}{2} \right| > \tau \right]$$

$$\leq 2^{k+1} \cdot \Pr_{w \in \{0,1\}^n} \left[\sum_{i=1}^n p_i w_i < \varepsilon \right]$$

$$\leq \frac{\mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\varepsilon)}{2^{n-k-1}}.$$

Thus, the probability that a random code is τ-balanced (under the p-distance) is at least $2/3$, and so, with probability at least $\frac{1}{3}$, a random code satisfies the proposition's hypothesis. ◀

We now establish a connection between the rate of p-weighted codes and the K-functional of p, as introduced in Section 6:

▶ **Claim 29.** *Let $p \in \Delta([n])$ be a probability distribution. Then, for any $\gamma \in (0, \frac{1}{2})$ we have*

$$n - \log \mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\gamma) \geq \frac{1}{2 \ln 2} \kappa_p^{-1}(1 - 2\gamma)^2$$

where $\kappa_p^{-1}(u) = \inf \{ t \in (0, \infty) : \kappa_a(t) \geq u \}$ for $u \in [0, \infty)$.

Proof. From the definition,

$$\mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\gamma) = \left| \left\{ w \in \mathbb{F}_2^n : \mathrm{weight}_p(w) \leq \gamma \right\} \right| = \left| \left\{ w \in \mathbb{F}_2^n : \sum_{i=1}^n p_i w_i \leq \gamma \right\} \right|$$

$$= 2^n \Pr_{Y \sim \{0,1\}^n} \left[\sum_{i=1}^n p_i Y_i \leq \gamma \right] = 2^n \Pr_{X \sim \{-1,1\}^n} \left[\sum_{i=1}^n p_i X_i \geq 1 - 2\gamma \right]$$

$$= 2^n \Pr_{X \sim \{-1,1\}^n} \left[\sum_{i=1}^n p_i X_i \geq \kappa_p(u_\gamma) \right]$$

where we set $u_\gamma \overset{\text{def}}{=} \kappa_p^{-1}(1 - 2\gamma)$. From Theorem 22, we then get $\mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\gamma) \leq 2^n e^{-\frac{u_\gamma^2}{2}}$, from which

$$n - \log \mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\gamma) \geq -\log e^{-\frac{u_\gamma^2}{2}} = \frac{1}{2 \ln 2} u_\gamma^2$$

as claimed.   ◀

7.1.2 The Reduction

Equipped with the nearly balanced p-weighted codes in Theorem 28, we are ready to prove Theorem 24. Assume there exists an s-sample ε-tester for identity to p, with error probability $1/6$, and assume, without loss of generality, that ε is a constant (independent of n).

Fix $\gamma = \varepsilon$ and $\tau = (1 - 2\varepsilon)/2$. For a sufficiently large $k \in \mathbb{N}$, let $C \colon \{0,1\}^k \to \{0,1\}^n$ be a τ-balanced p-weighted code with relative distance γ, as guaranteed by Theorem 28; namely, the code C satisfies the following conditions.
1. *Balance*: $\mathrm{weight}_p(C(x)) \in \left(\frac{1}{2} - \tau, \frac{1}{2} + \tau\right)$ for all $x \in \{0,1\}^k$;
2. *Distance*: $\mathrm{dist}_p(C(x), C(y)) > \gamma$ for all distinct $x, y \in \{0,1\}^k$;
3. *Rate*: $k = \Omega(n - \log \mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\varepsilon))$.

We reduce from the problem of equality in the (private coin) SMP model. Given their respective inputs $x, y \in \{0,1\}^k \times \{0,1\}^k$ from EQ_k, Alice and Bob separately create inputs $(a, b) = (C(x), C(y)) \in \{0,1\}^n \times \{0,1\}^n$. Let $A \subseteq [n]$ denote the set indicated by a, and let $B \subseteq [n]$ denote the set indicated by $\bar{b}$. Alice and Bob then each send to the referee the p-weight of their encoded input, $\mathrm{weight}_p(a) = p(A)$ and $\mathrm{weight}_p(\bar{b}) = p(B)$ respectively,[12] as well as a sequence of $6cs$ samples independently drawn from the distribution p *restricted* to the subsets A and B respectively, where c is the constant such that $\frac{1}{c} p(B) \cdot \leq p(A) \leq c \cdot p(B)$, guaranteed by the balance property of C. Finally, the referee checks that $p(A) + p(B) = 1$ (and otherwise rejects) and generates a sequence of q samples by choosing independently, for each of them, Alice's element with probability $p(A)$ and Bob's with probability $p(B)$, and feeds these samples to the ε-tester for identity to p.

By Markov's inequality, the above procedure indeed allows the referee to retrieve, with probability at least $1 - \frac{cs}{6cs} = \frac{5}{6}$, at least s independent samples from the distribution

$$q \overset{\text{def}}{=} p(A) \cdot p|_A + p(B) \cdot p|_B,$$

at the cost of $O(s \log n)$ bits of communication in total.

For correctness, note that if $x = y$, then $A = \bar{B}$, which implies $q = p$. On the other hand, if $x \neq y$, by the (p-weighted) distance of C we have $\mathrm{dist}_p(C(x), C(y)) > \gamma$, and so $p(A \cap B) + p(\overline{A \cup B}) > \gamma$. Note that every $i \in A \cap B$ satisfies $q_i = 2p_i$ and every $i \in \overline{A \cup B}$ is not supported in q. Therefore, we have $\|p - q\|_1 > \varepsilon$. The referee can therefore invoke the identity testing algorithm to distinguish between p and q with probability $1 - \left(\frac{1}{6} + \frac{1}{6}\right) = \frac{2}{3}$. This implies that the number of samples q used by any such tester must satisfy $s \log n = \Omega\left(\sqrt{k}\right)$. Finally, by Claim 29 we have

$$k = \Omega\big(n - \log \mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\varepsilon)\big) = \Omega\big(\kappa_p^{-1}(1 - 2\varepsilon)^2\big),$$

and therefore we obtain a lower bound of $s = \Omega(t_\varepsilon / \log(n))$.

---

[12] A standard argument shows it suffices to specify $p(A)$ and $p(B)$ with precision roughly $1/n^2$, and so sending the weights only costs $O(\log n)$ bits.

7.2 The Upper Bound

Inspired by the results of the previous section, it is natural to wonder whether the dependence on t_ε of the lower bound is the "right" one. Our next theorem shows that this is the case: the parameter t_ε does, in fact, capture the sample complexity of the problem.

▶ **Theorem 30.** *There exists an absolute constant $c > 0$ such that the following holds. Given any fixed distribution $p \in \Delta([n])$ and parameter $\varepsilon \in (0, 1]$, and granted sample access to an unknown distribution $q \in \Delta([n])$, one can test $p = q$ vs. $\|p - q\|_1 > \varepsilon$ with $O\big(\max\big(\frac{t_{c\varepsilon}}{\varepsilon^2}, \frac{1}{\varepsilon}\big)\big)$ samples from q. (Moreover, one can take $c = \frac{1}{18}$).*

7.2.1 High-level idea

As discussed in Section 2.4, the starting point of the proof is the connection between the K-functional and the "Q-norm" obtained in Lemma 18: indeed, this result ensures that for $T = 2t_{O(\varepsilon)}^2$, there exists a partition of the domain into sets $A_1, \ldots, A_T$ such that

$$1 - O(\varepsilon) \le \|p\|_{Q(T)} = \sum_{j=1}^T \sqrt{\sum_{i \in A_j} p_i^2} = \sum_{j=1}^T \|p_{A_j}\|_2$$

where p_{A_j} is the restriction of the sequence p to the indices in A_j. But by the monotonicity of ℓ_p norms, we know that $\sum_{j=1}^T \|p_{A_j}\|_2 \le \sum_{j=1}^T \|p_{A_j}\|_1 = \sum_{j=1}^T \sum_{i \in A_j} p_i = \|p\|_1 = 1$. Therefore, what we obtain is in fact that

$$0 \le \sum_{j=1}^T (\underbrace{\|p_{A_j}\|_1 - \|p_{A_j}\|_2}_{\ge 0}) \le O(\varepsilon).$$

Now, if the right-hand side were *exactly* 0, then this would imply $\|p_{A_j}\|_1 = \|p_{A_j}\|_2$ for all j, and thus that p has (at most) one non-zero element in each A_j. Therefore, testing identity to p would boil down to testing identity on a distribution with support size T, which can be done with $O(\sqrt{T}/\varepsilon^2)$ samples.

This is not actually the case, of course: the right-hand-side is only small and not exactly zero. Yet, one can show that a robust version of the above holds, making this intuition precise: in Lemma 31, we show that on average, *most* of the probability mass of p is concentrated on a single point from each A_j. This sparsity implies that testing identity to p on this set of T points is indeed enough – leading to the theorem.

7.2.2 Proof of Theorem 30

Let $p \in \Delta([n])$ be a fixed, known distribution, assumed monotone non-increasing without loss of generality: $p_1 \ge p_2 \ge \cdots \ge p_n$. Given $\varepsilon \in (0, 1/2)$, we let t_ε be as above, namely such that

$$\kappa_p(t_\varepsilon) \ge 1 - 2\varepsilon.$$

From this, it follows by Lemma 18 that

$$\|p\|_{Q(T)} \ge 1 - 2\varepsilon, \tag{15}$$

where we set $T \overset{\text{def}}{=} 2t_\varepsilon^2$. Choose $A_1, \ldots, A_T$ to be a partition of $[n]$ achieving the maximum (since we are in the finite, discrete case) defining $\|p\|_{Q(T)}$; and let $\tilde{p}$ be the subdistribution

on T elements defined as follows. For each $j \in [T]$, choose $i_j \stackrel{\text{def}}{=} \arg\max_{i \in A_j} p_i$, and set $\tilde{p}(j) \stackrel{\text{def}}{=} p(i_j)$.

▶ **Lemma 31** (Sparsity Lemma). *There exists an absolute constant $\kappa > 0$ such that $\tilde{p}([T]) = \sum_{j=1}^T p(i_j) \geq 1 - \kappa\varepsilon$. (Moreover, one can take $\kappa \stackrel{\text{def}}{=} \frac{2}{3-\sqrt{7}} \simeq 5.65$.)*

Proof. Fix any $j \in [T]$, and for convenience let $A \stackrel{\text{def}}{=} A_j$. Write a^* for the maximum element for p in A, so that $p(i_j) = \max_{a \in A} p(a) = p(a^*)$. We have by monotonicity $p(A) \geq \sqrt{\sum_{a \in A} p(a)^2}$, and moreover, letting $\alpha \stackrel{\text{def}}{=} p(A) - p(a^*) = p(A \setminus \{a^*\})$,

$$p(A) - \sqrt{\sum_{a \in A} p(a)^2} = p(a^*) + \alpha - \sqrt{p(a^*)^2 + \sum_{a \neq a^*} p(a)^2} \geq p(a^*) + \alpha - \sqrt{p(a^*)^2 + \alpha^2}.$$

We let $s > 1$ be a (non-integer) parameter to be chosen later. Suppose first that $\alpha \leq \frac{s}{s+1}p(A)$, or equivalenty $\alpha \leq sp(a^*)$. In that case, we have

$$p(A) - \sqrt{\sum_{a \in A} p(a)^2} \geq p(a^*) + \alpha - p(a^*)\sqrt{1 + \left(\frac{\alpha}{p(a^*)}\right)^2}$$

$$\geq p(a^*) + \alpha - p(a^*)\left(1 + \frac{\sqrt{s^2+1}-1}{s}\frac{\alpha}{p(a^*)}\right)$$

$$= \left(1 - \frac{\sqrt{s^2+1}-1}{s}\right)\alpha \stackrel{\text{def}}{=} L_1(s)\alpha$$

where we relied on the inequality $\sqrt{1+x^2} \leq 1 + \frac{\sqrt{s^2+1}-1}{s}x$ for $x \in [0, s]$. However, if $\alpha > sp(a^*)$, then we have

$$p(A) - \sqrt{\sum_{a \in A} p(a)^2} = p(a^*) + \alpha - \sqrt{p(a^*)^2 + \sum_{a \neq a^*} p(a)^2} \geq \alpha - \sqrt{\sum_{a \neq a^*} p(a)^2}$$

$$\geq \alpha - \sqrt{\lfloor s \rfloor \left(\frac{\alpha}{s}\right)^2 + 1 \cdot \left(\alpha - \frac{\lfloor s \rfloor}{s}\alpha\right)^2}$$

$$= \left(1 - \sqrt{\frac{\lfloor s \rfloor}{s^2} + \left(1 - \frac{\lfloor s \rfloor}{s}\right)^2}\right)\alpha \stackrel{\text{def}}{=} L_2(s)\alpha.$$

using the fact that $p(a^*)$ is the maximum probability value of any element, so that the total α has to be spread among at least $\lfloor s \rfloor + 1$ elements (recall that s will be chosen not to be an integer). Optimizing these two bounds leads to the choice of $s \stackrel{\text{def}}{=} \frac{4+\sqrt{7}}{3} \notin \mathbb{N}$, for which $L_1(s) = L_2(s) = 3 - \sqrt{7} \simeq 0.35$.

Putting it together, we obtain, summing over all $j \in [T]$, that

$$1 - \|p\|_{Q(T)} = \sum_{j=1}^T p(A_j) - \sum_{j=1}^T \sqrt{\sum_{i \in A_j} p(i)^2} = \sum_{j=1}^T \left(p(A_j) - \sqrt{\sum_{i \in A_j} p(i)^2}\right)$$

$$\geq (3 - \sqrt{7}) \sum_{j=1}^T (p(A_j) - p(i_j))$$

$$= (3 - \sqrt{7})(1 - \tilde{p}([T]))$$

which implies $\tilde{p}([T]) \geq \frac{1}{3-\sqrt{7}}\|p\|_{Q(T)} - \frac{1}{3-\sqrt{7}} + 1 \geq 1 - \frac{2}{3-\sqrt{7}}\varepsilon$ by Eq. (15).   ◀

▶ **Lemma 32.** *Fix p, ε as above, let $S \stackrel{\text{def}}{=} \{i_1, \ldots, i_T\}$ be the corresponding set of T elements, and take κ as in Lemma 31. For any $q \in \Delta([n])$, if (i) $\sum_{j=1}^{T} q(i_j) \geq 1 - (\kappa + \frac{1}{3})\varepsilon$ and (ii) $\sum_{j=1}^{T} \left| \frac{\tilde{p}(j)}{p(S)} - \frac{\tilde{q}(j)}{q(S)} \right| \leq \frac{1}{3}\varepsilon$, then $\|p - q\|_1 \leq (3\kappa + 1)\varepsilon$.*

Proof. Unrolling the definition, and as $p(\bar{S}) \leq \kappa\varepsilon$ by Lemma 31,

$$
\begin{aligned}
\|p - q\|_1 &= \sum_{i=1}^{n} |p(i) - q(i)| = \sum_{j=1}^{T} |p(i_j) - q(i_j)| + \sum_{i \notin S} |p(i) - q(i)| \\
&\leq \sum_{j=1}^{T} |p(i_j) - q(i_j)| + p(\bar{S}) + q(\bar{S}) \\
&\leq \sum_{j=1}^{T} |p(i_j) - q(i_j)| + \kappa\varepsilon + (\kappa + \tfrac{1}{3})\varepsilon = \sum_{j=1}^{T} \left| p(S)\frac{\tilde{p}(j)}{p(S)} - q(S)\frac{\tilde{q}(j)}{q(S)} \right| + (2\kappa + \tfrac{1}{3})\varepsilon \\
&\leq p(S) \sum_{j=1}^{T} \left| \frac{\tilde{p}(j)}{p(S)} - \frac{\tilde{q}(j)}{q(S)} \right| + \sum_{j=1}^{T} \frac{\tilde{q}(j)}{q(S)} |p(S) - q(S)| + (2\kappa + \tfrac{1}{3})\varepsilon \\
&= p(S) \cdot \sum_{j=1}^{T} \left| \frac{\tilde{p}(j)}{p(S)} - \frac{\tilde{q}(j)}{q(S)} \right| + |p(S) - q(S)| + (2\kappa + \tfrac{1}{3})\varepsilon \\
&\leq \tfrac{1}{3}\varepsilon + (\kappa + \tfrac{1}{3})\varepsilon + (2\kappa + \tfrac{1}{3})\varepsilon = (3\kappa + 1)\varepsilon
\end{aligned}
$$

concluding the proof of the lemma.   ◀

Let $\kappa > 0$ be the constant from Lemma 31. We let $\varepsilon' \stackrel{\text{def}}{=} \frac{\varepsilon}{3\kappa+1}$, and $T \stackrel{\text{def}}{=} 2t_{\varepsilon'}^2$, $\{i_1, \ldots, i_T\} \subseteq [n]$ the corresponding value and elements (i.e., T and the i_j's are as in the foregoing discussion (chosen with regard to ε' and the known distribution p)). For convenience, denote by $\tilde{q}$ the (unknown) subdistribution on $[T]$ defined by $\tilde{q}(j) \stackrel{\text{def}}{=} q(i_j)$ for $j \in [T]$.

We first verify that $\tilde{q}([T]) \geq 1 - \kappa\varepsilon'$, with $O(1/\varepsilon')$ samples (specifically, we distinguish, with probability at least $9/10$, between $\tilde{q}([T]) \geq 1 - \kappa\varepsilon'$ and $\tilde{q}([T]) \leq 1 - (\kappa + \frac{1}{3})\varepsilon'$; and reject in the latter case). Once this is done, we apply one of the known identity testing algorithms to $\bar{p}, \bar{q} \in \Delta([T])$, renormalized versions of $\tilde{p}, \tilde{q}$:

$$
\bar{p} = \frac{\tilde{p}}{\tilde{p}([T])}, \quad \bar{q} = \frac{\tilde{q}}{\tilde{q}([T])}
$$

using rejection sampling (note that we have the explicit description of $\bar{p}$; and, since $\tilde{q}([T]) \geq 1 - (\kappa + \frac{1}{3})\varepsilon'$ (conditioning on the first test meeting its guarantee), we can obtain m independent samples from $\bar{q}$ with an expected $O(m)$ number of samples from q). This is done with parameter ε' and failure probability $1/10$; and costs $O\left(\frac{\sqrt{T}}{\varepsilon'^2}\right) = O\left(\frac{t_{\varepsilon'}}{\varepsilon'^2}\right)$ samples from q.

Turning to the correctness: we condition on both tests meeting their guarantees, which by a union bound holds with probability at least $4/5$.

- If $p = q$, then $q(S) = p(S) \geq 1 - \kappa\varepsilon'$, and $\bar{q} = \bar{p}$: neither the first nor the second test reject, and the overall algorithm accepts.
- If the algorithm accepts, then $q(S) \geq 1 - (\kappa + \frac{1}{3})\varepsilon'$ (by the first test) and $\sum_{j=1}^{T} \left| \frac{\tilde{p}(j)}{p(S)} - \frac{\tilde{q}(j)}{p(S)} \right| \leq \varepsilon'$ (by the second): Lemma 32 then guarantees that $\|p - q\|_1 \leq 3\kappa + 1\varepsilon' = \varepsilon$.

Observing that for $\kappa = \frac{2}{3-\sqrt{7}}$ (as suggested by Lemma 31) we have $3\kappa + 1 \leq 18$ establishes the last part of the theorem.

7.3 Tightening the Lower Bound

As a last step, one may want to strenghten the lower bound obtained by the communication complexity reduction of Theorem 24. We here describe how this can be achieved using more standard arguments from distribution testing. However, we stress that these arguments in some sense are applicable "'after the fact," that is after Section 7.1 revealed the connection to the K-functional, and the bound we should aim for. Specifically, we prove the following:

▶ **Theorem 33.** *For any $p \in \Delta([n])$, and any $\varepsilon \in (0, 1/2)$ any algorithm testing identity to p must have sample complexity $\Omega\left(\frac{t_\varepsilon}{\varepsilon}\right)$.*

Proof. Fix $p \in \Delta([n])$ and $\varepsilon \in (0, 1/2)$ as above, and consider the corresponding value t_ε; we assume that $t_\varepsilon \geq 2$, as otherwise there is nothing to prove.[13] Without loss of generality – as we could always consider a sufficiently small approximation, and take the limit in the end, we further assume the infimum defining κ_p is attained: let $h, \ell \in [0, 1]^n$ be such that $p = h + \ell$ and $\kappa_p(t_\varepsilon) = \|h\|_1 + t_\varepsilon \|\ell\|_2 = 1 - 2\varepsilon$.

Since $\|\ell\|_1 = 1 - \|h\|_1$, from the definition of h, ℓ, we have that $1 - 2\varepsilon = 1 - \|\ell\|_1 + t_\varepsilon \|\ell\|_2$, from which

$$0 < \|\ell\|_2 = \frac{\|\ell\|_1 - 2\varepsilon}{t_\varepsilon} \leq \frac{1}{t_\varepsilon} \tag{16}$$

(note that the right inequality is strict because $\varepsilon > 0$: since if $\|\ell\|_2 = 0$, then $\|\ell\|_1 = 0$ and $h = p$; but then $\kappa_{t_\varepsilon} = \|p\|_1 = 1$.) In particular, this implies $\|\ell\|_1 - 2\varepsilon > 0$.

With this in hand, we will apply the following theorem, due to Valiant and Valiant:

▶ **Theorem 34** ([53, Theorem 4]). *Given a distribution $p \in \Delta([n])$, and associated values $(\varepsilon_i)_{i \in [n]}$ such that $\varepsilon_i \in [0, p_i]$ for each i, define the distribution over distributions $\mathcal{Q}$ by the process: independently for each domain element i, set uniformly at random $q_i = p_i \pm \varepsilon_i$, and then normalize q to be a distribution. Then there exists a constant $c > 0$ such that is takes at least $c\left(\sum_{i=1}^n \varepsilon_i^4/p_i^2\right)^{-1/2}$ samples to distinguish p from $\mathcal{Q}$ with success probability $2/3$. Further, with probability at least $1/2$ the ℓ_1 distance between p and a uniformly random distribution from $\mathcal{Q}$ is at least $\min\left(\sum_{i=1}^n \varepsilon_i - \max_i \varepsilon_i, \frac{1}{2} \sum_{i=1}^n \varepsilon_i\right)$.*

We want to invoke the above theorem with ℓ being, roughly speaking, the "random perturbation" to p. Indeed, since ℓ has small ℓ_2 norm of order $O(1/t_\varepsilon)$ by (16) (which gives a good lower bound) and has ℓ_1 sum $\Omega(\varepsilon)$ (which gives distance), this seems to be a natural choice.

In view of this, set $\alpha \overset{\text{def}}{=} \frac{2\varepsilon}{\|\ell\|_1} \in (0, 1)$ and, for $i \in [n]$, $\varepsilon_i \overset{\text{def}}{=} \alpha \ell_i \leq \ell_i \in [0, p_i]$. Theorem 33 will then be a direct consequence of the next two claims:

▶ **Claim 35** (Distance). *We have $\min\left(\sum_{i=1}^n \varepsilon_i - \max_i \varepsilon_i, \frac{1}{2} \sum_{i=1}^n \varepsilon_i\right) \geq \varepsilon$.*

Proof. Since by our choice of α it is immediate that $\sum_{i=1}^n \varepsilon_i = \frac{2\varepsilon}{\|\ell\|_1} \sum_{i=1}^n \ell_i = 2\varepsilon$, it suffices to show that $\max_i \varepsilon_i \leq \varepsilon$, or equivalently that $\max_i \ell_i \leq \frac{1}{2}\|\ell\|_1$. But this follows from the fact that $\|\ell\|_\infty \leq \|\ell\|_2 \leq \frac{\|\ell\|_1}{t_\varepsilon}$, and our assumption that $t_\varepsilon \geq 2$.   ◀

It then remains to analyze the lower bound obtained through the application of Theorem 34:

▶ **Claim 36** (Lower bound). *With the ε_i's defined as before, $\left(\sum_{i=1}^n \varepsilon_i^4/p_i^2\right)^{-1/2} \geq \frac{2t_\varepsilon}{\varepsilon}$.*

---

[13] Indeed, an immediate lower bound of $\Omega(1/\varepsilon)$ on this problem holds.

Proof. Unrolling the definition of the ε_i's,

$$\sum_{i=1}^n \frac{\varepsilon_i^4}{p_i^2} = \alpha^4 \sum_{i=1}^n \frac{\ell_i^4}{p_i^2} = \alpha^4 \sum_{i=1}^n \frac{\ell_i^2}{p_i^2}\ell_i^2 \leq \alpha^4 \sum_{i=1}^n \ell_i^2 = \frac{2^4 \varepsilon^4}{\|\ell\|_1^4}\|\ell\|_2^2 = \left(\frac{4\varepsilon^2}{\|\ell\|_1^2}\frac{\|\ell\|_1 - 2\varepsilon}{t_\varepsilon}\right)^2$$

where the last equality is (16). This yields

$$\left(\sum_{i=1}^n \frac{\varepsilon_i^4}{p_i^2}\right)^{-1/2} \geq \frac{t_\varepsilon}{4\varepsilon^2} \cdot \frac{\|\ell\|_1^2}{\|\ell\|_1 - 2\varepsilon} = \frac{t_\varepsilon}{2\varepsilon} \cdot \frac{\left(\frac{\|\ell\|_1}{2\varepsilon}\right)^2}{\frac{\|\ell\|_1}{2\varepsilon} - 1} \geq \frac{2t_\varepsilon}{\varepsilon}$$

where the last inequality comes from $f\colon x > 1 \mapsto \frac{x^2}{x-1}$ achieving its minimum, 4, at $x = 2$. ◄

Combining the two claims with Theorem 34 implies, by a standard argument, the lower bound of Theorem 33. ◄

▶ Remark. A straightforward modification of the proof of Theorem 33 allows one to prove a somewhat more general statement, namely a lower bound of $\Omega\big(\gamma t_\gamma/\varepsilon^2\big)$ for any $\gamma \in [\varepsilon, 1/2]$ such that $t_\gamma \geq 2$. In particular, this implies an incomparable bound of $\Omega\big(t_{1/4}/\varepsilon^2\big)$ as long as p does not put almost all its probability weight on $O(1)$ elements.

On the optimality of our bound. We conclude this section by briefly discussing the optimality of our bound, and specifically whether one could hope to strenghten Theorem 33 to obtain an $\Omega\big(t_\varepsilon/\varepsilon^2\big)$ lower bound. Unfortunately, it is easy to come up with simple (albeit contrived) counterexamples: e.g., fix $\varepsilon \in (0, 1/3)$, and let $p \in \Delta([n])$ be the distribution that puts mass $1 - 3\varepsilon$ on the first element and uniformly spreads the rest among the remaining $n - 1$ elements. A straightforward calculation shows that, for this distribution $p = p(\varepsilon)$, one has $\kappa_p^{-1}(1 - 2\varepsilon) = \Theta(\sqrt{n})$; and it is not hard to check that one can indeed test identity to p with $O(\sqrt{n}/\varepsilon)$ samples only,[14] and so the $\Omega(t_\varepsilon/\varepsilon)$ lower bound is tight in this case.

Although this specific instance is somewhat unnatural, as it fails to be a counterexample for any distance parameter $\varepsilon' \ll \varepsilon$, it does rule out an improvement of Theorem 33 for the full range of parameters. On the other hand, it is also immediate to see that the upper bound $O\big(t_\varepsilon/\varepsilon^2\big)$ cannot be improved in general, as demonstrated by choosing p to be the uniform distribution (yet, in this case, the extension provided by the above remark does provide the optimal bound).

8 Lower Bounds on Other Properties

In this section we demonstrate how our methodology can be used to easily obtain lower bounds on the sample complexity of various properties of distributions. To this end, we provide sketches of proofs of lower bounds for monotonicity testing, k-modality, and the "symmetric sparse support" property (that we define below). We remark that using minor variations on the reductions presented in Section 5 and Section 7, it is also straightforward to obtain lower bounds for properties of distributions such as being binomially distributed, Poisson binomially distributed, and having a log-concave probability mass function. Throughout this section, we fix ε to be a small constant and refer to testing with respect to proximity $\Theta(\varepsilon)$.

---

[14] Indeed, any distribution q such that $\|q - p\|_1 > \varepsilon$ must either be such that $|p(1) - q(1)| = \Omega(\varepsilon)$ or $\big|p|_{[n]\setminus\{1\}} - q|_{[n]\setminus\{1\}}\big| = \Omega(1)$. The first case only takes $O(1/\varepsilon)$ samples, while the second can be achieved by rejection sampling with $O(1/\varepsilon) \cdot O(\sqrt{n})$ samples.

Monotonicity on the integer line and the Boolean hypercube.   We start with the problem of testing monotonicity on the integer line, that is, testing whether a distribution $p \in \Delta([n])$ has a monotone probability mass function. Consider the "vanilla" reduction, presented in Section 5. Note that for yes-instances, we obtain the uniform distribution, which is monotone. For no-instances, however, we obtain a distribution p that has mass $1/n$ on a $(1 - \varepsilon)$-fraction of the domain, is unsupported on a $(\varepsilon/2)$-fraction of the domain, and has mass $2/n$ on the remaining $(\varepsilon/2)$-fraction. Typically, p is $\Omega(1)$-far from being monotone; however, it could be the case that the first (respectively, last) $\varepsilon n/2$ elements are of 0 mass, and the last (respectively, first) $\varepsilon n/2$ elements are of mass $2/n$, in which case p is perfectly monotone. To remedy this, all we have to do is let the referee emulate a distribution $p' \in \Delta([3n])$ such that $p'_i = \begin{cases} \frac{1}{3}p_{i-n} & i \in \{n+1, \ldots, 2n\} \\ \frac{1}{3n} & \text{otherwise} \end{cases}$. It is immediate to see that the probability mass functions of p' is $(\varepsilon/3)$-far from monotone.

The idea above can be extended to monotonicity over the hypercube as follows. We start with the uniformity reduction, this time over the domain $\{0,1\}^n$. As before, yes-instances will be mapped to the uniform distribution over the hypercube, which is monotone, and no-instances will be mapped to a distribution that has mass $1/2^n$ on a $(1 - \varepsilon)$-fraction of the domain, is unsupported on a $(\varepsilon/2)$-fraction of the domain, and has mass $1/2^{n-1}$ on the remaining $(\varepsilon/2)$-fraction – but could potentially be monotonously *strictly* increasing (or decreasing). This time, however, the "boundary" is larger than the "edges" of the integer line, and we cannot afford to pad it with elements of weight $1/2^n$. Instead, the referee, who receives for the players samples drawn from a distribution $p \in \Delta(\{0,1\}^n)$, emulates a distribution $p'' \in \Delta(\{0,1\}^{n+1})$ over a larger hypercube whose additional coordinate determines between a negated or regular copy of p; that is, $p'(z) = \begin{cases} p(z) & z_1 = 0 \\ \frac{1}{2^{n-1}} - p(z) & z_1 = 1 \end{cases}$ (where the referee chooses $z_1 \in \{0,1\}$ independently and uniformly at random for each new sample). Hence, even if p is monotonously increasing (or decreasing), the emulated distribution p'' is $\Omega(\varepsilon)$-far from monotone. By the above, we obtain $\tilde{\Omega}(\sqrt{n})$ and $\tilde{\Omega}(2^{n/2})$ lower bounds on the sample complexity of testing monotonicity on the line and on the hypercube, respectively.

***k*-modality.**   Recall that a distribution $p \in \Delta([n])$ is said to be *k-modal* if its probability mass function has at most k "peaks" and "valleys." Such distributions are natural generalizations of monotone (for $k = 0$) and unimodal (for $k = 1$) distributions. Fix a sublinear k, and consider the uniformity reduction presented in Section 5, with the additional step of letting the prover apply a random permutation to the domain $[n]$ (similarly to the reduction shown in Section 5.1). Note that yes-instances are still mapped to the uniform distribution (which is clearly k-modal), and no-instances are mapped to distributions with mass $1/n$, $2/n$, and 0 on a $(1 - \varepsilon)$, $(\varepsilon/2)$, and $(\varepsilon/2)$ (respectively) fractions of the domain. Intuitively, applying a random permutation of the domain to such a distribution "spreads" the elements with masses 0 and $2/n$ nearly uniformly, causing many level changes (i.e., high modality); indeed, it is straightforward to verify that with high probability over the choice of a random permutation of the domain, such a distribution will indeed be $\Omega(\varepsilon)$-far from k-modal. This yields an $\tilde{\Omega}(\sqrt{n})$ lower bound on the sample complexity of testing k-modality, nearly matching the best known lower bound of $\Omega(\max(\sqrt{n}, k/\log k))$ following from [19], for $k/\log(k) = O(\sqrt{n})$.

Symmetric sparse support.   Consider the property of distributions $p \in \Delta([n])$ such that when projected to its support, p is mirrored around the middle of the domain. That is, p is said to have a *symmetric sparse support* if there exists $S = \{i_0 < i_2 < \cdots < i_{2\ell}\} \subseteq [n]$ with

$i_\ell = \frac{n}{2}$ such that: (1) $p(i) = 0$ for all $i \in [n] \setminus S$, and (2) $p(i_{\ell+1-j}) = p(i_{\ell+j})$ for all $0 \le j \le \ell$. We sketch a proof of an $\tilde{\Omega}(\sqrt{n})$ lower bound on the sample complexity of testing this property. Once again, we shall begin with the uniformity reduction presented in Section 5, obtaining samples from a distribution $p \in \Delta([n/2])$. Then the referee emulates samples from the distribution $p' \in \Delta([n])$ that is distributed as p on its left half, and uniformly distributed on its right half; that is, $p'_i = \begin{cases} p_i/2 & i \in [n] \\ 1/n & \text{otherwise} \end{cases}$. Note that yes-instances are mapped to the uniform distribution, which has symmetric sparse support, and no-instances are mapped to distributions in which the right half is uniformly distributed and the left half contains $\varepsilon n/2$ elements of mass $2/n$, and hence it is $\Omega(\varepsilon)$-far from having symmetric sparse support.

Other properties. As aforementioned, similar techniques as in the reductions above (as well as in the identity testing reduction of Section 7, invoked on a specific p, e.g., the $\mathrm{Bin}(n, 1/2)$ distribution) can be applied to obtain nearly-tight lower bounds of $\tilde{\Omega}(\sqrt{n})$ (respectively $\tilde{\Omega}(n^{1/4})$) for the properties of being log-concave and monotone hazard rate (respectively Binomially and Poisson Binomially distributed). See e.g., [20] for the formal definitions of these properties.

9 Testing with Conditional Samples

In this section we show that reductions from communication complexity protocols can be used to obtain lower bounds on the sample complexity of distribution testers that are augmented with conditional samples. These testing algorithms, first introduced in [22, 21], aim to address scenarios that arise both in theory and practice yet are not fully captured by the standard distribution testing model.

In more detail, algorithms for testing with conditional samples are distribution testers that, in addition to sample access to a distribution $p \in \Delta(\Omega)$, can ask for samples from p conditioned on the sample belonging to a subset $S \subseteq \Omega$. It turns out that testers with conditional samples are much stronger than standard distribution testers, leading in many cases to exponential savings (or even more) in the sample complexity. In fact, these testing algorithms can often maintain their power even if they only have the ability to query subsets of a particular structure.

One of the most commonly studied restricted conditional samples models is the PAIRCOND model [21]. In this model, the testers can either obtain standard samples from p, or specify two distinct indices $i, j \in \Omega$ and get a sample from p conditioned on membership in $S = \{i, j\}$. As shown in [21, 18], even under this restriction one can obtain constant- or $\mathrm{poly}\log(n)$-query testers for many properties, such as uniformity, identity, closeness, and monotonicity (all of which require $\Omega(\sqrt{n})$ or more samples in the standard sampling setting). This, along with the inherent difficulty of proving hardness results against *adaptive* algorithms, makes proving lower bounds in this setting a challenging task; and indeed the PAIRCOND lower bounds established in the aforementioned works are quite complex and intricate.

We will prove, via a reduction from communication complexity, a strong lower bound on the sample complexity of any PAIRCOND algorithm for testing *junta distributions*, a class of distributions introduced in [5] (see definition below).

Since PAIRCOND algorithms are stronger than standard distribution testers (in particular, they can make adaptive queries), we shall reduce from the general randomized communication complexity model (rather than from the SMP model, as we did for standard distribution testers). In this model, Alice and Bob are given inputs x and y as well as a common random

string, and the parties aim to compute a function $f(x, y)$ using the minimum amount of communication.

We say that a distribution $p \in \Delta(\{0,1\}^n)$ is a *k-junta distribution* (with respect to the uniform distribution) if its probability mass function is only influenced by k of its variables. We outline below a proof of the following lower bound.

▶ **Theorem 37.** *Every* PAIRCOND *algorithm for testing k-junta distributions must make* $\Omega(k)$ *queries.*

Sketch of Proof. We closely follow the k-linearity lower bound in [15] and reduce from the unique $(k/2)$-disjointness problem. In this promise problem, Alice and Bob get inputs $x \in \{0,1\}^n$ and $y \in \{0,1\}^n$ (respectively) of Hamming weight $k/2$ each, and the parties are required to decide whether $\sum_{i=1}^n x_i y_i = 1$ or $\sum_{i=1}^n x_i y_i = 0$. It is well-known that in every randomized protocol for this problem the parties must communicate $\Omega(k)$ bits.

Assume there exists a PAIRCOND algorithm for testing k-junta distributions, with query complexity q. The reduction is as follows. Alice sets $A = \{ i \in [n] : x_i = 1 \}$ and considers the character function $\chi_A(z) = \oplus_{i \in A} z_i$, and similarly Bob sets $B = \{ i \in [n] : y_i = 1 \}$ and considers the character function $\chi_B(z) = \oplus_{i \in B} z_i$. Both players then invoke the tester for k-junta distributions, feeding it samples emulated from the distribution $p \in \Delta(\{0,1\}^n)$ given by $p(z) = \chi_{A \triangle B}(z)/2^{n-1}$ (where $\chi_{A \triangle B}(z) = \oplus_{i \in A \triangle B} z_i$); note that since the non-zero character functions are balanced, p is indeed a probability distribution. Recall that each query of a PAIRCOND algorithm is performed by either setting $S = \{0,1\}^n$, or choosing $z, z' \in \{0,1\}^n$ and setting $S = \{z, z'\}$, then sampling from $p|_S$. The players emulate each PAIRCOND query by the following rejection sampling procedure:

Sampling query ($S = \{0,1\}^n$): Alice and Bob proceed as follows.

    1. Choose $z \in S$ uniformly at random, using shared randomness;

    2. Exchange $\chi_A(z)$ and $\chi_B(z)$ between the players, and compute $\chi_{A \triangle B}(z) = \chi_A(z) \cdot \chi_B(z)$;

    3. If $\chi_{A \triangle B}(z) = 1$, feed the tester with the sample z. Otherwise repeat the process.

    Note that since $\chi_{A \triangle B}(z)$ is a balanced function, then on expectation each PAIRCOND query to p can be emulated by exchanging $O(1)$ bits.

Pairwise query ($S = \{z, z'\}$ for some $z, z' \in \{0,1\}^n$): exchange $\quad \chi_A(z), \chi_A(z') \quad$ and $\chi_B(z), \chi_B(z')$ between the players, compute $\chi_{A \triangle B}(z)$ and $\chi_{A \triangle B}(z')$, and use shared randomness to sample from S with the corresponding (now fully known) conditional probabilities.

The above gives a protocol with *expected* communication complexity $O(q)$, correct with probability $5/6$. To convert it to a honest-to-goodness protocol with communication complexity $O(q)$ and success probability $2/3$, it suffices for Alice and Bob to run the above protocol and stop (and output reject) as soon as they go over Ck bits of communication, for some absolute constant $C > 0$. An application of Markov's inequality guarantees that this happens with probability at most $1/6$, yielding the claimed bound on the error probability of the protocol.

Finally, note that on the one hand, if (x, y) is such that $\sum_{i=1}^n x_i y_i = 0$, then $\chi_{A \triangle B}(z)$ is a degree-k character, and in particular, a k-junta. Hence, by definition p is a k-junta distribution. On the other hand, if (x, y) is such that $\sum_{i=1}^n x_i y_i = 1$, then $\chi_{A \triangle B}(z)$ is a degree-$(k-2)$ character, which in particular disagrees with every k-junta on $\Omega(1)$-fraction of the inputs. Therefore, since p is uniform over its support, we can deduce that that p is $\Omega(1)$-far in ℓ_1-distance from any k-junta distribution. ◀

―――― **References** ――――

1   Jayadev Acharya, Clément L. Canonne, and Gautam Kamath. A Chasm Between Identity and Equivalence Testing with Conditional Queries. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2015)*, volume 40 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 449–466. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2015. `doi:10.4230/LIPIcs.APPROX-RANDOM.2015.449`.

2   Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *Proceedings of COLT*, pages 47–68, 2011.

3   Jayadev Acharya and Constantinos Daskalakis. Testing Poisson Binomial Distributions. In *Proceedings of SODA*, pages 1829–1840, 2015.

4   Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Proceedings of NIPS*, pages 3577–3598, 2015.

5   Maryam Aliakbarpour, Eric Blais, and Ronitt Rubinfeld. Learning and testing junta distributions. In *Proceedings of COLT*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 19–46. JMLR.org, 2016.

6   Sergey V. Astashkin. Rademacher functions in symmetric spaces. *Journal of Mathematical Sciences*, 169(6):725–886, sep 2010. `doi:10.1007/s10958-010-0074-z`.

7   Tuğkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

8   Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of FOCS*, pages 442–451, 2001.

9   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.

10   Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *ArXiV*, abs/1009.5397, 2010. This is a long version of [9].

11   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. `doi:10.1145/1007352.1007414`.

12   Colin Bennett and Robert C. Sharpley. *Interpolation of Operators*. Pure and Applied Mathematics. Elsevier Science, 1988. URL: `https://books.google.com/books?id=HpqF9zjZWMMC`.

13   Bhaswar B. Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. In *Proceedings of NIPS*, pages 2611–2619, 2015.

14   Arnab Bhattacharyya, Eldar Fischer, Ronitt Rubinfeld, and Paul Valiant. Testing monotonicity of distributions over general partial orders. In *Proceedings of ITCS*, pages 239–252, 2011. URL: `http://conference.itcs.tsinghua.edu.cn/ICS2011/content/papers/38.html`.

15   Eric Blais, Joshua Brody, and Kevin Matulef. Property testing lower bounds via communication complexity. *Computational Complexity*, 21(2):311–358, 2012. `doi:10.1007/s00037-012-0040-x`.

16   Joshua Brody, Kevin Matulef, and Chenggang Wu. Lower bounds for testing computability by small width OBDDs. In *TAMC*, volume 6648 of *Lecture Notes in Computer Science*, pages 320–331. Springer, 2011.

17   Clément L. Canonne. Big Data on the Rise? Testing Monotonicity of Distributions. In *Proceedings of ICALP*, pages 294–305. Springer, 2015. `doi:10.1007/978-3-662-47672-7_24`.

18   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. URL: `https://eccc.weizmann.ac.il/report/2015/063/`.

19   Clément L. Canonne. Are Few Bins Enough: Testing Histogram Distributions. In *Proceedings of PODS*. Association for Computing Machinery (ACM), 2016. `doi:10.1145/2902251.2902274`.

20   Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing Shape Restrictions of Discrete Distributions. In *33rd Symposium on Theoretical Aspects of Computer Science (STACS 2016)*, volume 47 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 25:1–25:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2016. `doi:10.4230/LIPIcs.STACS.2016.25`.

21   Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 44(3):540–616, 2015. Also available on arXiv at abs/1211.2664. `doi:10.1137/130945508`.

22   Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of ITCS*, pages 561–580, New York, NY, USA, 2013. ACM. `doi:10.1145/2422436.2422497`.

23   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. Society for Industrial and Applied Mathematics (SIAM), 2014.

24   Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of SODA*, pages 1833–1852. Society for Industrial and Applied Mathematics (SIAM), 2013. URL: `http://dl.acm.org/citation.cfm?id=2627817.2627948`.

25   Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of FOCS*. IEEE Computer Society, 2016.

26   Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Optimal Algorithms and Lower Bounds for Testing Closeness of Structured Distributions. In *Proceedings of FOCS*. Institute of Electrical and Electronics Engineers (IEEE), oct 2015. `doi:10.1109/focs.2015.76`.

27   Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of SODA*, pages 1841–1854. Society for Industrial and Applied Mathematics (SIAM), January 2015.

28   Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapathi, and Ananda Theertha Suresh. Faster algorithms for testing under conditional sampling. *ArXiV*, abs/1504.04103, April 2015.

29   Eldar Fischer, Oded Lachish, and Yadu Vasudev. Improving and extending the testing of distributions for shape-restricted properties. *ArXiV*, abs/1609.06736, September 2016.

30   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.

31   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.

32   Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7:20, 2000. URL: `https://eccc.weizmann.ac.il/report/2000/020/`.

33   Paweł Hitczenko and Stanisław Kwapień. On the Rademacher series. In *Probability in Banach spaces, 9 (Sandjberg, 1993)*, volume 35 of *Progr. Probab.*, pages 31–36. Birkhäuser

Boston, Boston, MA, 1994.

34   Tord Holmstedt. Interpolation of Quasi-Normed Spaces. *Mathematica Scandinavica*, 26(0):177–199, 1970. URL: `http://www.mscand.dk/article/view/10976`.

35   Piotr Indyk, Reut Levi, and Ronitt Rubinfeld. Approximating and Testing k-Histogram Distributions in Sub-linear Time. In *Proceedings of PODS*, pages 15–22, 2012.

36   Jiantao Jiao, Kartik Venkat, and Tsachy Weissman. Order-optimal estimation of functionals of discrete distributions. *ArXiV*, abs/1406.6956, 2014. URL: `http://arxiv.org/abs/1406.6956`.

37   Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9:295–347, 2013. `doi:10.4086/toc.2013.v009a008`.

38   Stephen J. Montgomery-Smith. The distribution of Rademacher sums. *Proceedings of the American Mathematical Society*, 109(2):517–522, 1990.

39   Ilan Newman. Property testing of massively parametrized problems – A survey. In *Property Testing*, volume 6390 of *Lecture Notes in Computer Science*, pages 142–157. Springer, 2010.

40   Ilan Newman and Mario Szegedy. Public vs. private coin flips in one round communication games. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 561–570. ACM, 1996.

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

42   Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. `doi:10.1109/TIT.2008.928987`.

43   Jaak Peetre. *A theory of interpolation of normed spaces.* Notas de Matemática, No. 39. Instituto de Matemática Pura e Aplicada, Conselho Nacional de Pesquisas, Rio de Janeiro, 1968.

44   David Pollard. Asymptopia. `http://www.stat.yale.edu/~pollard/Books/Asymptopia`, 2003. Manuscript.

45   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.

46   Ronitt Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24, sep 2012. `doi:10.1145/2331042.2331052`.

47   Ronitt Rubinfeld and Rocco A. Servedio. Testing monotone high-dimensional distributions. *Random Structures and Algorithms*, 34(1):24–44, January 2009. `doi:10.1002/rsa.v34:1`.

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

49   Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:179, 2010. URL: `https://eccc.weizmann.ac.il/report/2010/179/`.

50   Gregory Valiant and Paul Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:180, 2010. URL: `https://eccc.weizmann.ac.il/report/2010/180/`.

51   Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of STOC*, pages 685–694. ACM, 2011.

52   Gregory Valiant and Paul Valiant. The power of linear estimators. In *Proceedings of FOCS*, pages 403–412, October 2011. See also [49] and [50]. `doi:10.1109/FOCS.2011.81`.

53   Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of FOCS*, pages 51–60, 2014.

54   Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*,

40(6):1927–1968, 2011.

55 Wikipedia. Hypergeometric distribution – wikipedia, the free encyclopedia, 2016. [Online; accessed 28-May-2016]. URL: `https://en.wikipedia.org/w/index.php?title=Hypergeometric_distribution&oldid=702419153#Multivariate_hypergeometric_distribution`.

56 Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *ArXiV*, abs/1407.0381, 2014. URL: `http://arxiv.org/abs/1407.0381`.

57 Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997. `doi:10.1007/978-1-4612-1880-7_29`.