Quantum algorithms for testing properties of distributions

Sergey Bravyi[*], Aram W. Harrow[†] and Avinatan Hassidim[‡]

July 8, 2009

Abstract

Suppose one has access to oracles generating samples from two unknown probability distributions p and q on some N-element set. How many samples does one need to test whether the two distributions are close or far from each other in the L_1-norm? This and related questions have been extensively studied during the last years in the field of property testing. In the present paper we study quantum algorithms for testing properties of distributions. It is shown that the L_1-distance $\|p-q\|_1$ can be estimated with a constant precision using only $O(N^{1/2})$ queries in the quantum settings, whereas classical computers need $\Omega(N^{1-o(1)})$ queries. We also describe quantum algorithms for testing Uniformity and Orthogonality with query complexity $O(N^{1/3})$. The classical query complexity of these problems is known to be $\Omega(N^{1/2})$. A quantum algorithm for testing Uniformity has been recently independently discovered by Chakraborty et al [1].

1 Introduction

Suppose one has access to a black box generating independent samples from an unknown probability distribution p on some N-element set. If the number of available samples grows linearly with N, one can use the standard Monte Carlo method to simultaneously estimate the probability p_i of every element $i = 1, \ldots, N$ and thus obtain a good approximation to the entire distribution p. On the other hand, many important questions that one usually encounters in statistical analysis can be answered using only a *sublinear* number of samples. For example, deciding whether p is close in the L_1-norm to another distribution q requires approximately $N^{1/2}$ samples if q is known [2] and approximately $N^{2/3}$ samples if q is also specified by a black-box [3]. Another example is estimating the Shannon entropy $H(p) = -\sum_i p_i \log_2 p_i$. It was shown in [4, 8] that distinguishing whether $H(p) \le a$ or $H(p) \ge b$ requires approximately $N^{\frac{a}{b}}$ samples. Other examples include deciding whether p is close to a monotone or a unimodal distribution [5], and deciding whether a pair of distributions have disjoint supports [6]. These and other questions fall into the field of *distribution testing* [7, 8] that studies how many samples one needs to decide whether an unknown distribution has a certain property or is far from having this property. The purpose of the present paper is to explore whether quantum computers are capable of solving distribution testing problems more efficiently.

[*]IBM Watson Research Center, Yorktown Heights, NY 10598 (USA); *email:* sbravyi@us.ibm.com
[†]Department of Mathematics, University of Bristol, Bristol, U.K.
[‡]MIT, *email:* avinatanh@gmail.com

The black-box sampling model adopted in [2, 3, 4, 5, 7, 8] assumes that a tester is presented with a list of samples drawn from an unknown distribution. What does it mean to sample from an unknown distribution in the quantum settings? Let us start by casting the black-box sampling model into a form that admits a quantum generalization. Suppose p is an unknown distribution on an N-element set $[N] \equiv \{1, \ldots, N\}$ and let S be some specified integer. We shall assume that p is represented by an *oracle* $O_p : [S] \to [N]$ such that a probability p_i of any element $i \in [N]$ is proportional to the number of elements in the pre-image of i, that is, the number of inputs $s \in [S]$ such that $O_p(s) = i$. In other words, one can sample from p by querying the oracle O_p on a random input $s \in [S]$ drawn from the uniform distribution[1]. Note that a tester interacting with an oracle can potentially be more powerful due to the possibility of making adaptive queries which could allow him to learn the internal structure of the oracle as opposed to the black-box model. However, it will be shown below (see Lemma 9 in Section 6) that the oracle model and the black-box model are in fact equivalent. More precisely, for any fixed N one can always choose sufficiently large S such that a tester will need the same number of queries in both models.

The oracle model admits a standard quantum generalization. Specifically, we shall transform the oracle O_p into a reversible form by keeping a copy of the input and writing the output of O_p into an ancillary register. A quantum oracle generating p is a unitary operator whose action on basis vectors coincides with the reversible version of O_p, see Section 2 for technical details.

The present paper focuses on testing three particular properties of distributions, namely, *Statistical Difference*, *Orthogonality*, and *Uniformity*. The corresponding property testing problems are promise problems so that a tester is required to give a correct answer (with a bounded error probability) only for those instances that satisfy the promise.

Problem 1 (Testing Uniformity).
Instance: *Integers N, S, precision $\epsilon > 0$. Access to an oracle generating a distribution p on $[N]$.*
Promise: *Either p is the uniform distribution or the L_1-distance between p and the uniform distribution is at least ϵ.*
Decide which one is the case.

Problem 2 (Testing Orthogonality).
Instance: *Integers N, S, precision $\epsilon > 0$. Access to oracles generating distributions p, q on $[N]$.*
Promise: *Either p and q are orthogonal or the L_1-distance between p and q is at most $2 - \epsilon$.*
Decide which one is the case.

Problem 3 (Testing Statistical Difference).
Instance: *Integers N, S, thresholds $0 \le a < b \le 2$. Access to oracles generating distributions p and q on $[N]$.*
Promise: *Either $\|p - q\|_1 \le a$ or $\|p - q\|_1 \ge b$.*
Decide which one is the case.

We assume that the precision ϵ is bounded from below by a fixed constant independent of N, for instance, $\epsilon \ge 1/10$. The same applies to the decision gap $b - a$ for testing Statistical Difference. Given a function $f(N)$

[1] Although in this model probabilities p_i can only take values that are multiples of $1/S$, choosing sufficiently large S allows one to represent any distribution p with an arbitrarily small error.

we shall say that a property is testable in $f(N)$ queries if there exists a testing algorithm making at most $f(N)$ queries that gives a correct answer with a sufficiently high probability (say 2/3) for any distributions p, q satisfying the promise and for any oracles[2] specifying p and q. If a promise is violated, a tester can give an arbitrary answer.

Our main results are the following theorems.

Theorem 1. *Statistical Difference is testable on a quantum computer in $O(N^{1/2})$ queries.*

Theorem 2. *Uniformity is testable on a quantum computer in $O(N^{1/3})$ queries.*

Theorem 3. *Orthogonality is testable on a quantum computer in $O(N^{1/3})$ queries.*

It is known that classically testing Orthogonality and Uniformity requires $\Omega(N^{1/2})$ queries, see Sections 6.2 and 6.3, while Statistical Difference is not testable in $O(N^{\alpha})$ queries for any $\alpha < 1$, see [8]. Therefore quantum computers provide a polynomial speedup for testing Uniformity, Orthogonality, and Statistical Difference in terms of query complexity.

More interesting than the mere fact of a polynomial speedup is the way in which our algorithms achieve it. Classically, it is trivially true that with M queries to the oracle the best strategy is to query it on a random set of M distinct inputs. Additionally, the Wishful Thinking theorem of Ref. [**?**] gives a simple characterization of any asymptotically optimal estimation algorithm. By contrast, our algorithms use a variety of different strategies both to query the oracles and to analyze the results of those queries. These strategies appear not to be special cases of the quantum walk framework which has been responsible for most of the polynomial quantum speedups found to date[19, 18]. A major challenge for future research is to give a quantum version of Ref. [**?**]'s Wishful Thinking theorem; in other words, we would like to characterize optimal quantum algorithms for any symmetric property testing problem.

Testing Orthogonality is closely related to the Collision Problem studied in [14, 13]. In Section 6.2 we describe a randomized reduction from the Collision Problem to testing Orthogonality. Using the quantum lower bound for the Collision Problem due to Aaronson and Shi [15] we obtain the following result.

Theorem 4. *Testing Orthogonality on a quantum computer requires $\Omega(N^{1/3})$ queries.*

Quite recently Chakraborty, Fischer, Matsliah, and de Wolf [1] independently discovered a quantum Uniformity testing algorithm with query complexity $O(N^{1/3})$ and proved a lower bound $\Omega(N^{1/3})$ for testing Uniformity. These authors also presented a quantum algorithm for testing whether an unknown distribution p coincides with a known distribution q with query complexity $\tilde{O}(N^{1/3})$.

The rest of the paper is organized as follows. Section 2 introduces necessary notations and basic facts about the quantum counting algorithm by Brassard, Hoyer, Mosca, and Tapp [17]. The distribution testing algorithms described in the rest of the paper are actually classical probabilistic algorithms using the quantum counting as a subroutine. Theorem 1 is proved in Section 3. Theorem 2 is proved in Section 4. Theorem 3 is proved in Section 5. We discuss lower bounds for the above distribution testing problems in Section 6.

[2]Note that according to this definition a tester needs at most $f(N)$ queries even in the limit $S \to \infty$.

2 Preliminaries

Let $\mathcal{D}_N$ be a set of probability distributions $p = (p_1, \ldots, p_N)$ such that a probability p_i of any element $i \in [N]$ is a rational number. Let us say that an oracle $O : [S] \to [N]$ generates a distribution $p \in \mathcal{D}_N$ iff for all $i \in [N]$ the probability p_i equals the fraction of inputs $s \in [S]$ such that $O(s) = i$,

$$p_i = \frac{1}{S} \#\{s \in [S] \, : \, O(s) = i\}.$$

Note that the identity of elements in the domain of an oracle O is irrelevant, so if O generates p and σ is any permutation on $[S]$ then $O \circ \sigma$ also generates p. By definition, any map $O : [S] \to [N]$ generates some distribution $p \in \mathcal{D}_N$.

For any oracle $O : [S] \to [N]$ we shall define a quantum oracle $\hat{O}$ by transforming O into a reversible form and allowing it to accept coherent superpositions of queries. Specifically, a quantum oracle $\hat{O}$ is a unitary operator acting on a Hilbert space $\mathbb{C}^S \otimes \mathbb{C}^{N+1}$ equipped with a standard basis $\{|s\rangle \otimes |i\rangle\}$, $s \in [S]$, $i \in \{0\} \cup [N]$ such that

$$\hat{O}\, |s\rangle \otimes |0\rangle = |s\rangle \otimes |O(s)\rangle \quad \text{for all } s \in [S]. \tag{1}$$

In other words, querying $\hat{O}$ on a basis vector $|s\rangle \otimes |0\rangle$ one gets the output of the classical oracle $O(s)$ in the second register while the first register keeps a copy of s to maintain unitarity. The action of $\hat{O}$ on a subspace in which the second register is orthogonal to the state $|0\rangle$ can be arbitrary. We shall assume that a quantum tester can execute operators $\hat{O}$, $\hat{O}^\dagger$ and the controlled versions of them. Execution of any one of these operators counts as one query.

We shall see that all testing problems posed in Section 1 can be reduced (via classical randomized reductions) to the following problem.

Problem 4 (**Probability Estimation**). *Given integers S, N, description of a subset $A \subset [N]$, precision δ, error probability ω, and access to an oracle generating some distribution $p \in \mathcal{D}_N$. Let $p_A = \sum_{i \in A} p_i$ be the total probability of A. One needs to generate an estimate $\tilde{p}_A$ satisfying*

$$\Pr\left[|\tilde{p}_A - p_A| \le \delta\right] \ge 1 - \omega. \tag{2}$$

Our main technical tool will be the quantum counting algorithm by Brassard et al. [17]. Specifically, we shall use the following version of Theorem 12 from [17].

Theorem 5. *There exists a quantum algorithm* **EstProb**(p, A, M) *taking as input a distribution $p \in \mathcal{D}_N$ specified by an oracle, a subset $A \subset [N]$, and an integer M. The algorithm makes exactly M queries to the oracle generating p and outputs an estimate $\tilde{p}_A$ such that*

$$\Pr\left[|\tilde{p}_A - p_A| \le \delta\right] \ge 1 - \omega \tag{3}$$

for all $\delta > 0$ and $0 \le \omega \le 1/2$ satisfying

$$M \ge \frac{c\sqrt{p_A}}{\omega\delta} \quad \text{and} \quad M \ge \frac{c}{\omega\sqrt{\delta}}. \tag{4}$$

Here $c = O(1)$ is some constant. If $p_A = 0$ then $\tilde{p}_A = 0$ with certainty.

The proof can be found in Ref. [20] and is omitted from this extended abstract.

3 Quantum algorithm for estimating statistical difference

In this section we prove Theorem 1. Let $p, q \in \mathcal{D}_N$ be unknown distributions specified by oracles. Define an auxiliary distribution $r \in \mathcal{D}_N$ such that $r_i = (p_i + q_i)/2$ for all $i \in [N]$. If we can sample i from both p and q then by choosing randomly between these two options we can also sample i from r. Let $x \in [0, 1]$ be a random variable which takes value

$$x_i = \frac{|p_i - q_i|}{p_i + q_i}$$

with probability r_i. It is evident that

$$\mathbb{E}(x) = \sum_{i \in [N]} r_i x_i = \frac{1}{2} \sum_{i \in [N]} |p_i - q_i| = \frac{1}{2} \|p - q\|_1. \tag{5}$$

Thus in order to estimate the distance $\|p - q\|_1$ it suffices to estimate the expectation value $\mathbb{E}(x)$ which can be done using the standard Monte Carlo method. Since we have to estimate $\mathbb{E}(x)$ only with a constant precision, it suffices to generate $O(1)$ samples of x_i. Given a sample of i (which is easy to generate classically) we can estimate x_i by calling the probability estimation algorithm to get estimates of p_i and q_i. It suggests the following algorithm for estimating the distance $\|p - q\|_1$.

EstDist(p, q, ϵ, τ)
Set $n = 27/\tau\epsilon^2$, $M = c\sqrt{N}/\epsilon^6\tau^4$.
Let $i_1, \ldots, i_n \in [N]$ be a list of n independent samples drawn from r.
For $a = 1, \ldots, n$
{
    Let $\tilde{p}_{i_a}$ be estimate of p_{i_a} obtained using **EstProb**$(p, \{i_a\}, M)$.
    Let $\tilde{q}_{i_a}$ be estimate of q_{i_a} obtained using **EstProb**$(q, \{i_a\}, M)$.
    Let $\tilde{x}_{i_a} = |\tilde{p}_{i_a} - \tilde{q}_{i_a}|/(\tilde{p}_{i_a} + \tilde{q}_{i_a})$ be estimate of x_{i_a}.
}
Output $\tilde{x} = (1/n) \sum_{a=1}^{n} \tilde{x}_{i_a}$.

Here $c = O(1)$ is a constant whose precise value will not be important for us.

Lemma 1. *The algorithm* **EstDist**(p, q, ϵ, τ) *outputs an estimate* $\tilde{x}$ *satisfying*

$$\Pr\left[|\tilde{x} - \mathbb{E}(x)| < \epsilon\right] \geq 1 - \tau, \tag{6}$$

where $\mathbb{E}(x) = (1/2)\|p - q\|_1$.

Proof. Define a random variable

$$\bar{x} = \frac{1}{n} \sum_{a=1}^{n} x_{i_a},$$

where $i_1, \ldots, i_n$ is a list of samples generated at the first step of the algorithm. Note that $\mathbb{E}(\bar{x}) = \mathbb{E}(x)$ and $\mathrm{Var}\,(\bar{x}) = \mathrm{Var}\,(x)/n$. As $|p_i - q_i| \le p_i + q_i$ we have $0 \le x_i \le 1$ and so one can bound the variance of x as $\mathrm{Var}\,(x) \le \mathbb{E}(x^2) \le 1$. Therefore $\mathrm{Var}\,(\bar{x}) \le 1/n$. Applying the Chebyshev inequality to $\bar{x}$ one gets

$$\Pr\left[|\bar{x} - \mathbb{E}(x)| \ge \epsilon/3\right] \le \frac{9\,\mathrm{Var}\,(\bar{x})}{\epsilon^2} \le \frac{9}{n\epsilon^2} \le \frac{\tau}{3}. \tag{7}$$

Let $\tilde{x}$ be the output of **EstDist**(p, q, ϵ, τ). The union bound implies that

$$\Pr\left[|\tilde{x} - \bar{x}| \ge \epsilon/3\right] \le \Pr\left[\exists a \,:\, |\tilde{x}_{i_a} - x_{i_a}| \ge \epsilon/3n\right] \le n\Pr\left[|\tilde{x}_i - x_i| \ge \epsilon/3n\right], \tag{8}$$

where $i \equiv i_a$ is a sample drawn from r. Therefore it suffices to verify that

$$\Pr\left[|\tilde{x}_i - x_i| \ge \epsilon/3n\right] \le \frac{2\tau}{3n}. \tag{9}$$

Let us say that an element i is *bad* iff

$$\max\,(p_i, q_i) \le \frac{\tau}{3nN} \quad \text{(bad element)}. \tag{10}$$

The probability that i is bad is at most

$$p_{bad} = \sum_{i \text{ is bad}} r_i \le \frac{\tau}{3n}.$$

Therefore it suffices to get a bound

$$\Pr\left[|\tilde{x}_i - x_i| \ge \epsilon/3n \,|\, i \text{ is good}\right] \le \frac{\tau}{3n}, \tag{11}$$

where we conditioned on i being a good (not bad) element.

Let us translate the precision up to which one needs to estimate x_i into a precision up to which one needs to estimate p_i and q_i.

Proposition 1. *Consider a real-valued function $f(p,q) = (p-q)/(p+q)$ where $0 \le p, q \le 1$. Assume that $|p - \tilde{p}|, |q - \tilde{q}| \le \delta(p+q)$ for some $\delta \le 1/5$. Then*

$$|f(p,q) - f(\tilde{p},\tilde{q})| \le 5\delta. \tag{12}$$

The proof can be found in Ref. [20] and is omitted from this extended abstract. Note that

$$|\tilde{x}_i - x_i| = |\,|f(\tilde{p}_i,\tilde{q}_i)| - |f(p_i,q_i)|\,| \le |f(\tilde{p}_i,\tilde{q}_i) - f(p_i,q_i)|.$$

Since we want to estimate x_i with a precision $\epsilon/3n$, it suffices to estimate p_i and q_i with a precision $\delta(p_i + q_i) \ge \delta \max\,(p_i, q_i)$ where $5\delta = \epsilon/3n$, that is, $\delta = \epsilon/(15n)$. Summarizing,

$$|\tilde{p}_i - p_i|, |\tilde{q}_i - q_i| \le \frac{\epsilon}{15n}\max\,(p_i, q_i) \quad \Rightarrow \quad |\tilde{x}_i - x_i| \le \frac{\epsilon}{3n}. \tag{13}$$

Thus it suffices to estimate p_i and q_i with precision

$$\delta \sim \epsilon n^{-1} \max(p_i, q_i) \sim \tau\epsilon^3 \max(p_i, q_i). \tag{14}$$

We are going to get these estimates by calling $\mathbf{EstProb}(p, \{i\}, M)$ and $\mathbf{EstProb}(q, \{i\}, M)$. The number of queries M has to be chosen sufficiently large such that conditions Eq. (4) are satisfied for precision δ defined in Eq. (14) and error probability determined by Eq. (11), that is,

$$\omega \sim \tau n^{-1} \sim \tau^2\epsilon^2. \tag{15}$$

It leads to the condition

$$M \geq \Omega\left(\frac{1}{\tau^3\epsilon^5 \max(\sqrt{p_i}, \sqrt{q_i})}\right). \tag{16}$$

Recall that we are interested in the case when i is good. In this case $\max(p_i, q_i) \geq \tau/(3nN) \sim N^{-1}\tau^2\epsilon^2$. Therefore Eq. (16) is satisfied whenever

$$M \geq \Omega\left(\frac{1\sqrt{N}}{\tau^4\epsilon^6}\right).$$

□

Theorem 1 follows directly from Lemma 1 since $\mathbf{EstDist}(p, q, \epsilon, \tau)$ makes $O(\sqrt{N})$ queries to the quantum oracles generating p and q.

4 Quantum algorithm for testing Uniformity

In this section we prove Theorem 2. Let $p \in \mathcal{D}_N$ be an unknown distribution specified by an oracle. We are promised that either p is the uniform distribution, or p is ϵ-*nonuniform*, that is, the L_1-distance between p and the uniform distribution is at least ϵ. The algorithm described below is based on the following simple observation. Choose some integer $M \ll N$ and let $S = (i_1, \ldots, i_M)$ be a list of M independent samples drawn from the distribution p. Define a random variable $p_S = \sum_{a=1}^{M} p_{i_a}$. It coincides with the total probability of all elements in S unless S contains a collision (that is, $i_a = i_b$ for some $a \neq b$). The characteristic property of the uniform distribution is that $p_S = M/N$ with certainty. On the other hand, we shall see that for any ϵ-nonuniform distribution p_S takes values greater than $(1+\delta)M/N$ for some constant $\delta > 0$ depending on ϵ with a non-negligible probability. This observation suggests the following algorithm for testing uniformity (the constants K and M below will be chosen later).

$\mathbf{UTest}(p, K, M, \epsilon)$
Let $S = (i_1, \ldots, i_M)$ be a list of M independent samples drawn from p.
Reject unless all elements in S are distinct.
Let $p_S = \sum_{a=1}^{M} p_{i_a}$ be the total probability of elements in S.
Let $\tilde{p}_S$ be an estimate of p_S obtained using $\mathbf{EstProb}(p, S, K)$.
If $\tilde{p}_S > (1 + \epsilon^2/8)M/N$ then reject. Otherwise accept.

This procedure will need to be repeated several times to achieve the desired bound on the error probability, see the proof of Theorem 2 below.

The main technical result of this section is the following lemma.

Lemma 2. *Let $p \in \mathcal{D}_N$ be an ϵ-nonuniform distribution. Let $S = (i_1, \ldots, i_M)$ be a list of M independent samples drawn from p, where*

$$M^3 = \frac{32N}{\epsilon^4}. \tag{17}$$

Let $p_S = \sum_{a=1}^{M} p_{i_a}$ and $\alpha = 2^8 \epsilon^{-4}$. Then

$$\Pr\left[p_S \geq (1+\epsilon^2/2)\frac{M}{N}\right] \geq \frac{1}{2}\exp(-\alpha). \tag{18}$$

Theorem 1 follows straightforwardly from the above lemma and Theorem 5.

Proof of Theorem 1. Let M be chosen as in Eq. (17) and

$$K = c\frac{e^{\alpha} N^{1/3}}{\epsilon^{4/3}},$$

where $c = O(1)$ is a constant to be chosen later. Consider the following algorithm:

Perform $L = 4\exp(\alpha)$ independent tests $\mathbf{UTest}(p, K, M, \epsilon)$. If at least one of the tests outputs 'reject' then reject. Otherwise accept.

Let us show that this algorithm rejects any ϵ-nonuniform distribution with probability at least 2/3 and accepts the uniform distribution with probability at least 2/3.

Part 1: Any ϵ-nonuniform distribution is rejected with high probability. Let P_s be the probability that for at least one of the **UTest**s one has

$$p_S \geq (1+\epsilon^2/2)\frac{M}{N} \tag{19}$$

Using Lemma 2 we conclude that

$$P_s \geq 1 - \left(1 - \frac{1}{2e^{\alpha}}\right)^{4e^{\alpha}} \geq 1 - e^{-2} \geq \frac{5}{6}. \tag{20}$$

In what follows we shall focus on a single test $\mathbf{UTest}(p, K, M, \epsilon)$ that satisfies Eq. (19) and show that it outputs 'reject' with high probability. Indeed, let S be the sample list generated by this **UTest**. If S contains a collision, the test outputs 'reject'. Otherwise p_S coincides with the total probability of all elements in S. The test outputs 'reject' whenever p_S is estimated with a precision

$$\delta = p_S \frac{\epsilon^2}{4}. \tag{21}$$

In this case

$$\tilde{p}_S \geq \left(1 - \frac{\epsilon^2}{4}\right) p_S \geq \left(1 - \frac{\epsilon^2}{4}\right)\left(1 + \frac{\epsilon^2}{2}\right)\frac{M}{N} > \left(1 + \frac{\epsilon^2}{8}\right)\frac{M}{N}.$$

(Here we assumed for simplicity that $\epsilon \leq 1$.) Suppose we want the **UTest** to output 'reject' with probability at least 5/6. Applying Eq. (4) with δ defined in Eq. (21) and $\omega = 1/6$ we arrive at

$$K \geq \frac{c}{\epsilon^2 \sqrt{p_S}} \tag{22}$$

for some constant $c = O(1)$. Using Eq. (19) it suffices to choose

$$K = O\left(\frac{\sqrt{N}}{\epsilon^2\sqrt{M}}\right) = O\left(\frac{N^{1/3}}{\epsilon^{4/3}}\right) \tag{23}$$

Summarizing, if p is an ϵ-nonuniform distribution it will be rejected with probability at least $(5/6)^2 \geq 2/3$.

Part 2: The uniform distribution is accepted with high probability. Note that the uniform distribution can be rejected for two possible reasons: (i) for some **UTest** the sample list S contains a collision; (ii) for some **UTest** the estimate $\tilde{p}_S$ is sufficiently large, $\tilde{p}_S > (1 + \epsilon^2/8)\, M/N$. We analyze these two possible sources of errors below.

(i) For any fixed **Utest** let $S = (i_1, \ldots, i_M)$ be a list of M samples drawn from p. Let C be the number of collisions in S, that is, the number of pairs $1 \leq a < b \leq M$ such that $i_a = i_b$. Then,

$$\mathbb{E}(C) = \binom{M}{2} \sum_{i=1}^{N} p_i^2 \leq \frac{M^2}{2N}.$$

Markov's inequality implies that $\Pr[C \geq 1] \leq \mathbb{E}(C) \leq M^2/(2N)$. Then the probability that at least one of the **UTests** will find a collision can be bounded using the union bound as

$$P_c \leq \frac{LM^2}{2N} = O\left(\frac{1}{N^{1/3}}\right)$$

since we have chosen $M = O(N^{1/3})$ and $L = O(1)$. Thus the error probability associated with finding collisions can be neglected.

(ii) Let $\tilde{p}_S$ be the estimate of p_S obtained in some fixed **UTest**. Since $p_S = M/N$ with certainty, the test outputs 'accept' whenever the estimate $\tilde{p}_S$ returned by **EstProb**(p, S, K) satisfies $|\tilde{p}_S - p_S| \leq \delta$, where

$$\delta = \frac{\epsilon^2 M}{8N}. \tag{24}$$

Since the total number of **Utest**s is $L = 4e^\alpha$, we would like the estimate $\tilde{p}_S$ to have precision δ with error probability $\omega \leq \frac{1}{12}e^{-\alpha}$. Applying Eq. (4) with δ, ω defined above and taking into account that $p_S = M/N$, we find that we can take the number of queries K to be

$$K = O\left(\frac{\sqrt{p_S}}{\omega\delta}\right) = O\left(\frac{e^\alpha N^{1/3}}{\epsilon^{4/3}}\right). \tag{25}$$

It remains to choose the largest of Eq. (23) and Eq. (25). □

In the rest of this section we prove Lemma 2. We shall adopt notations introduced in the statement of Lemma 2, that is, the number of samples M is defined by

$$M^3 = 32\epsilon^{-4}N,$$

$\alpha \equiv 2^8\epsilon^{-4}$, $S = (i_1, \ldots, i_M)$ is a list of M independent samples drawn from p, and $p_S = \sum_{a=1}^{M} p_{i_a}$.

Definition 1. *An element $i \in [N]$ is called big iff $p_i > 1/(2M^2)$.*

Define the set $\mathrm{Big} \subset [N]$ of all big elements and their total probability:

$$\mathrm{Big} = \{i \in [N] \,:\, p_i > 1/(2M^2)\}, \quad w_{\mathrm{big}} = \sum_{i \in \mathrm{Big}} p_i. \tag{26}$$

We shall start in see subsection 4.1 by proving Lemma 2 for the special case when p has no big elements. The proof is based on Chebyshev's inequality. Then we shall leverage this result in subsection 4.2 to show that distributions with a few big elements (small w_{big}) also satisfy Lemma 2. Finally in subsection 4.3, we shall treat distributions with many big elements (large w_{big}) using a completely different technique.

4.1 Proof of Lemma 2: no big elements

Lemma 3 (**No big elements**)**.** *Suppose $p \in \mathcal{D}_N$ is ϵ-nonuniform and has no big elements. Then*

$$\Pr\left[p_S \ge \left(1 + \frac{\epsilon^2}{2}\right)\frac{M}{N}\right] \ge \frac{3}{4}. \tag{27}$$

Proof. One can easily check that

$$\mathbb{E}(p_S) = M\langle p|p\rangle, \quad \mathrm{Var}\,(p_S) = M\left(\sum_{i=1}^{N} p_i^3 - \langle p|p\rangle^2\right). \tag{28}$$

Proposition 2. *Suppose $p \in \mathcal{D}_N$ is ϵ-nonuniform. Then*

$$\langle p|p\rangle \ge \frac{1+\epsilon^2}{N}. \tag{29}$$

Proof. Let u be the uniform distribution. Then $\epsilon \le \|p-u\|_1 \le \sqrt{N}\,\|p-u\|_2 = \sqrt{N}\,\sqrt{\langle p|p\rangle - N^{-1}}$ which gives the desired bound. □

Using the proposition and the assumption that p has no big elements we get

$$\mathbb{E}(p_S) \geq \frac{M}{N}(1+\epsilon^2), \quad \mathrm{Var}\,(p_S) \leq M\|p\|_\infty \langle p|p\rangle \leq \frac{1}{2M}\langle p|p\rangle. \tag{30}$$

Chebyshev's inequality implies that

$$\Pr\left[|p_S - \mathbb{E}(p_S)| \geq t\mathbb{E}(p_S)\right] \leq \frac{\mathrm{Var}\,(p_S)}{\mathbb{E}(p_S)^2\, t^2}. \tag{31}$$

Assuming for simplicity that $\epsilon^2 \leq 1/3$ we can use the bound $(1+\epsilon^2)^{-1} \leq 1 - 3\epsilon^2/4$ and thus

$$\Pr\left[p_S \leq \left(1+\frac{\epsilon^2}{2}\right)\cdot\left(\frac{M}{N}\right)\right] \leq \Pr\left[p_S \leq \mathbb{E}(p_S)\,\frac{(1+\epsilon^2/2)}{(1+\epsilon^2)}\right] \leq \Pr\left[p_S \leq (1-\epsilon^2/4)\mathbb{E}(p_S)\right].$$

Using Eq. (31) with $t = \epsilon^2/4$ and Eqs. (28,30) we arrive at

$$\Pr\left[p_S \leq (M/N)(1+\epsilon^2/2)\right] \leq \frac{\langle p|p\rangle}{2M}\frac{1}{M^2\langle p|p\rangle^2 t^2} \leq \frac{8N}{M^3\epsilon^4} \leq \frac{1}{4}$$

since $\langle p|p\rangle \geq N^{-1}$ for any distribution $p \in \mathcal{D}_N$ and since we have chosen $M^3 = 32\epsilon^{-4}N$. □

4.2 Proof of Lemma 2: a few big elements

Lemma 4 (A few big elements). *Suppose $p \in \mathcal{D}_N$ is ϵ-nonuniform and has only a few big elements such that*

$$w_{\mathrm{big}} \leq \frac{\alpha}{M}, \quad \alpha \equiv 2^8\epsilon^{-4}. \tag{32}$$

Then

$$\Pr\left[p_S \geq (1+\epsilon^2/2)\frac{M}{N}\right] \geq \frac{1}{2}\,\exp\,(-\alpha). \tag{33}$$

Proof. Let $S = (i_1, \ldots, i_M)$ be a list of M samples drawn from p. We can get a constant lower bound on the probability that S contains no big elements:

$$\Pr\left[S \cap \mathrm{Big} = \emptyset\right] \quad = \quad (1-w_{\mathrm{big}})^M \approx \exp\,(-Mw_{\mathrm{big}}) \geq e^{-\alpha}. \tag{34}$$

(Strictly speaking, one gets a lower bound $e^{-\alpha}(1-o(1))$.) It suffices to show that $p_S \geq (1+\epsilon^2/2)M/N$ with probability at least $1/2$ conditioned on S having no big elements.

The conditional distribution of the random variable p_S given that S contains no big elements can be obtained by setting the probability of all big elements to zero and renormalizing p by a factor $(1-w_{\mathrm{big}})^{-1}$. In other words, we can repeat all arguments of Lemma 3 if we replace p by a new distribution $p' \in \mathcal{D}_N$ such that

$$p'_i = \begin{cases} \frac{p_i}{(1-w_{\mathrm{big}})} & \text{if} \quad i \notin \mathrm{Big}, \\ 0 & \text{if} \quad i \in \mathrm{Big}. \end{cases} \tag{35}$$

We have to check that p' is also ϵ-nonuniform.

Proposition 3. *The distribution p' is ϵ'-nonuniform, where $\epsilon' \geq \epsilon - O(N^{-1/3})$.*

Proof.

$$\|p - p'\|_1 = \sum_{i \in \mathrm{Big}} p_i + \sum_{i \notin \mathrm{Big}} \left[(1 - w_{\mathrm{big}})^{-1} - 1\right] p_i \leq w_{\mathrm{big}} + \frac{w_{\mathrm{big}}}{(1 - w_{\mathrm{big}})} = O(N^{-1/3}).$$

Let u be the uniform distribution. Using the triangle inequality we get

$$\|p' - u\|_1 \geq \|p - u\|_1 - \|p - p'\|_1 \geq \epsilon - O(N^{-1/3}).$$

□

To simplify notations we shall neglect the correction of order $N^{-1/3}$ and assume that p' is ϵ-nonuniform. By construction,

$$\|p'\|_\infty \leq \frac{1}{(1 - w_{\mathrm{big}})2M^2} = 1/(2M^2) + O(N^{-1}).$$

Neglecting the correction of order N^{-1} we can assume that p' has no big elements. Then Lemma 3 implies that $p'_S \geq (1+\epsilon^2/2)M/N$ with probability at least $3/4$. Combining it with Eq. (34) we arrive at Eq. (33). □

4.3 Proof of Lemma 2: many big elements

Lemma 5 (Many big elements). *Suppose p is ϵ-nonuniform and has many big elements such that*

$$w_{\mathrm{big}} > \frac{\alpha}{M}, \quad \alpha \equiv 2^8 \epsilon^{-4}. \tag{36}$$

Then

$$\Pr\left[p_S \geq 2\frac{M}{N}\right] \geq \frac{1}{2}. \tag{37}$$

Proof. Let $S = (i_1, \ldots, i_M)$ be a list of M independent samples drawn from p. Since each big element contained in S contributes at least $1/(2M^2)$ to p_S, the inequality $p_S \geq 2M/N$ is satisfied whenever S contains at least n big elements where

$$\frac{n}{2M^2} \geq \frac{2M}{N}.$$

Since $M^3 = 2^5 \epsilon^{-4} N$, we can choose

$$n = 2^7 \epsilon^{-4} = \alpha/2. \tag{38}$$

The total number of samples $a \in [M]$ such that i_a is big can be represented as $\xi = \sum_{i=1}^M \xi_i$, where $\xi_i \in \{0, 1\}$ is a random variable such that $\xi_i = 1$ iff i is a big element. Note that $\mathbb{E}(\xi) = M w_{\mathrm{big}} > \alpha$. Using Chebyshev's inequality we get

$$\Pr\left[\xi < n\right] \leq \Pr\left[|\xi - \mathbb{E}(\xi)| \geq \frac{1}{2}\mathbb{E}(\xi)\right] \leq \frac{4\,\mathrm{Var}\,(\xi)}{\mathbb{E}(\xi)^2} \leq \frac{4}{\mathbb{E}(\xi)} \leq \frac{4}{\alpha} \leq \frac{1}{2}. \tag{39}$$

□

5 Quantum algorithm for testing orthogonality

Consider distributions $p, q \in \mathcal{D}_N$ and let $S = (i_1, \ldots, i_M)$ be a list of M independent samples drawn from p. Let $A \subseteq [N]$ be the set of all elements that appear in S at least once. Define the *collision probability*

$$q_A = \sum_{i \in A} q_i.$$

Note that q_A is a deterministic function of A, so the probability distribution of q_A is determined by probability distribution of A (which depends on p and M). For a fixed A the variable q_A is the probability that a sample drawn from q belongs to A.

Clearly if p and q are orthogonal then $q_A = 0$ with probability 1. On the other hand, if p and q have a constant overlap, we will show that q_A takes values of order M/N with constant probability. Specifically, we shall prove the following lemma.

Lemma 6. *Consider a pair of distributions $p, q \in \mathcal{D}_N$ such that $\|p - q\|_1 \le 2 - \epsilon$. Let q_A be a collision probability constructed using M samples. Suppose $M \ge 2^9 \epsilon^{-2}$. Then*

$$\Pr\left[q_A \ge \frac{\epsilon^3 M}{2^{11} N}\right] \ge \frac{1}{2}. \tag{40}$$

It suggests the following algorithm for testing orthogonality.

OTest(p, q, M, K)
Let $S = \{i_1, \ldots, i_M\}$ be a list of M independent samples drawn from p.
Let $A \subseteq [N]$ be the set of elements that appear in S at least once.
Let $q_A = \sum_{i \in A} q_i$ be the total probability of elements in A with respect to q.
Let $\tilde{q}_A$ be estimate of q_A obtained using **EstProb**(q, A, K).
If $\tilde{q}_A \ge \frac{\epsilon^3 M}{2^{12} N}$ then reject. Otherwise accept.

We note that if $q_A = 0$ then $\tilde{q}_A = 0$ with certainty (see Theorem 5) and so **OTest** accepts any pair of orthogonal distributions with certainty. Theorem 3 is a direct consequence of the following lemma.

Lemma 7. *Choose*

$$M = K = O\left(\frac{N^{1/3}}{\epsilon}\right). \tag{41}$$

Then **OTest**(p, q, M, K) *rejects any distributions $p, q \in \mathcal{D}_N$ such that $\|p - q\|_1 \le 2 - \epsilon$ with probability at least* $1/4$.

Proof. According Eq. (40), $q_A \ge \epsilon^3 M/(2^{11} N)$ with probability $\ge 1/2$. When this holds, the algorithm rejects whenever

$$|\tilde{q}_A - q_A| \le \frac{q_A}{2}$$

since this implies $\tilde{q}_A \geq q_A/2 \geq \epsilon^3 M/(2^{12}N)$. Applying Theorem 5 with precision $\delta = q_A/2$ and error probability $\omega = 1/2$, we find (according to Eq. (4)), that K should be

$$K \geq \Omega\left(\frac{1}{\sqrt{q_A}}\right) \tag{42}$$

Taking into account Eq. (40) it suffices to choose

$$K = \Omega\left(\frac{N^{1/2}}{\epsilon^{3/2}M^{1/2}}\right)$$

to guarantee that **Otest** outputs 'reject' with probability at least $(1/2)\cdot(1/2) = 1/4$. Minimizing the total number of queries $K + M$ we arrive at Eq. (41). □

In the rest of this section we prove Lemma 6.

Proof. Begin by defining two sets of indices:

$$B \equiv \{i : q_i < \frac{\epsilon}{4}p_i\} \tag{43}$$

$$C \equiv \{i : p_i \leq \frac{\epsilon}{32}N^{-1}\} \tag{44}$$

Let B^c, C^c denote the complements of B and C respectively. We will prove that

$$\Pr\left[|A \cap B^c \cap C^c| \geq \frac{\epsilon}{16}M\right] \geq 1/2, \tag{45}$$

which will imply the Lemma since

$$q_A \geq \sum_{i\in A\cap B^c\cap C^c} q_i \geq \frac{\epsilon}{4}\sum_{i\in A\cap B^c\cap C^c} p_i \geq \frac{\epsilon^2}{2^7 N}|A \cap B^c \cap C^c|. \tag{46}$$

First, we show that $|A \cap B|$ is likely to not be too big. Observe that $q_B < \frac{\epsilon}{4}p_B \leq \frac{\epsilon}{4}$. Next use the fact that $\frac{1}{2}\|p - q\|_1 = \max_{U\subset[N]} p_U - q_U \leq 1 - \frac{\epsilon}{2}$ to bound $p_B \leq 1 - \frac{\epsilon}{2} + \frac{\epsilon}{4} = 1 - \frac{\epsilon}{4}$. Now we state a Chernoff-Hoeffding bound.

Lemma 8. *Let $X_1, \ldots, X_M$ be independent $0, 1$ random variables with $X \equiv \sum_{i=1}^M X_i$. Then for any $\delta > 0$,*

$$\Pr\left[X \geq \mathbb{E}(X) + M\delta\right] \leq \exp(-2M\delta^2). \tag{47}$$

Recall that A consists of the unique elements of $S = \{i_1, \ldots, i_M\}$. For $j = 1, \ldots, M$, define $X_j = 1$ if $i_j \in B$ and $X_j = 0$ if not. Then $|A \cap B| \leq \sum_{j=1}^M X_j$, with the possibility of an inequality in case there are repeats. We can now use Lemma 8 with $\mathbb{E}(X_j) = p_B \leq 1 - \epsilon/4$ and $\delta = \epsilon/8$ to prove that

$$\Pr\left[|A \cap B| \geq \left(1 - \frac{\epsilon}{8}\right)M\right] \leq \exp\left(-2M\left(\frac{\epsilon}{8}\right)^2\right) = \exp\left(-\frac{M\epsilon^2}{32}\right). \tag{48}$$

Next, we observe that $p_C \leq \epsilon/32$. We can use the same method to show that $|A \cap C|$ is likely to not be too big. This time we define $X_j = 1$ iff $i_j \in C$, so that $|A \cap C| \leq \sum_{j=1}^{M} X_j$ and $\mathbb{E}(X_j) = p_C \leq \epsilon/16$. Setting $\delta = \epsilon/32$ we get

$$\Pr\left[|A \cap C| \geq \frac{\epsilon}{16} M\right] \leq \exp\left(-\frac{M\epsilon^2}{2^9}\right). \tag{49}$$

When $M \geq 2^9/\epsilon^2$, we can combine (48) and (49) to find that with probability $\geq 1/2$, both $|A \cap B^c| \geq \frac{\epsilon}{8} M$ and $|A \cap C^c| \geq (1 - \frac{\epsilon}{16})M$. Thus $|A \cap B^c \cap C^c| \geq \frac{\epsilon}{16} M$ with probability at least $1/2$. This establishes (45), and completes the proof of the lemma. □

6 Lower bounds

6.1 Sampling vs query complexity

Let $p \in \mathcal{D}_N$ be any distribution and $O : [S] \to [N]$ be an oracle generating p. Recall that p_i coincides with the fraction of inputs $s \in [S]$ such that $O(s) = i$. It does not matter which particular inputs s are mapped to i. The only thing that matters is the number of such inputs. Therefore one can choose an arbitrary permutation of inputs $\sigma : [S] \to [S]$ and construct a new oracle $O' = O \circ \sigma$ that generates the same distribution p. We shall see below that if a classical testing algorithm $\mathcal{A}$ gives a correct answer with high probability for any choice of S and σ then $\mathcal{A}$ cannot take any advantage from making adaptive queries to $\mathcal{O}$. Let us transform $\mathcal{A}$ into a 'sampling' algorithm $\mathcal{A}_s$ such that each query made in $\mathcal{A}$ is replaced by a random query drawn from the uniform distribution on $[S]$.

Lemma 9. *Let $\mathcal{A}$ be any classical testing algorithm and $p \in \mathcal{D}_N$ be some distribution such that $\mathcal{A}$ accepts (rejects) p with probability at least $2/3$ for any oracle $O : [S] \to [N]$ generating p. Then the corresponding sampling algorithm $\mathcal{A}_s$ accepts (rejects) p with probability at least $2/3$.*

Proof. Let $P_{acc}(\sigma)$ be a probability that $\mathcal{A}$ accepts while interacting with the oracle $O \circ \sigma$, where σ is a permutation on $[S]$. Without loss of generality $P_{acc}(\sigma) \geq 2/3$ for all σ. It implies that the average acceptance probability

$$P_{acc} = \frac{1}{S!} \sum_{\sigma} P_{acc}(\sigma) \geq \frac{2}{3}. \tag{50}$$

An execution of the algorithm $\mathcal{A}$ can be represented by a history of queries $Q = (s_1, \ldots, s_T) \in [S]^{\times T}$. Let $P(Q)$ be a probability that an execution of $\mathcal{A}$ leads to a history Q. We can assume without loss of generality that the output of $\mathcal{A}$ (accept or reject) is a deterministic function of Q. Let Ω_{acc} be a set of histories Q that make $\mathcal{A}$ to accept. We have $P_{acc}(\sigma) = \sum_{Q \in \Omega_{acc}} P(\sigma^{-1} Q)$, where

$$\sigma^{-1} Q \equiv (\sigma^{-1}(s_1), \ldots, \sigma^{-1}(s_T)),$$

and thus

$$P_{acc} = \sum_{Q \in \Omega_{acc}} \frac{1}{S!} \sum_{\sigma} P(\sigma^{-1} Q) \geq \frac{2}{3}.$$

Let $\bar{P}(Q) = \mathbb{E}(P(\sigma^{-1}Q))$ where σ is drawn from the uniform distribution. Let $U(Q)$ be the uniform distribution on the set $[S]^{\times T}$. We claim that

$$\|\bar{P} - U\|_1 = O(TS^{-1}). \tag{51}$$

Assume without loss of generality that all queries in Q are different. Then

$$\bar{P}(Q) = \frac{(S-T)!}{S!} = S^{-T}(1 + O(T^2/S)).$$

A probability that a history drawn from the uniform distribution contains two or more equal queries can be bounded by $O(T^2/S)$ and thus we arrive at Eq. (51). Therefore in the limit $S \to \infty$ the acceptance probability is at least 2/3 if Q is drawn from the uniform distribution. But this implies that the sampling algorithm $\mathcal{A}_s$ accepts p with probability at least 2/3. □

6.2 Reduction from the Collision Problem to testing Orthogonality

One can get lower bounds on the query complexity of testing Orthogonality using the lower bounds for the Collision problem [14]. Indeed, let $H : [N] \to [3N/2]$ be an oracle function such that either H is one-to-one (yes-instance) or H is two-to-one (no-instance). The Collision Problem is to decide which one is the case. It was shown by Aaronson and Shi [15] that the quantum query complexity of the Collision problem is $\Omega(N^{1/3})$. Below we show that the Collision problem can be reduced to testing Orthogonality[3]. It implies that testing Orthogonality requires $\Omega(N^{1/2})$ queries classically and $\Omega(N^{1/3})$ queries quantumly.

Indeed, choose a random permutation $\sigma : [N] \to [N]$ and define functions $O_p, O_q : [N/2] \to [3N/2]$ by restricting the composition $H \circ \sigma$ to the subsets of odd and even integers respectively:

$$O_p(s) = H(\sigma(2s-1)), \quad O_q(s) = H(\sigma(2s)), \quad s \in [N/2].$$

For any yes-instance (i.e. H is one-to-one), the distributions $p, q \in \mathcal{D}_{3N/2}$ generated by O_p and O_q are uniform distributions on some pair of disjoint subsets of $[3N/2]$; that is, p and q are orthogonal.

We need to show that for any no-instance (H is two-to-one) the distance $\|p - q\|_1$ takes values smaller than $2 - \epsilon$ with a sufficiently high probability for some constant ϵ.

Lemma 10. *Let $H : [N] \to [3N/2]$ be any two-to-one function. Let $\sigma : [N] \to [N]$ be a random permutation drawn from the uniform distribution. Then*

$$\Pr\left[\|p - q\|_1 \le \frac{7}{4}\right] \ge \frac{1}{2}.$$

Proof. Given the promise on H we can define a perfect matching $\mathcal{M}$ on the set $[N]$ (considered as a complete graph with N vertices) such that $H(u) = H(v)$ iff u and v are matched. Let $\mathcal{M}_\sigma = \sigma^{-1} \circ \mathcal{M}$. Clearly, $\mathcal{M}_\sigma$ is a random perfect matching on $[N]$ drawn from the uniform distribution on the set of all perfect matchings.

[3] In order to apply the lower bound proved in [15] one has to choose the range of H of size $3N/2$ rather than N which would be more natural.

Let $(u,v) \in \mathcal{M}_\sigma$ be some pair of matched vertices and $w = H(\sigma(u)) = H(\sigma(v))$. Note that if u and v have different parity then $p_w = q_w = 2/N$. On the other hand, if u and v have the same parity then $p_w = 4/N$, $q_w = 0$ or vice verse. Thus

$$\|p - q\|_1 = 2 - \frac{4}{N}\#\{(u,v) \in \mathcal{M}_\sigma \, : \, u \text{ and } v \text{ have different parity}\}. \tag{52}$$

A nice property of the uniform distribution on the set of perfect matchings on $[N]$ is that a conditional distribution given that $(u,v) \in \mathcal{M}_\sigma$ is the uniform distribution on the set of perfect matchings on $[N]\backslash\{u,v\}$. Thus we can generate $\mathcal{M}_\sigma$ using the following algorithm. Let $U \subseteq [N]$ be the set of all unpaired vertices (in the beginning $U = [N]$). Let U_{even} and U_{odd} be the subsets of all even and all odd integers in U. The algorithm starts from an empty matching $\mathcal{M}_\sigma = \emptyset$. Suppose at some step of the algorithm we have some matching $\mathcal{M}_\sigma$ and some sets of unpaired vertices $U = U_{even} \cup U_{odd}$. If $|U_{even}| \ge |U_{odd}|$ choose a random vertex $u \in U_{odd}$. If $|U_{even}| < |U_{odd}|$ choose a random vertex $u \in U_{even}$. Pair u with a random vertex $v \in U\backslash\{u\}$ and update

$$\mathcal{M}_\sigma \to \mathcal{M}_\sigma \cup \{u,v\}, \quad U \to U\backslash\{u,v\}$$

with the corresponding update for U_{even} and U_{odd}. After $N/2$ steps of the algorithm we generate a random uniform $\mathcal{M}_\sigma$.

By construction, at each step of the algorithm we pair a vertex u to a vertex v with the opposite parity with probability at least $1/2$. Thus the probability $P(k)$ of having a matching $\mathcal{M}_\sigma$ with less than k pairs having opposite parity is

$$P(k) \le \sum_{i=0}^{k} \binom{N/2}{i} 2^{-\frac{N}{2}+k} \le 2^{\frac{N}{2}[H(x)+x-1+o(1)]},$$

where $x = 2k/N$. One can check that $H(x) + x - 1 < 0$ for $x \le 1/8$ and thus $P(N/16) \le 1/2$ for sufficiently large N. Thus Eq. (52) implies that $\|p - q\|_1 \le 2 - 1/4 = 7/4$ with probability at least $1/2$. □

6.3 Classical lower bound for testing Uniformity

In this section we prove that classically testing Uniformity requires $\Omega(N^{1/2})$. A proof uses the machinery developed by Valiant in [8]. Valiant's techniques apply to testing *symmetric* properties of distributions, that is, properties that are invariant under relabeling of elements in the domain of a distribution. Clearly, Uniformity is a symmetric property.

We shall need two technical tools from [8], namely, the Positive-Negative Distance lemma and Wishful Thinking theorem (see Theorem 4 and Lemma 3 in [8]). Let us start from introducing some notations. Let $p \in \mathcal{D}_N$ be an unknown distribution and $S = (i_1, \ldots, i_M)$ be a list of M independent samples drawn from p. We shall say that S has a collision of order r iff some element $i \in [N]$ appears in S exactly r times. Let c_r be the total number of collisions of order r, where $r \ge 1$. A sequence of integers $\{c_r\}_{r\ge 1}$ is called a *fingerprint* of S. Define a probability distribution D_p^M on a set of fingerprints as follows: (1) draw k from the Poisson distribution $\mathrm{Poi}(k) = e^{-M}M^k/k!$. (2) Generate a list S of k independent samples drawn from p. (3) Output a fingerprint of S.

An important observation made in [8] is that a fingerprint contains all relevant information about a sample list as far as testing symmetric properties is concerned. Thus without loss of generality, a testing algorithm has to make its decision by looking only on a fingerprint of a sample list. Applying Positive-Negative Distance lemma from [8] to testing Uniformity we get the following result.

Lemma 11 (**[8]**). *Let u be the uniform distribution on $[N]$ and $p \in \mathcal{D}_N$ be any distribution such that $\|p-u\|_1 \geq 1$. If for some integer M*

$$\|D_p^M - D_u^M\|_1 < \frac{1}{12} \tag{53}$$

then Uniformity is not testable in M samples.

The second technical tool is a usable upper bound on the distance between the distributions of fingerprints. For any integer k define an k-th moment of p as

$$m_k(p) = \sum_{i=1}^{N} p_i^k. \tag{54}$$

Clearly $m_k(u) = N^{1-k}$ which is the smallest possible value of a k-th moment for distributions on $[N]$. Applying Wishful Thinking theorem from [8] to testing Uniformity we get the following result.

Lemma 12 (**[8]**). *Let $p \in \mathcal{D}_N$ be any distribution such that $\|p\|_\infty \leq \delta/M$ for some $\delta > 0$. Then*

$$\|D_p^M - D_u^M\|_1 \leq 40\delta + 10 \sum_{k\geq 2} M^k \, \frac{m_k(p) - N^{1-k}}{\lfloor k/2 \rfloor! \sqrt{1 + M^k \, m_k(p)}}. \tag{55}$$

Corollary 1. *Uniformity is not testable classically in $32^{-1} N^{1/2}$ queries.*

Proof. Consider a distribution

$$p_i = \begin{cases} 2/N & \text{if} \quad 1 \leq i \leq N/2, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly $\|p-u\|_1 = 1$ and

$$m_k(p) = 2^{k-1} N^{1-k}.$$

In particular, choosing $M = 2^{-a} N^{1/2}$ we have

$$M^k m_k(p) = 2^{-k(a-1)-1} \, N^{1-\frac{k}{2}} \leq 2^{-2a+1} \quad \text{for all } k \geq 2.$$

Taking into account that

$$\sum_{k\geq 2} \frac{1}{\lfloor k/2 \rfloor!} \leq 2(e-1) \leq 4$$

we can use Eq. (55) to infer that

$$\|D_p^M - D_u^M\|_1 \leq 40\delta + 10 \cdot 2^{-2a+3}. \tag{56}$$

Clearly, condition $\|p\|_\infty \le \delta/M$ can be satisfied for any constant $\delta > 0$ and sufficiently large N. Then Lemma 11 implies that Uniformity is not testable in M samples whenever $10 \cdot 2^{-2a+3} < 1/12$. It suffices to choose $a = 5$. Finally, Lemma 9 implies that Uniformity is not testable in M queries in the oracle model. $\square$

Acknowledgments

We are grateful to Ronald de Wolf for numerous comments that helped to improve the paper. We would like to thank Sourav Chakraborty for informing us about the results in [1]. S.B. received support from the DARPA QUEST program under contract no. HR0011-09-C-0047. A.W.H. was funded by EPSRC grant "QIP IRC" and is grateful to IBM and MIT for their hospitality while this work was being done. A.H received support from the xQIT Keck fellowship.

References

[1] S. Chakraborty, E. Fischer, A. Matsliah, and R. de Wolf, "Quantum Queries for Testing Distributions," unpublished.

[2] T. Batu, E. Fischer, L. Fortnow, R. Rubinfeld, and P. White, "Testing random variables for independence and identity," Proceedings of FOCS 42nd, p. 442 (2002).

[3] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," Proceedings of FOCS 41st, p. 259 (2000).

[4] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld, "The complexity of approximating the entropy," Proceedings of STOC 34th, p. 678 (2002).

[5] T. Batu, R. Kumar, and R. Rubinfeld, "Sublinear algorithms for testing monotone and unimodal distributions," Proceedings of STOC 36th, p. 381 (2004).

[6] O. Goldreich and D. Ron, "A sublinear bipartiteness tester for bounded degree graphs," STOC 1998, p. 289.

[7] T. Batu, "Testing properties of distributions," Ph.D. thesis, Cornell University, 2001.

[8] P. Valiant, "Testing symmetric properties of distributions," STOC 2008, p. 383.

[9] A. Sahai and S. Vadhan, "A complete promise problem for statistical zero-knowledge," FOCS 1997, p. 448.

[10] O. Goldreich and D. Ron, "Property testing in bounded degree graphs," STOC 1997, p. 406.

[11] S. Aaronson, Complexity Zoo, URL *http://qwiki.stanford.edu/wiki/ComplexityZoo*.

[12] D. Aharonov and A. Ta-Shma, "Adiabatic quantum state generation and Statistical Zero Knowledge," quant-ph/0301023.

[13] S. Aaronson, "Quantum lower bound for the collision problem," STOC 2002, p. 635.

[14] G. Brassard, P. Hoyer, and A. Tapp, "Quantum algorithm for the collision problem," quant-ph/9705002.

[15] S. Aaronson and Y. Shi, "Quantum lower bounds for the collision and the element distinctness problems," Journal of the ACM (JACM), Vol. 51, Iss. 4, p. 595 (2004)

[16] G. Brassard, P. Hoyer, and A. Tapp, "Quantum Counting," quant-ph/9805082.

[17] G. Brassard, P. Hoyer, M. Mosca, and A. Tapp, "Quantum Amplitude Amplification and Estimation," arXiv:quant-ph/0005055.

[18] M. Santha, "Quantum walk based search algorithms." In *Proc. of the 5th Theory and Applications of Models of Computation (TAMC08)*, Xian, April 2008, LNCS 4978, pp. 31–46, arXiv:0808.0059.

[19] M. Szegedy, "Quantum speed-up of Markov chain based algorithms." In *Proc. of the 45th FOCS*, pp. 32–41, 2004. arXiv:quant-ph/0401053.

[20] S. Bravyi, A. W. Harrow, A. Hassidim. "Quantum algorithms for testing properties of distributions." arXiv:0907.XXXX, 2009.