Testing Juntas

Eldar Fischer* Guy Kindler† Dana Ron‡ Shmuel Safra§
Alex Samorodnitsky¶

April 17, 2005

Abstract

We show that a boolean valued function over n variables, where each variable ranges in an arbitrary probability space, can be tested for the property of depending on only J of them using a number of queries that depends only polynomially on J and the approximation parameter ϵ. We present several tests that require a number of queries that is polynomial in J and linear in ϵ^{-1}. We show a non-adaptive test that has one-sided error, an adaptive version of it that requires fewer queries, and a non-adaptive two-sided version of the test that requires the least number of queries. We also show a two-sided non-adaptive test that applies to functions over n boolean variables, and has a more compact analysis.

We then provide a lower bound of $\tilde{\Omega}(\sqrt{J})$ on the number of queries required for the non-adaptive testing of the above property; a lower bound of $\Omega(\log(J+1))$ for adaptive algorithms naturally follows from this. In establishing this lower bound we also prove a result about random walks on the group Z_2^q that may be interesting in its own right. We show that for some $t(q) = \tilde{O}\left(q^2\right)$, the distributions of the random walk at times t and $t+2$ are close to each other, independently of the step distribution of the walk.

We also discuss related questions. In particular, when given in advance a known J-junta function h, we show how to test a function f for the property of being identical to h up to a permutation of the variables, in a number of queries that is polynomial in J and ϵ^{-1}.

*Faculty of Computer Science, The Technion, Haifa, Israel. Research supported by a Technion VPR fund – Dent Charitable Trust – non-military research fund, and by a joint Haifa University – Technion research fund.

†School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel.

‡Department of Electrical Engineering, Tel-Aviv University, Tel-Aviv, Israel; research supported by the Israel Science Foundation (grant number 32/00-1).

§School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel. Research supported by an Israeli Science Foundation grant and a United States – Israel Binational Science Foundation grant.

¶School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel.

1 Introduction

Combinatorial property testing deals with the following task: For a fixed property P and any given input f, one has to distinguish with high probability between the case where f satisfies P and the case where f is 'far' from satisfying it, accessing the least possible number of bits from the input.

A property P is said to be *ϵ-testable using q queries*, or simply (ϵ, q)-testable, if there exists a probabilistic algorithm that makes at most q queries on any given input f (it is assumed that the input is accessed using an oracle), such that

- if f satisfies P, then the algorithm accepts it with probability at least 2/3, and
- if f is ϵ-far from P, that is, if it must be changed in more than an ϵ-fraction of the places in order to make it satisfy P, then the algorithm rejects it with probability at least 2/3.

A testing algorithm is said to be 1-*sided* if it accepts with probability 1 (rather than 2/3) any input that satisfies P. A testing algorithm that determines all its queries in advance, and uses the answers only in deciding whether to accept the input (and not in planning some of the queries) is called a *non-adaptive test*.

The general notion of property testing was first formulated by Rubinfeld and Sudan [RS96], who were motivated mainly by its connection to the study of program checking. The study of this notion for combinatorial objects, and mainly for graphs, was introduced by Goldreich, Goldwasser and Ron [GGR98].

Property testing has recently become a very active research area, see for example the surveys [Ron01] and [Fis01]. In addition to its theoretical appeal, it emerges in the context of PAC learning [GGR98], program checking [RS96], probabilistically checkable proofs [ALM+98, AS98, RS97], approximation algorithms [GGR98] and more. Properties of boolean functions were given particular consideration from the point of view of property testing, and especially properties related to monotonicity [GGL+00, DGL+99, FLN+02]. Perhaps the work most closely related to ours is [PRS01]. That paper presents testing algorithms that perform $O(1/\epsilon)$ queries for the following properties of boolean functions: Being a singleton function (a function of a single variable), being a J-monomial (a conjunction of at most J literals), and being a monotone DNF function with a bounded number of terms.

1.1 Boolean functions and juntas

In this paper we consider properties of boolean functions over n variables, namely functions over n variables that admit only two values. It will be convenient for us to assume that the values of boolean functions range in $\{-1, 1\}$.

While some of our results consider functions over boolean variables, other results apply to functions over variables that range in general domains. When the type of the boolean function f being discussed is known, we denote the range of the i'th variable of f by Ω_i (in the case of boolean variables, $\Omega_i = \{0, 1\}$). Denoting

$$\mathcal{P}([n]) \stackrel{\text{def}}{=} \prod_{i=1}^{n} \Omega_i \ ,$$

we have that all the boolean functions that we consider here can be written in the form $\mathsf{f} : \mathcal{P}([n]) \to \{-1, 1\}$, and that any assignment x for such a function is a vector $(x_1, \ldots, x_n)$, where $x_i \in \Omega_i$ for every i. In the following we will also consider a probability measure μ_i associated with every Ω_i, and the corresponding product measure associated with $\mathcal{P}([n])$.

Juntas. The main property of boolean functions we focus on is that of depending on only J (or less) of the variables.

Definition 1 (juntas, dominating sets). *A boolean function* $\mathsf{f} : \mathcal{P}([n]) \to \{-1, 1\}$ *is called a* J-junta *if there exists a set* $\mathcal{J} \subseteq [n]$ *of size at most* J*, such that* $\mathsf{f}(x) = \mathsf{f}(y)$ *for every two assignments* $x, y \in \mathcal{P}([n])$ *that agree on* $\mathcal{J}$*, namely that satisfy* $x_i = y_i$ *for all* $i \in \mathcal{J}$*. In this case it is said that* f is dominated by $\mathcal{J}$*. Somewhat abusing notation,* $\mathcal{J}$ *is also referred to as* the junta that dominates f*.*

1.2 Preview of results

Knowing that a function depends on only a small number of variables can be especially useful in the context of learning. For various function classes there exist algorithms that are attribute efficient (*cf.* [Lit87, BHL95, BL89, UTW97]). That is, they have polynomial dependence on the number of relevant variables of the function being learned and only logarithmic dependence on the total number of variables. One should also mention here the work of [MOS02] concerning computationally efficient learning of such functions when the algorithm is restricted to uniform samples.

As part of this effort, [GTT99] presented an algorithm that, for any input function f over boolean variables, uses $O(J(\log(J+1)/\epsilon + \log n))$ queries to completely determine a J-junta that dominates a function f' which is ϵ-close to f, if such a J-junta exists. In particular, their algorithm can be used to test for the property of being a J-junta. We show here the existence of a test for being a J-junta, for functions over arbitrary product spaces, whose number of queries does not depend on n at all.

Theorem 1 (the main result). *For every fixed* J *the property of being a* J*-junta is* $(\epsilon, \mathrm{poly}(J)/\epsilon))$*-testable for any given* ϵ*.*

1.2.1 Almost juntas

Let us review the definition of testable properties, with respect to the property of being a J-junta. To prove that this property is ϵ-testable, a test is to be shown, that distinguishes between J-juntas, and functions that must be changed in more than an ϵ-fraction of the places in order for them to become J-juntas. This is made more formal and somewhat more general using the following definition of a function that is ϵ-close to being a junta. Instead of just counting the number of values of f that need to be changed in order to make it a J-junta, giving the same weight to the value at every assignment, we allow weighing the assignments using a product probability measure.

Definition 2 ((ϵ, J)-juntas). *Let $\mathsf{f} : \mathcal{P}([n]) \to \{-1, 1\}$ be a boolean function, and assume that the range Ω_i of every variable of f is equipped with a probability measure μ_i. This determines a probability measure $\mu_{[n]} = \prod_{i=1}^{n} \mu_i$ over $\mathcal{P}([n])$.*

f *is said to be an* (ϵ, J)-junta *if there exists a boolean J-junta $\mathsf{g} : \mathcal{P}([n]) \to \{-1, 1\}$ such that for a random assignment $x \in \mathcal{P}([n])$ (chosen according to $\mu_{[n]}$),*

$$\Pr[\mathsf{f}(x) = \mathsf{g}(x)] \geq 1 - \epsilon \ .$$

In terms of the above definition, an (ϵ, q)-test for the property of being a J-junta is given a product measure $\mu_{[n]}$ on a domain $\mathcal{P}([n]) = \prod_{i=1}^{n} \Omega_i$ (specifically, we assume that the testing algorithm can for each i, select a random element in Ω_i according to the distribution μ_i), and an oracle access to an input function $\mathsf{f} : \mathcal{P}([n]) \to \{-1, 1\}$. It then uses q queries to distinguish between the case where the input function is a J-junta, and the case where it is not an (ϵ, J)-junta. We require that the number of queries made to f is entirely independent of $\mu_{[n]}$.

Note that the above definition includes the standard case where f is defined over boolean variables – one should just take $\Omega_i = \{0, 1\}$ for every i, and μ_i to be the uniform measure over Ω_i. By supplying a biased measure μ_i for every i, a J-junta test can, however, use the same number of queries to distinguish between the case where a given f is a J-junta, and the case where it must be changed on a set of $\mu_{[n]}$-measure more than ϵ in order to become a J-junta. Applying our results for other probability measures μ_i, one can test functions over variables that range over non-boolean domains, even infinite ones.

1.2.2 Junta tests

In order to establish Theorem 1 we describe several testing algorithms. The first algorithm is non-adaptive, requires $O(J^4 \ln(J+1)/\epsilon)$ queries, and in addition is 1-sided. We also provide an adaptive variant of this algorithm that requires only $O(J^3 \ln^2(J+1)/\epsilon)$ queries. Another algorithm presented here is a non-adaptive variant of the first algorithm that has a 2-sided error, but requires only $O(J^2 \ln^2(J+1)/\epsilon)$ queries. In the case of functions over boolean variables, and where the product measure $\mu_{[n]}$ is uniform, we present a non-adaptive testing algorithm with a compact, algebraically oriented analysis, that makes $O(J^4 \ln(J+1)/\epsilon)$ queries and has a 2-sided error.

1.2.3 Lower bound

On the other hand, at least with regards to non-adaptive algorithms, we show that the query complexity has to be a power of J (the tilde notation in the following is used to hide polylogarithmic factors), even if the test is restricted to functions over boolean variables with respect to the uniform measure.

Theorem 2. *For every $\alpha > 0$, a non-adaptive $\left(\frac{1}{2} - \alpha, q\right)$-test for the property of being a J-junta requires at least $q \geq \tilde{\Omega}\left(\sqrt{J}\right)$ queries, even if restricted to functions over boolean variables equipped with the uniform measure over their domain.*

Recently, Chockler and Gutfreund [CG02] have proven a better $\Omega(J)$ lower bound, which holds for adaptive testing algorithms as well. However, the proof given here may have significance beyond the lower bound itself, since during its course we prove a result about random walks on the group Z_2^q that may be of independent interest. In addition, the proof here also provides a lower bound for permutation testing (see below) for an explicit interesting function.

1.2.4 Random walks

Given any (finite) group G and a distribution P on G, a *random walk on G with step distribution P* starts with the identity element, and at each step t, denoting its current position by X_t, picks a random element ξ_t of G according to P and goes to $X_{t+1} = \xi_t X_t$. This definition of a random walk generalizes the more familiar notion of a random walk on a Cayley graph of a group, which is obtained by setting P to be a uniform distribution on the elements of a generating set for G.

A fundamental result of Markov [Mar06] from 1906 (see also [AD86]) states that this random walk converges to the uniform distribution on G, unless P is concentrated on a coset. A more recent question of interest is to estimate the rate of convergence of the random walk to its limit distribution. It is easy to see that this rate depends on the step distribution P, and therefore all the results in this direction concentrate on particular families of distributions for which good bounds can be obtained.

Here we ask a different question: Given a distance parameter $\delta > 0$, when do the distributions of X_t and X_{t+c} (for an appropriate constant c) become δ-close to each other with respect to the variation distance? Here we give a bound for the group Z_2^q (and $c = 2$), that does not depend on the step distribution P.

We remark that for any $\delta < 2$, such a bound has a chance to hold only if the order of any element ξ of G divides c. Otherwise taking P to be concentrated in ξ will give a counterexample. In this sense, the following theorem is optimal, since it turns out that for Z_2^q we can choose $c = 2$. It is tempting to conjecture that for any finite group we may choose c to be the least common multiple of the orders of the elements (it seems possible that the argument we give for the proof of the theorem might be extended for a general finite Abelian group; the case of non-Abelian groups seems to be more challenging).

Theorem 3. *Let P be a distribution on Z_2^q, and let X be the random walk on Z_2^q with step distribution P. Let P_t be the distribution of X at step t. There is an absolute constant C, such that for every $\delta > 0$, if $t \geq C \cdot \frac{\log \frac{1}{\delta}}{\delta} \cdot q^2 \log^2(q+1)$ then $|P_t - P_{t+2}| \leq \delta$, where $|P_t - P_{t+2}|$ denotes the variation distance between the two distributions.*

1.2.5 Testing for being a permutation of a given function

Finally, we consider the question of testing that a function f is identical to a fixed function h up to a permutation of its variables. We only consider functions over boolean variables here, whose domains are equipped with the uniform measure. Similar questions were given consideration already in [PRS01]. Here we construct a test for any function h which is a J-junta that is given in advance.

Some notation about restrictions and permutations of vectors is needed for the exact formulation of this result: Suppose that $\mathcal{J} = \{j_1, \ldots, j_J\}$ is some subset of $[n]$, whose elements are given in ascending order, $j_1 < \cdots < j_J$. For every permutation $\sigma : [J] \to [J]$ and every vector $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, we denote by $x|_{\sigma(\mathcal{J})}$ the vector $x = (x_{j_{\sigma(1)}}, \ldots, x_{j_{\sigma(J)}}) \in \{0,1\}^J$.

Theorem 4. *Let* $\mathsf{g} : \{0,1\}^J \to \{-1,1\}$ *be a function. The property, that* $\mathsf{f}(x) = \mathsf{g}(x|_{\sigma(\mathcal{J})})$ *for some* $\mathcal{J} \subset [n]$ *of size* J *and some permutation* $\sigma : [J] \to [J]$, *is* $(\epsilon, \mathrm{poly}(\epsilon, J))$*-testable for every* ϵ.

1.3 Organization of the paper

We start with Section 2, where we give some preliminaries and notation required for the subsequent sections, and introduce the notion of the variation of a function f on a set I of coordinates.

Section 3 presents our first junta test, called the size test. It randomly partitions the coordinates of the input function f, and applies a simple test to each subset in the partition, to discover whether f depends on any of its coordinates. The size test is non-adaptive, and has a 1-sided error. In Section 4, we present two variants of the size test, which achieve better query complexity. One of these variants has a 1-sided error but is adaptive, and the other is non-adaptive but has a 2-sided error. In Section 5 we present another junta test, that is restricted to functions defined over the discrete cube with the uniform distribution. This test is 2-sided, and its query complexity does not match that of the first 2-sided test. However, its algebraic approach yields a nice and compact analysis.

We then provide the lower bound for non-adaptive junta testing in Section 6, deriving it from the result concerning random walks in Z_2^q that is also proven there. In Section 7 we show how to test a function f for the property of being identical to a permutation of a given function h. We end the presentation with Section 8, which contains a discussion of some possible directions for future research, and some open problems.

2 Preliminaries

First, let us define some notation that will simplify the following exposition.

Partial assignments. Suppose that $\mathsf{f} : \mathcal{P}([n]) \to \{-1,1\}$ is a boolean function, where $\mathcal{P}([n]) = \prod_{i=1}^n \Omega_i$, and each set Ω_i is equipped with a probability measure μ_i. Each element

$x \in \mathcal{P}([n])$ is thus an assignment to the variables of f, where the i'th coordinate of x determines the value of the i'th variable. To specify assignments for only some of the variables of f, we define for each set $I \subseteq [n]$ of coordinates,

$$\mathcal{P}(I) \stackrel{\text{def}}{=} \prod_{i\in I} \Omega_i$$

and equip it with the probability measure $\mu_I \stackrel{\text{def}}{=} \prod_{i\in I} \mu_i$. An element $w \in \mathcal{P}(I)$ is thus a partial assignment for the variables of f. Whenever an element $w \in \mathcal{P}(I)$ is chosen randomly, it is chosen with respect to μ_I unless stated otherwise.

Assignment manipulation. If $w \in \mathcal{P}(I)$ and $z \in \mathcal{P}(H)$ are two partial assignments, and I and H are disjoint, let $w \sqcup z \in \mathcal{P}(I \cup H)$ denote the partial assignment whose i'th coordinate is w_i if $i \in I$, and z_i if $i \in H$. For a set $I \subseteq [n]$ of coordinates and an assignment $x \in \mathcal{P}([n])$, it is possible to obtain a partial assignment by restricting x to the coordinates of I, obtaining $x|_I \in \mathcal{P}(I)$. For simplicity we somewhat abuse notation, writing $x \cap I$ instead of $x|_I$. Similarly, we let $x \setminus I \in \mathcal{P}([n] \setminus I)$ denote the partial assignment obtained from x by taking the coordinates from $[n] \setminus I$.

2.1 Probability, some notation and lemmas

We use $\mathbb{E}$ and $\mathbb{V}$ to denote expectation and variance respectively. Specifically, suppose that g is a function of the form $\mathsf{g} : \mathcal{P}([n]) \to \mathbb{R}$. For a fixed partial assignment $w \in \mathcal{P}([n] \setminus I)$, we denote by

$$\mathbb{E}_{z\in\mathcal{P}(I)}[\mathsf{g}(w \sqcup z)]$$

the expectation of the value of $\mathsf{g}(w \sqcup z)$, where z is randomly drawn according to μ_I. Whenever the context is clear we may also use the shorthand $\mathbb{E}_z[\mathsf{g}(w \sqcup z)]$. Similarly, we denote the variance of $\mathsf{f}(w \sqcup z)$ where w is fixed and z is distributed according to μ_I, by

$$\begin{aligned}\mathbb{V}_{z\in\mathcal{P}(I)}[\mathsf{g}(w \sqcup z)] &= \mathbb{E}_z\left[(\mathsf{g}(w \sqcup z))^2\right] - (\mathbb{E}_z[\mathsf{g}(w \sqcup z)])^2 \\ &= \mathbb{E}_z\left[\Big(\mathsf{g}(w \sqcup z) - \mathbb{E}_z\big[\mathsf{g}(w \sqcup z)\big]\Big)^2\right].\end{aligned}$$

The following lemma immediately follows from the law of conditional variance (the lemma is also not hard to prove directly).

Lemma 2.1 (conditional variance). *For every $\mathsf{g} : \mathcal{P}([n]) \to \mathbb{R}$, two disjoint sets $I_1 \subset [n]$ and $I_2 \subset [n]$, and $w \in \mathcal{P}([n] \setminus (I_1 \cup I_2))$,*

$$\mathbb{V}_{z_1\in\mathcal{P}(I_1), z_2\in\mathcal{P}(I_2)}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)] = \mathbb{E}_{z_1}\left[\mathbb{V}_{z_2}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)]\right] + \mathbb{V}_{z_1}\left[\mathbb{E}_{z_2}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)]\right].$$

Another lemma that will be important to our arguments is the following inequality relation between expectation and variance.

Lemma 2.2. *For every* $\mathsf{g} : \mathcal{P}([n]) \to \mathbb{R}$*, two disjoint sets* $I_1 \subset [n]$ *and* $I_2 \subset [n]$*, and* $w \in \mathcal{P}([n] \setminus (I_1 \cup I_2))$*,*

$$\mathbb{V}_{z_1}\left[\mathbb{E}_{z_2}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)]\right] \leq \mathbb{E}_{z_2}\left[\mathbb{V}_{z_1}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)]\right].$$

Proof. The proof follows directly from the definitions, together with one application of the Cauchy-Schwarz inequality, which implies that $\left(\mathbb{E}_z[h(w \sqcup z)]\right)^2 \leq \mathbb{E}_z\left[\left(h(w \sqcup z)\right)^2\right]$ for every $w \in \mathcal{P}([n] \setminus I)$ and $\mathsf{h} : \mathcal{P}([n]) \to \mathbb{R}$.

$$\begin{aligned}
\mathbb{V}_{z_1}\left[\mathbb{E}_{z_2}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)]\right] &= \mathbb{E}_{z_1}\left[\left(\mathbb{E}_{z_2}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)] - \mathbb{E}_{z_1,z_2}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)]\right)^2\right] \\
&= \mathbb{E}_{z_1}\left[\left(\mathbb{E}_{z_2}\left[\mathsf{g}(w \sqcup z_1 \sqcup z_2) - \mathbb{E}_{z_1}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)]\right]\right)^2\right] \\
&\leq \mathbb{E}_{z_1,z_2}\left[\left(\mathsf{g}(w \sqcup z_1 \sqcup z_2) - \mathbb{E}_{z_1}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)]\right)^2\right] \\
&= \mathbb{E}_{z_2}\left[\mathbb{V}_{z_1}[\mathsf{g}(w \sqcup z_1 \sqcup z_2)]\right].
\end{aligned}$$

■

2.2 Variation

We now turn to define a measure called *variation*, of the dependency of a function f on a given subset of its coordinates (variables). Although we are mostly interested in boolean functions, we define the variation for all real valued functions.

Definition 3 (variation). *Let* $\mathsf{f} : \mathcal{P}([n]) \to \mathbb{R}$ *be a real valued function, and fix a subset* $I \subseteq [n]$ *of coordinates. The* variation *of* f *on* I *is defined to be the expectation of the variance of the restrictions of the form* $\mathsf{f}(w \sqcup \cdot\,)$*, where* $w \in \mathcal{P}([n] \setminus I)$*. That is, we define*

$$\mathsf{Vr}_{\mathsf{f}}(I) \stackrel{\text{def}}{=} \mathbb{E}_{w \in \mathcal{P}([n] \setminus I)}\left[\mathbb{V}_{z \in \mathcal{P}(I)}[\mathsf{f}(w \sqcup z)]\right]$$

In the case of boolean valued functions, we have an alternative definition for the variation. The variation of f on a set I is proportional to the probability that f yields different values, when evaluated on two random assignments which differ only on coordinates from I.

Proposition 2.3. *Let* $\mathsf{f} : \mathcal{P}([n]) \to \{-1, 1\}$ *be a boolean function, and fix a set* $I \subseteq [n]$ *of coordinates. Let* $w \in \mathcal{P}([n] \setminus I)$ *and let* $z_1, z_2 \in \mathcal{P}(I)$ *be chosen independently at random. Then*

$$\mathsf{Vr}_{\mathsf{f}}(I) \stackrel{\text{def}}{=} 2\Pr[\mathsf{f}(w \sqcup z_1) \neq \mathsf{f}(w \sqcup z_2)]$$

Proof. It is easy to observe that for two identically distributed independent random variables X, Y, taking values in $\{-1, 1\}$, it holds that $\mathbb{V}(X) = 2\Pr[X \neq Y]$, and hence

$$\mathsf{Vr}_{\mathsf{f}}(I) = \mathbb{E}_w\left[\mathbb{V}_{z_1}[\mathsf{f}(w \sqcup z_1)]\right] = \mathbb{E}_w\left[2\Pr_{z_1,z_2}[\mathsf{f}(w \sqcup z_1) \neq \mathsf{f}(w \sqcup z_2)]\right] = 2\Pr[\mathsf{f}(w \sqcup z_1) \neq \mathsf{f}(w \sqcup z_2)]$$

■

The next proposition shows that the variation is monotone and sub-additive. We also note that for functions defined over the discrete cube with the uniform measure, the monotonicity and sub-additivity of the variation follow directly from the Fourier-analytic formula for the variation in Proposition 2.6 below.

Proposition 2.4 (monotonicity and sub-additivity). *Let* $\mathsf{f} : \mathcal{P}([n]) \to \mathbb{R}$*, and let* A *and* B *be subsets of* $[n]$*. Then*

$$\mathsf{Vr}_{\mathsf{f}}(B) \leq \mathsf{Vr}_{\mathsf{f}}(A \cup B) \leq \mathsf{Vr}_{\mathsf{f}}(A) + \mathsf{Vr}_{\mathsf{f}}(B).$$

Proof. Both cases are consequences of Lemma 2.1. We begin by proving the monotonicity of the variation. To make the formal argument, we let w be a random element in $\mathcal{P}([n] \setminus (A \cup B))$, let z_1 and z_2 be independent random elements in $\mathcal{P}(A)$ and $\mathcal{P}(B \setminus A)$ respectively, and let $y = z_1 \sqcup z_2$ be the resulting random element in $\mathcal{P}(A \cup B)$. Then

$$\begin{aligned}\mathsf{Vr}_{\mathsf{f}}(A \cup B) &= \mathbb{E}_w\Big[\mathbb{V}_y[\mathsf{f}(w \sqcup y)]\Big] = \mathbb{E}_w\Big[\mathbb{E}_{z_1}\big[\mathbb{V}_{z_2}[\mathsf{f}(w \sqcup z_1 \sqcup z_2)]\big] + \mathbb{V}_{z_1}\big[\mathbb{E}_{z_2}[\mathsf{f}(w \sqcup z_1 \sqcup z_2)]\big]\Big] \\ &\geq \mathbb{E}_w\Big[\mathbb{E}_{z_1}\big[\mathbb{V}_{z_2}[\mathsf{f}(w \sqcup z_1 \sqcup z_2)]\big]\Big] = \mathbb{E}_{w,z_1}\Big[\mathbb{V}_{z_2}[\mathsf{f}(w \sqcup z_1 \sqcup z_2)]\Big] = \mathsf{Vr}_{\mathsf{f}}(B)\end{aligned}$$

and we have the monotonicity property.

Having proven the monotonicity, we may assume in proving the sub-additivity property that the sets A and B are disjoint. Using the notation above, we now prove the sub-additivity with the aid of Lemma 2.2.

$$\begin{aligned}\mathsf{Vr}_{\mathsf{f}}(A \cup B) &= \mathbb{E}_w\Big[\mathbb{V}_y[\mathsf{f}(w \sqcup y)]\Big] = \mathbb{E}_w\Big[\mathbb{E}_{z_1}\big[\mathbb{V}_{z_2}[\mathsf{f}(w \sqcup z_1 \sqcup z_2)]\big] + \mathbb{V}_{z_1}\big[\mathbb{E}_{z_2}[\mathsf{f}(w \sqcup z_1 \sqcup z_2)]\big]\Big] \\ &\leq \mathbb{E}_w\Big[\mathbb{E}_{z_1}\big[\mathbb{V}_{z_2}[\mathsf{f}(w \sqcup z_1 \sqcup z_2)]\big] + \mathbb{E}_{z_2}\big[\mathbb{V}_{z_1}[\mathsf{f}(w \sqcup z_1 \sqcup z_2)]\big]\Big] \\ &= \mathbb{E}_{w,z_2}\Big[\mathbb{V}_{z_1}[\mathsf{f}(w \sqcup z_1 \sqcup z_2)]\Big] + \mathbb{E}_{w,z_1}\Big[\mathbb{V}_{z_2}[\mathsf{f}(w \sqcup z_1 \sqcup z_2)]\Big] = \mathsf{Vr}_{\mathsf{f}}(A) + \mathsf{Vr}_{\mathsf{f}}(B)\end{aligned}$$

as required. ∎

We close this section with a lemma that generalizes the sub-additivity of the variation, and plays a crucial role in the proof of Theorem 1.

Lemma 2.5 (diminishing marginal variation). *Let* $\mathsf{f} : \mathcal{P}([n]) \to \mathbb{R}$ *be a real valued function, and let* A, B, C *be disjoint subsets of* $[n]$*. Then*

$$\mathsf{Vr}_{\mathsf{f}}(A \cup B) - \mathsf{Vr}_{\mathsf{f}}(B) \geq \mathsf{Vr}_{\mathsf{f}}(A \cup B \cup C) - \mathsf{Vr}_{\mathsf{f}}(B \cup C).$$

Proof. In the following we let w be a random member of $\mathcal{P}([n] \setminus (A \cup B \cup C))$, and x, y, z be random members of $\mathcal{P}(A), \mathcal{P}(B), \mathcal{P}(C)$ respectively, all independent. We first note from the definitions and a direct use of Lemma 2.1 that

$$\begin{aligned}\mathsf{Vr}_{\mathsf{f}}(A \cup B) - \mathsf{Vr}_{\mathsf{f}}(B) &= \mathbb{E}_{w,z}\big[\mathbb{V}_{x,y}[\mathsf{f}(w \sqcup x \sqcup y \sqcup z)]\big] - \mathbb{E}_{w,x,z}\big[\mathbb{V}_y[\mathsf{f}(w \sqcup x \sqcup y \sqcup z)]\big] \\ &= \mathbb{E}_{w,z}\Big[\mathbb{V}_x\big[\mathbb{E}_y[\mathsf{f}(w \sqcup x \sqcup y \sqcup z)]\big]\Big],\end{aligned}$$

and similarly

$$\mathsf{Vr}_{\mathsf{f}}(A \cup B \cup C) - \mathsf{Vr}_{\mathsf{f}}(B \cup C) = \mathbb{E}_w\Big[\mathbb{V}_x\big[\mathbb{E}_{y,z}[\mathsf{f}(w \sqcup x \sqcup y \sqcup z)]\big]\Big].$$

A direct application of Lemma 2.2, over $\mathsf{g}(w \sqcup x \sqcup z) = \mathbb{E}_y[\mathsf{f}(w \sqcup x \sqcup y \sqcup z)]$, shows that

$$\mathbb{E}_w\Big[\mathbb{V}_x\big[\mathbb{E}_{y,z}[\mathsf{f}(w \sqcup x \sqcup y \sqcup z)]\big]\Big] \leq \mathbb{E}_{w,z}\Big[\mathbb{V}_x\big[\mathbb{E}_{y}[\mathsf{f}(w \sqcup x \sqcup y \sqcup z)]\big]\Big],$$

concluding the proof. ■

2.3 Norms, distances, and inner products

Although our main concern here is the set of boolean functions over $\mathcal{P}([n])$, it is useful to consider such functions as elements in the space of real-valued functions $\mathsf{f} : \mathcal{P}([n]) \to \mathbb{R}$. For such a function f, and any parameter $1 \leq q < \infty$, the normalized ℓ_q-norm of f is defined by

$$\|\mathsf{f}\|_q \stackrel{\text{def}}{=} \left(\mathbb{E}_{x \in \mathcal{P}([n])}\left[|\mathsf{f}(x)|^q\right]\right)^{1/q}$$

(x is randomly chosen in $\mathcal{P}([n])$ according to $\mu_{[n]}$). An inner product between two functions $\mathsf{f}, \mathsf{g} : \mathcal{P}([n]) \to \mathbb{R}$, is defined by

$$\langle \mathsf{f}, \mathsf{g} \rangle \stackrel{\text{def}}{=} \mathbb{E}_{x \in \mathcal{P}([n])}[\mathsf{f}(x)\mathsf{g}(x)]$$

This inner product is related to the ℓ_2 norm, satisfying $\langle \mathsf{f}, \mathsf{f} \rangle = \|\mathsf{f}\|_2^2$ for every real-valued function f.

We also define another norm, that is used in Section 6 to measure the distance between two probability measures $P, Q : \{0,1\}^n \to \mathbb{R}$ over the discrete cube. The *variation distance* between two such measures is defined by $|P - Q| \stackrel{\text{def}}{=} \frac{1}{2}\sum_{x \in \{0,1\}^n} |P(x) - Q(x)|$ (this is not related to the notion of variation discussed above).

2.4 Harmonic analysis

Let us now focus on functions defined over the discrete cube $\{0,1\}^n$, equipped with the uniform measure. Real-valued functions defined over this domain can be expressed by their Fourier expansion as follows.

Definition 4 (characters and weights). *Let $S \subseteq [n]$. The* character χ_S *is the function over $\{0,1\}^n$ defined by $\chi_S(x) \stackrel{\text{def}}{=} (-1)^{|x \cap S|}$ (in other words, $\chi_S(x) = -1$ if the number of 1's in $\langle x_i | i \in S\rangle$ is odd, and $\chi_S(x) = 1$ if it is even).*

Given a function $\mathsf{f} : \{0,1\}^n \to \mathbb{R}$, its expansion as a linear combination of characters

$$\mathsf{f}(x) = \sum_{S \subseteq [n]} \hat{\mathsf{f}}(S)\chi_S(x)$$

is called the Fourier expansion of f (such an expansion always exists and is unique, since the set of characters forms a linear basis for the set of real functions over $\{0,1\}^n$).

Properties of Characters. The set of all characters forms an orthonormal basis for the space of real-valued functions over $\{0,1\}^n$, with respect to the inner product defined above. In addition, every character χ_S satisfies $\chi_S(x \oplus y) = \chi_S(x)\chi_S(y)$ for every $x, y \in \{0,1\}^n$, where '$x \oplus y$' denotes the coordinate-wise addition of x and y in Z_2^n.

Variation and Fourier expansion. The variation of a function f, defined over the discrete cube, can be written in terms of its Fourier expansion as follows.

Proposition 2.6. *Let $\mathsf{f} : \{0,1\}^n \to \{-1,1\}$ be a Boolean function, where $\{0,1\}^n$ is equipped with the uniform measure, and let $I \subseteq [n]$ be a set of coordinates. Then*

$$\mathsf{Vr}_{\mathsf{f}}(I) = \sum_{S \cap I \neq \emptyset} \hat{\mathsf{f}}^2(S)$$

The proof of Proposition 2.6 is straightforward, and we omit it. Note, however, that this directly implies Proposition 2.4 and Lemma 2.5 for functions over the discrete cube (with the uniform measure).

Convolution. The *convolution* of two functions (or distributions) $\mathsf{f}, \mathsf{g} : \{0,1\}^n \to \mathbb{R}$ is denoted by $\mathsf{f} * \mathsf{g}$, and is defined by $(\mathsf{f} * \mathsf{g})(y) \stackrel{\text{def}}{=} \sum_{x \in \{0,1\}^n} \big(\mathsf{f}(x) \cdot \mathsf{g}(x \oplus y)\big)$. We will need the following important property of convolution:

$$\widehat{(\mathsf{f} * \mathsf{g})}(S) = 2^n \cdot \hat{\mathsf{f}}(S) \cdot \hat{\mathsf{g}}(S).$$

3 The size test

The size test, described here, is a one-sided non-adaptive $(\epsilon, \Theta(J^4 \ln(J+1)/\epsilon))$-test for the property of being a J-junta. The independence test, presented next, is its main component. Given a set I of coordinates, the independence test is used to determine whether a given boolean function f is independent of the coordinates in I. It is a simple two-query test as follows.

The independence test. Choose a random $w \in \mathcal{P}([n] \setminus I)$, and choose $z_1, z_2 \in \mathcal{P}(I)$ randomly and independently. Verify that $\mathsf{f}(w \sqcup z_1) = \mathsf{f}(w \sqcup z_2)$.

Properties of the independence test. It is obvious that the independence test always accepts if f is independent of the coordinates in I, and by Proposition 2.3 its rejection probability equals $\frac{1}{2}\mathsf{Vr}_{\mathsf{f}}(I)$.

If f is a J-junta, then it clearly has the following property: for every partition $I_1, \ldots, I_r$ of the set of coordinates, all but at most J of them have zero variation. Hence the independence test when applied to f must accept all but at most J of the subsets. This consideration motivates the following size test.

The size test. The test has two parameters, r and h, that are to be chosen later. The test first chooses a random partition $I_1, \ldots, I_r$ of the set $[n]$ of coordinates, by choosing for every $i \in [n]$ independently and uniformly the set I_j to which it belongs. It then identifies on which of the I_j's f has a non-negligible variation, using $2rh$ queries, by going over every j from 1 to r and applying h iterations of the independence test to I_j. If f is found to be dependent on more than J subsets, the test rejects, and otherwise it accepts.

Properties of the test. The size test obviously accepts every J-junta, thus having perfect completeness. We show in the next subsection that, for a proper setting of the parameters r and h, the size test rejects f with probability at least $1/2$ if it is not an (ϵ, J)-junta (since the test is 1-sided this can easily be amplified to $2/3$). Before we prove this, let us set the parameters r and h.

The parameters of the test. We set $r \stackrel{\text{def}}{=} 16J^2$ and $h \stackrel{\text{def}}{=} 4er(\ln(J+1)+2)/\epsilon = \Theta(J^2 \ln(J+1)/\epsilon)$. Hence overall the test makes $2rh = \Theta(J^4 \ln(J+1)/\epsilon)$ queries to f, as required.

3.1 Soundness of the size test

Assuming that f passes the test with probability $1/2$, we prove that f must be an (ϵ, J)-junta in two steps. We first take $\mathcal{J}$ to be the set of coordinates on which f has variation larger than some threshold t, and prove that $|\mathcal{J}| \leq J$. We then show that the total variation of f on coordinates outside $\mathcal{J}$ is bounded by 2ϵ. This implies, by a simple argument, that f is ϵ-close to a junta dominated by $\mathcal{J}$.

Let $t \stackrel{\text{def}}{=} \frac{2(\ln(J+1)+2)}{h} = \frac{\epsilon}{2er}$, and let $\mathcal{J}$ denote the set of all coordinates i for which $\mathsf{Vr}_{\mathsf{f}}(\{i\}) > t$. We also denote $\bar{\mathcal{J}} \stackrel{\text{def}}{=} [n] \setminus \mathcal{J}$.

Proposition 3.1. *If the size test succeeds on* f *with probability* $1/2$*, then* $|\mathcal{J}| \leq J$.

Proof. The key observation here is that if a set I of coordinates contains a member of $\mathcal{J}$, then the variation of f on that set is at least t (by Proposition 2.4), and therefore each iteration of the independence test on I detects this dependence with probability at least $t/2$.

Suppose, for the sake of contradiction, that $|\mathcal{J}| > J$. Since $r = 16J^2$, it is easy to verify that with probability at least $3/4$ the number of subsets in the partition $I_1, \ldots, I_r$ that contain an element from $\mathcal{J}$ is at least $J+1$. When this occurs, the probability that any of the first $J+1$ subsets which intersect $\mathcal{J}$ will *not* be identified by the size test is bounded by $(J+1)(1-t/2)^h \leq (J+1)e^{-\ln(J+1)-2} < 1/4$, since $h = 2(2+\ln(J+1))/t$. Overall we have that with probability at least $1/2$ the size test rejects. ∎

Having shown that $|\mathcal{J}| \leq J$, the proof of soundness will be completed by showing that f is ϵ-close to a junta dominated by $\mathcal{J}$. We actually show that $\mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}}) < 2\epsilon$. This is sufficient to complete the proof, according to the following proposition.

Proposition 3.2. *Let $\mathcal{J}$ be a set of coordinates satisfying $\mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}}) < 2\epsilon$. Then there exists a boolean function h, that depends only on coordinates from $\mathcal{J}$, and agrees with f on a set of assignments of measure at least $(1-\epsilon)$.*

Proof. We define the function $\mathsf{h} : \mathcal{P}([n]) \to \{-1, 1\}$ by

$$\mathsf{h}(x) \stackrel{\text{def}}{=} \text{sign}\left(\mathbb{E}_{z\in\mathcal{P}(\bar{\mathcal{J}})}\Big[\ \mathsf{f}\big(\,(x\cap\mathcal{J})\sqcup z\,\big)\ \Big]\right)$$

where we arbitrarily set $\text{sign}(0) \stackrel{\text{def}}{=} 1$.

It is easy to observe that h only depends on coordinates from $\mathcal{J}$. To show that f and h are equal for most assignments, we take x to be a random element in $\mathcal{P}([n])$, y to be random in $\mathcal{P}(\mathcal{J})$, z to be random in $\mathcal{P}(\bar{\mathcal{J}})$, and we assume that they are all independent. Then

$$\begin{aligned}
2\Pr_x[\mathsf{f}(x) = \mathsf{h}(x)] - 1 &= \mathbb{E}_x[\mathsf{f}(x)\mathsf{h}(x)] = \mathbb{E}_y\Big[\mathbb{E}_z[\mathsf{f}(y\sqcup z)\mathsf{h}(y\sqcup z)]\Big] \\
&= \mathbb{E}_y\Big[\mathbb{E}_z[\mathsf{f}(y\sqcup z)]\cdot\text{sign}\big(\mathbb{E}_z[\mathsf{f}(y\sqcup z)]\big)\Big] \\
&= \mathbb{E}_y\left[\Big|\mathbb{E}_z[\mathsf{f}(y\sqcup z)]\Big|\right] \geq \mathbb{E}_y\Big[\big(\ \mathbb{E}_z[\,\mathsf{f}(y\sqcup z)\,]\ \big)^2\Big] \\
&= \mathbb{E}_y\Big[1 - \mathbb{V}_z[\,\mathsf{f}(y\sqcup z)\,]\ \Big] = 1 - \mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}}) > 1 - 2\epsilon
\end{aligned}$$

This immediately implies

$$\Pr_x[\mathsf{f}(x) = \mathsf{h}(x)] \geq 1 - \epsilon,$$

which completes the proof. ■

3.1.1 Bounding $\mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}})$

It is left to show that $\mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}}) < 2\epsilon$. Assume otherwise, and let us prove that the test rejects with probability at least 1/2.

Idea of the proof. The sum $\sum_{j=1}^{r}\mathsf{Vr}_{\mathsf{f}}(I_j\setminus\mathcal{J})$ is never less than $\mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}})$, as follows from the sub-additivity of the variation (see Proposition 2.4). Since we assume that $\mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}}) \geq 2\epsilon$, we have

$$\sum_{j=1}^{r}\mathbb{E}[\mathsf{Vr}_{\mathsf{f}}(I_j\setminus\mathcal{J})] = \mathbb{E}\left[\sum_{j=1}^{r}\mathsf{Vr}_{\mathsf{f}}(I_j\setminus\mathcal{J})\right] \geq 2\epsilon$$

where the expectation is taken over the random choice of the partition. Using the fact that the (unconditioned on other sets) distribution of any set in the partition is equal to that of any other, it follows that for any fixed j,

$$\mathbb{E}[\mathsf{Vr}_{\mathsf{f}}(I_j\setminus\mathcal{J})] \geq 2\epsilon/r$$

Since I_j is a random set of coordinates, we can obtain a concentration property for its variation, using the fact that every coordinate can contribute at most t to the variation of

$I_j \setminus \mathcal{J}$. In fact, we show that $\mathsf{Vr_f}(I_j \setminus \mathcal{J})$ (and therefore $\mathsf{Vr_f}(I_j)$) is with high probability at least a sizable portion of the bound for its expectation. This implies that with high probability, there are many sets I_j in the partition whose variation is relatively high. Since such sets are detected with high probability by the independence test, the size test rejects f with high probability.

Definition 5. *A set I_j in the partition is said to be* detectable *if* $\mathsf{Vr_f}(I_j) \geq \frac{\epsilon}{er}$.

Lemma 3.3. *Fix j, $1 \leq j \leq r$. The probability that I_j is detectable, over the choice of the partition $I_1, \ldots, I_r$, is at least* $3/4$.

Before we prove Lemma 3.3, we show how it completes the proof of the soundness of the size test. Let q denote the probability that the number of detectable subsets in the partition is smaller than $r/4$. Since the number of detectable subsets is bounded by r, Lemma 3.3 implies that

$$\frac{1}{4}rq + r(1-q) \geq \mathbb{E}[\,\text{number of detectable } I_j\text{'s}\,] \geq \frac{3}{4}r$$

from which we have $q \leq 1/3$. Hence with probability at least $2/3$, there are at least $r/4 = 4J^2 > J+1$ subsets in the partition, whose variation is larger than $\epsilon/er = 2t$. The size test fails in this case with probability at least $15/16$, as follows from an argument similar to that in the proof of Proposition 3.1. Therefore, the size test rejects f with an overall probability at least $1/2$, as required.

It is only left to prove Lemma 3.3. The proof requires extending our tools concerning the variation of a function, and occupies the remainder of this section.

Proof of Lemma 3.3: As mentioned above, the expectation of the variation of f on $I_j \setminus \mathcal{J}$ is at least $2\epsilon/r$. Lemma 3.3 will follow by showing that with probability at least $3/4$, $\mathsf{Vr_f}(I_j \setminus \mathcal{J}) \geq \epsilon/er$.

I_j is a random subset, obtained by going over the coordinates $i \in [n]$ and taking each of them into I_j independently with probability $1/r$. We can thus view the random variable $\mathsf{Vr_f}(I_j \setminus \mathcal{J})$ as the sum of the gradual donation of every coordinate,

$$\mathsf{Vr_f}(I_j \setminus \mathcal{J}) = \sum_{i=1}^{n} \Big(\mathsf{Vr_f}\big([i] \cap (I_j \setminus \mathcal{J})\big) - \mathsf{Vr_f}\big([i-1] \cap (I_j \setminus \mathcal{J})\big) \Big)$$

In order to use standard deviation bounds for $\mathsf{Vr_f}(I_j \setminus \mathcal{J})$, we would like the summands on the right-hand side to be independent and bounded by a small number. Note that the i'th summand is zero if $i \in \mathcal{J}$, and if $i \notin \mathcal{J}$ then it is bounded by t, as follows from the sub-additivity of the variation (and of course, all the summands are non-negative). The summands are thus indeed bounded by a small number, but they are *not* independent. This is tackled by introducing a technical tool that we call the unique-variation. While related to the variation, the unique-variation of I_j can be written as the sum of independent non-negative bounded random variables.

Definition 6 (unique-variation). *Define the unique-variation (with respect to $\mathcal{J}$) of every coordinate $i \in [n]$ by*

$$\mathsf{Ur_f}(i) \stackrel{\text{def}}{=} \mathsf{Vr_f}([i] \setminus \mathcal{J}) - \mathsf{Vr_f}([i-1] \setminus \mathcal{J}),$$

where $[0]$ denotes the empty set. Now for every set $I \subseteq \mathcal{P}([n])$ define its unique-variation by

$$\mathsf{Ur_f}(I) \stackrel{\text{def}}{=} \sum_{i \in I} \mathsf{Ur_f}(i)$$

The following lemma shows that the unique-variation of a subset I bounds the variation of I from below.

Lemma 3.4. *For every set $I \subseteq [n]$ of coordinates, $\mathsf{Ur_f}(I) \leq \mathsf{Vr_f}(I \setminus \mathcal{J})$.*

Proof. In fact we show that the unique-variation of $I \subseteq [n]$ is bounded from above by $\mathsf{Vr_f}(I \setminus \mathcal{J})$. For every $i \in [n]$ and $I \subseteq [n]$, it follows from Lemma 2.5 that

$$\mathsf{Vr_f}([i] \setminus \mathcal{J}) - \mathsf{Vr_f}([i-1] \setminus \mathcal{J}) \leq \mathsf{Vr_f}(([i] \cap I) \setminus \mathcal{J}) - \mathsf{Vr_f}(([i-1] \cap I) \setminus \mathcal{J}),$$

by substituting $A = \{i\} \setminus (I \cup \mathcal{J})$, $B = ([i-1] \cap I) \setminus \mathcal{J}$ and $C = [i-1] \setminus (I \cup \mathcal{J})$ in its formulation. From this it follows that

$$\begin{aligned}
\mathsf{Ur_f}(I) &= \sum_{i \in I} \mathsf{Ur_f}(i) = \sum_{i \in I} \Big(\mathsf{Vr_f}([i] \setminus \mathcal{J}) - \mathsf{Vr_f}([i-1] \setminus \mathcal{J}) \Big) \\
&\leq \sum_{i \in I} \Big(\mathsf{Vr_f}(([i] \cap I) \setminus \mathcal{J}) - \mathsf{Vr_f}(([i-1] \cap I) \setminus \mathcal{J}) \Big) \\
&= \sum_{i=1}^{n} \Big(\mathsf{Vr_f}(([i] \cap I) \setminus \mathcal{J}) - \mathsf{Vr_f}(([i-1] \cap I) \setminus \mathcal{J}) \Big) = \mathsf{Vr_f}(I \setminus \mathcal{J}),
\end{aligned}$$

concluding the proof. ■

By the above lemma, it remains to show that $\Pr[\mathsf{Ur_f}(I_j) \leq \epsilon/er] < 1/4$ in order to complete the proof of Lemma 3.3.

Note that the unique-variation of the coordinates in $\mathcal{J}$ is zero, and that $\mathsf{Ur_f}(i) \leq \mathsf{Vr_f}(i) \leq t$ for coordinates i outside $\mathcal{J}$, as follows from the sub-additivity property of the variation. The unique-variation of I_j is therefore a sum of independent non-negative random variables, each of which is bounded by t, and its expectation is given by

$$\mathbb{E}[\mathsf{Ur_f}(I_j)] = \frac{1}{r} \sum_{i \in [n]} \mathsf{Ur_f}(i) = \mathsf{Vr_f}(\bar{\mathcal{J}})/r \geq 2\epsilon/r$$

We can therefore apply standard deviation bounds to it, such as the following Chernoff-like bound, proven in Appendix A.

Proposition 3.5. *Let $X = \sum_{i=1}^{l} X_i$ be a sum of non-negative independent random variables X_i, and denote the expectation of X by α. If every X_i is bounded above by t, then*

$$\Pr[X < \eta\alpha] < \exp\left(\frac{\alpha}{et}(\eta e - 1)\right)$$

for every $\eta > 0$.

Since $\mathbb{E}[\mathsf{Ur}_{\mathsf{f}}(I_j)] \geq 2\epsilon/r$, Proposition 3.5 yields

$$\Pr[\mathsf{Ur}_{\mathsf{f}}(I_j) < \epsilon/er] < \exp\left(-\frac{\epsilon}{ert}\right) = e^{-2} < 1/4\,,$$

thus completing the proof of Lemma 3.3. ∎

4 Improving the query complexity

In this section we present two tests for the property of being a J-junta, that obtain an improved query complexity relative to that of the size test presented in Section 3. The first test uses a simple adaptive search method in order to reduce the query complexity. The second test checks possibly overlapping groupings of the coordinates for independence; it is two-sided, namely it may also reject a J-junta with some small but positive probability.

4.1 Improving the query complexity using adaptivity

The size test applies several iterations of the independence test to every subset in the partition, in order to detect whether it has a non-negligible variation. Here we show how, using an adaptive search, it is possible to detect all the subsets in the partition that have non-negligible variation using fewer queries, reducing a factor of J in the query complexity.

Theorem 5. *Set $r = 16J^2$ (as in the size test). Then there exists an adaptive one-sided J-junta test, that uses*

$$\frac{32erJ(1+\log_2 r)\ln(32J(1+\log_2 r))}{\epsilon} = \Theta\big(J^3 \ln^2(J+1)/\epsilon\big)$$

queries.

Proof. The idea of the adaptive test is to speed up the finding of the subsets of the partition with non-negligible variation as follows: Instead of applying the independence test to each subset individually, we apply it to blocks, each of which is a union of several such subsets. If f is not found to depend on a block, then all of its elements are declared to be 'variation free' at once. When f is found to depend on a block, the algorithm divides the block into two equally sized sub-blocks, for which the process is repeated.

Definition 7 (blocks). *Fix a partition $I_1, \ldots, I_r$ of the coordinates. A set B of coordinates is called a* block*, if it is the union of a positive number of subsets in the partition. The* size *of the block is the number of subsets in the partition that take part in this union.*

The adaptive test. The adaptive test begins by randomly partitioning the coordinates into subsets $I_1, \ldots, I_r$. The test maintains, throughout its operation, a set $S = \{B_1, \ldots, B_l\}$ of at most J disjoint blocks with respect to this partition. The blocks in S supposedly contain all the sets I_j in the partition that have non-negligible variation. Initially S is set to have only one block which contains all coordinates, namely $S = \{[n]\}$. At each step, the test performs the following.

- If all the blocks in S are of size one, accept (in this case at most J elements of the partition supposedly have non-negligible variation).

- Otherwise, choose a block $B \in S$ whose size is maximal. Remove B from S, and partition it arbitrarily into two sub-blocks $B = B' \cup B''$, whose sizes differ by at most 1 (remember that the size of a block is the number of sets I_j that are contained in it).

- Apply $\frac{4er\ln(32J(1+\log_2 r))}{\epsilon}$ iterations of the independence test to B'. If f is found to depend on B', then insert B' into S, and otherwise discard it. Apply the same treatment to B''.

- If the size of S is now greater than J, reject (f depends on each of the subsets in S, so it cannot be a J-junta in this case). Otherwise continue to the next step.

The adaptive test obviously accepts with probability 1 if f is a J-junta. To bound the number of rounds, we note that if after round T the maximum size of the blocks is m, then clearly after round $T + J$ the maximum size of the blocks is no more than $\lceil \frac{m}{2} \rceil$. This implies that the algorithm terminates after at most $2J(1 + \log_2 r)$ steps, and that each step uses $\frac{16er\ln(32J(1+\log_2 r))}{\epsilon}$ queries. The total number of queries made is therefore as required.

To prove Theorem 5, it is left to show that if f passes the test with probability at least 1/2, then it is an (ϵ, J)-junta.

Proposition 4.1 (soundness). *If f passes the adaptive-test with probability $1/2$, then it is an (ϵ, J)-junta.*

Proof. Let $t = \frac{\epsilon}{2er}$ and let $\mathcal{J}$ be defined as the set of coordinates i for which $\mathsf{Vr_f}(\{i\}) > t$ (as in Subsection 3.1). It suffices to prove that $|\mathcal{J}| \leq J$ and that $\mathsf{Vr_f}(\bar{\mathcal{J}}) \leq 2\epsilon$. Assume on the contrary that this is not the case, and let us prove that the adaptive-test rejects with probability at least 1/2.

According to the proof of Proposition 3.1, if $|\mathcal{J}| > J$ then with probability at least 3/4 there are at least $J + 1$ subsets in the partition $I_1, \ldots, I_r$ whose variation is at least t. Moreover, it is shown in Subsection 3.1 that if $\mathsf{Vr_f}(\bar{\mathcal{J}}) > 2\epsilon$, then with probability at least 2/3 there are at least $J + 1$ subsets in the partition, whose variation is at least $\epsilon/er = 2t$. In both cases, with probability at least 2/3 there are at least $J + 1$ subsets in the partition whose variation is at least t.

To complete the proof we show that if there are at least $J + 1$ subsets with variation at least t in the partition $I_1, \ldots, I_r$ chosen by the adaptive test, then the probability that it accepts is at most 1/8. This holds since in order to accept, the test must at some point discard a block whose variation is at least t. The probability of discarding each such block is at most

$$(1-\frac{t}{2})^{\frac{4er\ln(32J(1+\log_2 r))}{\epsilon}} \le e^{-\ln(32J(1+\log_2 r))} = \frac{1}{32J(1+\log_2 r)}$$

The test encounters two blocks at each step, so summing over all steps bounds the probability that such a block is discarded throughout the test by 1/8. ■

This concludes the proof of Theorem 5. ■

4.2 Improving the query complexity using two-sidedness

In this subsection we present a test with a significantly reduced query complexity. It makes $\Theta(J^2\ln^2(J+1)/\epsilon)$ queries, reducing a J^2 factor in the query complexity of the size test. The test is two-sided, namely we allow it to reject a J-junta with probability at most 1/3, on the condition that it rejects any input that is not an (ϵ, J)-junta with probability at least 2/3.

Theorem 6. *Let $\epsilon > 0$ be any positive number, and fix $r \stackrel{\text{def}}{=} 16J^2$, $s \stackrel{\text{def}}{=} 20J(3+\ln r)$, and $h \stackrel{\text{def}}{=} \frac{6er(3+2\ln s)}{\epsilon J}$. Then there exists a non-adaptive J-junta test, which makes $2sh = \Theta(J^2\ln^2(J+1)/\epsilon)$ queries, and satisfies the following.*

- *Every J-junta is accepted with probability at least* 2/3.
- *Any input which is not an (ϵ, J)-junta is rejected with probability at least* 2/3.

Proof. As in the size test, the two-sided test randomly partitions the coordinates into r subsets. In order to reduce the number of queries, the two-sided test finds subsets in the partition that have non-negligible variation by applying the independence test to blocks of such subsets (see Definition 7), like the adaptive test presented above, only here these blocks are chosen differently and may overlap.

The two-sided test. First, the test randomly partitions the coordinates into r subsets $I_1, \dots, I_r$. Then it picks s random subsets $\Lambda_1, \dots, \Lambda_s \subseteq [r]$ of size J independently, each by uniformly choosing without repetitions J members of $[r]$. Each set Λ_l determines a block $B_l \stackrel{\text{def}}{=} \bigcup_{j\in\Lambda_l} I_j$, to which the test applies h iterations of the independence test.

Acceptance conditions. The test declares a block B_l to be variation-free if none of the independence test iterations applied to B_l finds f to depend on it. If B_l is declared variation-free, then all the subsets I_j contained in it are declared to be variation-free on its behalf. The test accepts f if both of the following conditions hold.

- At least half of the blocks $B_1, \dots, B_s$ are declared variation free.
- Except for at most J subsets, every subset in the partition $I_1, \dots, I_r$ is declared variation-free on behalf of *some* block.

Properties of the test. It is obvious that the test performs $2sh$ queries, as required. It is left to show that a J-junta is accepted by the test with probability at least $2/3$, and that an input which is not an (ϵ, J)-junta is rejected with probability at least $2/3$. This is proven in the next two lemmas.

Lemma 4.2 (completeness). *If* f *is a* J*-junta, then it passes the two-sided test with probability at least* $2/3$.

Proof. Fix any partition $I_1, \ldots, I_r$. If f is a J-junta, then it is independent of all subsets in the partition, except for at most J of them. Hence for any fixed l, the probability over the selection of the blocks that f is independent of B_l is at least

$$\binom{r-J}{J}\Big/\binom{r}{J} > \left(\frac{r-2J}{r-J}\right)^J = \left(1 - \frac{J}{r-J}\right)^J > 1 - \frac{J^2}{r-J} \geq \frac{14}{15}$$

The probability that f depends on more than half of the blocks is therefore smaller than $\frac{2}{15} < \frac{1}{6}$, using the Markov inequality. Hence with probability at least $1 - \frac{1}{6}$, at least half of the blocks are declared variation-free, and the first acceptance condition holds.

Now fix j such that f does not depend on I_j, and let us bound the probability that it is not declared variation-free. Conditioned on the event that f does not depend on B_l, the probability that in addition B_l contains I_j is at least $J/r = 1/16J$. Hence I_j is declared variation-free on behalf of B_l with probability at least $1/20J$, for every fixed l. The probability that I_j is not declared variation-free is therefore bounded by

$$\left(1 - \frac{1}{20J}\right)^s = \left(1 - \frac{1}{20J}\right)^{20J(3+\ln r)} < \frac{1}{6r}$$

It follows that with probability at least $1 - \frac{1}{6}$, all the subsets in the partition on which f does not depend are declared variation-free (and in this case the second acceptance condition is fulfilled). Overall we have that with probability at least $2/3$, both conditions for acceptance are satisfied. ∎

Lemma 4.3 (soundness). *If* f *passes the two-sided test with probability higher than* $1/3$*, then it is an* (ϵ, J)*-junta.*

Proof. Let $t = \frac{\epsilon J}{3er}$ and let $\mathcal{J}$ denote the set of all coordinates i for which $\mathsf{Vr}_{\mathsf{f}}(\{i\}) > t$. As shown in Section 3, it suffices to prove that $|\mathcal{J}| \leq J$ and that $\mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}}) < 2\epsilon$. Assume on the contrary that this is not the case, and let us prove that the two-sided test rejects with probability at least $2/3$.

First case, $|\mathcal{J}| > J$. As in the proof of proposition 3.1, if $|\mathcal{J}| > J$ then with probability at least $3/4$ there are at least $J + 1$ subsets in the partition $I_1, \ldots, I_r$ with variation at least t. To conclude this case, we show that the probability of each such subset being declared variation-free is bounded by $\frac{1}{12(J+1)}$.

Let I_j be a subset whose variation is at least t, and let B_l be a block that contains it. By the monotonicity of the variation we have $\mathsf{Vr}_{\mathsf{f}}(B_l) > t$, so each iteration of the independence test on B_l detects a dependency of f on B_l with probability at least $t/2$. The probability of B_l being declared variation-free is therefore bounded by

$$(1-t/2)^h = (1-t/2)^{2\cdot(3+2\ln s)/t} < \frac{1}{12s(J+1)}$$

Since I_j is contained in at most s blocks, the probability of it being declared variation-free is bounded by $1/12(J+1)$, as required.

Second case, $\mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}}) \geq 2\epsilon$. Let us fix one index l, and show that B_l has high variation with very high probability. This will imply that with high-probability, the number of blocks *not* declared variation-free is larger than $s/2$, and the test rejects.

It follows from the procedure of choosing the partition and the blocks, that B_l is in fact a random set of coordinates, independently containing each coordinate $i \in [n]$ with probability J/r (to see this, note that its choice is equivalent to first choosing Λ_l and only then choosing the partition $I_1, \ldots, I_r$). We now consider the unique-variation as in Definition 6, only with respect to the set $\mathcal{J}$ as defined here. Then the expectation of $\mathsf{Ur}_{\mathsf{f}}(B_l)$ is given by

$$\mathbb{E}[\mathsf{Ur}_{\mathsf{f}}(B_l)] = \frac{J}{r}\sum_{i\in[n]} \mathsf{Ur}_{\mathsf{f}}(i) = \frac{J}{r}\mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}}) \geq 2\epsilon J/r$$

Moreover, the unique-variation of B_l is a sum of non-negative independent random variables, each bounded by t. It thus follows from Lemma 3.4 and Proposition 3.5 that

$$\Pr\left[\mathsf{Vr}_{\mathsf{f}}(B_l) < \frac{\epsilon J}{er}\right] \leq \Pr\left[\mathsf{Ur}_{\mathsf{f}}(B_l) < \frac{\epsilon J}{er}\right] < \exp\left(-\frac{\epsilon J}{ert}\right) = e^{-3} < 1/12$$

We say that a block B_l is detectable if its variation is at least $\epsilon J/er$. The expected number of non-detectable blocks is therefore smaller than $s/12$. It follows from the Markov inequality that with probability at least $1-\frac{1}{6}$, there are less that $s/2$ non-detectable blocks, and therefore there are more than $s/2$ detectable blocks. The probability of a detectable block being declared variation-free is bounded by

$$\left(1-\frac{\epsilon J}{2er}\right)^h < \exp\left(-(9+6\ln s)\right) < \frac{1}{6s} \quad ,$$

and therefore with probability at least $1-\frac{1}{6}$, none of the detectable blocks are declared variation-free. Overall we have that with probability at least $2/3$, the number of detectable blocks is more than $s/2$, and none of them is declared variation-free, and therefore the test rejects. ■

This concludes the proof of Theorem 6. ■

5 The compact test

In this section we describe and analyze a two-sided $(\epsilon, O(J^4 \ln(J+1)/\epsilon))$-test for the property of being a J-junta. This test is restricted to boolean functions defined over the discrete cube (namely $\mathcal{P}([n]) = \{0,1\}^n$) with the uniform measure. The algebraic approach of this test, combined with the fact that we do not insist on a 1-sided error, allows for a more compact analysis.

An overview of the testing algorithm. Let $\mathsf{f} : \{0,1\}^n \to \{-1,1\}$ be a J-junta. Let $V \stackrel{\text{def}}{=} V(\mathsf{f})$ be the set of all elements $v \in \{0,1\}^n$ that are 0 on all the variables that f depends on. Then V is clearly a subspace of $\{0,1\}^n$ (when viewed as a vector-space over the field $\{0,1\}$) of co-dimension at most J, and, moreover, it is an *ideal* under the bitwise AND operation, namely $x \in V$ implies that $x \wedge y \in V$ for every y. The crucial property of V is that any $x \in V$ is an *invariant shift* for f: for any $z \in \{0,1\}^n$ we have $\mathsf{f}(x \oplus z) = \mathsf{f}(z)$. Given f, our test looks for evidence to the existence a large ideal of invariant shifts for f. Specifically, we sample points in $\{0,1\}^n$ and check whether they lie in such an ideal V. Since $|V|$ could be exponentially small in relation to $|\{0,1\}^n|$, we sample according to a biased product distribution over $\{0,1\}^n$:

Definition 8. *Let $\mu_{1/J}$ denote the product measure on $\{0,1\}^n$, assigning to each bit 1 with probability $\frac{1}{J+1}$, and 0 with probability $1 - \frac{1}{J+1}$.*

It is easy to see that for any choice of a J-junta f we have $\mu_{1/J}(V) = (1 - \frac{1}{J+1})^J \geq e^{-1}$.

Given a point x chosen according to $\mu_{1/J}$, we randomly choose a logarithmic number of points $y \in \{0,1\}^n$ according to the uniform distribution. For each of these choices we test that $x \wedge y$ is an invariant shift for f by choosing uniformly at random a quadratic number of points $z \in \{0,1\}^n$, and checking whether $\mathsf{f}(z) = \mathsf{f}(z \oplus (x \wedge y))$.

Our testing algorithm will estimate the probability that a point x selected according to $\mu_{1/J}$ behaves like an invariant shift, and accept f only if this estimate is sufficiently large.

Alternative Algorithm for Testing J-juntas

Let C be a sufficiently large constant. Set $m = C \cdot J^2$, $t_1 = C \log(J+1)$, and $t_2 = C \cdot \frac{J^2}{\epsilon}$. We perform the following.

> Choose m points x according to $\mu_{1/J}$. For every selected x, choose t_1 points y uniformly from $\{0,1\}^n$. For each choice of x and y choose t_2 points z uniformly from $\{0,1\}^n$. All the choices are independent.
>
> For every selected point x check whether $\mathsf{f}(z) = \mathsf{f}(z \oplus (x \wedge y))$ for every z and y that were selected for x. If this equality holds for every z and y then we say that x *passed the check*.
>
> If the fraction of points x that passed the check is at least $(1 - \frac{1}{J+1})^J - \frac{1}{20(J+1)}$ then return "ACCEPT". Otherwise return "REJECT".

One observes that the query complexity of the algorithm is $O(m \cdot t_1 \cdot t_2) = O\left(J^4 \ln(J+1)/\epsilon\right)$, as required.

We next show that the test accepts every J-junta with probability at least $2/3$.

Definition 9. *For $v \in \{0,1\}^n$ let $s(v) \stackrel{\text{def}}{=} \Pr_z[\mathsf{f}(z) = \mathsf{f}(z \oplus v)]$.*

For $x \in \{0,1\}^n$, let $p(x)$ denote the probability that x passes the check, that is $\mathsf{f}(z) = \mathsf{f}(z \oplus (x \wedge y))$ for every z and y selected by the algorithm.

Let $p(\mathsf{f}) \stackrel{\text{def}}{=} \mathbb{E}_{x \sim \mu_{1/J}}[p(x)]$ be the probability that a point x selected according to $\mu_{1/J}$ passes the check.

Lemma 5.1 (completeness). *If f is a J-junta then the test returns "ACCEPT" with probability at least $2/3$.*

Proof. Note that $p(x) = 1$ for every $x \in V(\mathsf{f})$, and that $\mu_{1/J}(V(\mathsf{f})) \geq \left(1 - \frac{1}{J+1}\right)^J$. Therefore $p(\mathsf{f}) \geq \left(1 - \frac{1}{J+1}\right)^J$. By Chernoff's inequality, for a sufficiently large constant C, if we take $m = CJ^2$ points x, then with high probability, the fraction of points that pass the check is at least $\left(1 - \frac{1}{J+1}\right)^J - \frac{1}{20(J+1)}$, causing the test to return "ACCEPT" with high probability. ∎

5.1 Soundness of the compact test

From this point on we focus on showing that if f is accepted with probability greater than $1/3$, then it is ϵ-close to being a J-junta. Suppose that indeed the test returns "ACCEPT" with probability greater than $1/3$. Then by Chernoff's inequality, (assuming the constant C in the expression for $m = C \cdot J^2$ is sufficiently large), necessarily, $p(\mathsf{f}) \geq \left(1 - \frac{1}{J+1}\right)^J - \frac{1}{10(J+1)}$. The next definition will be useful in our analysis.

Definition 10. *For two points $v, x \in \{0,1\}^n$, we denote $v \leq x$ if $v_i \leq x_i$ for every $i \in [n]$. A point $x \in \{0,1\}^n$ is said to be* good*, if for a uniformly distributed $v \leq x$,*

$$\Pr_{v \leq x}\left[s(v) \geq 1 - \frac{\epsilon}{80(J+1)^2}\right] > \frac{1}{2}$$

Let $\mathcal{G} \subseteq \{0,1\}^n$ denote the set of all good x's.

Note that choosing $v \leq x$ uniformly is the same as choosing $y \in \{0,1\}^n$ uniformly and then setting $v = x \wedge y$. It is not hard to choose the constant C (defined in the testing algorithm) so that if x is not good, then $p(x) \leq \left(\frac{1}{2} + (1 - \frac{\epsilon}{80(J+1)^2})^{t_2}\right)^{t_1} \leq \left(\frac{3}{4}\right)^{t_1} \leq \frac{1}{10(J+1)}$. Let $\mathbf{1}_{\{\mathcal{G}\}}$ denote the characteristic function of the set $\mathcal{G}$. Then we have

$$\begin{aligned}
\mu_{1/J}(\mathcal{G}) &= \mathbb{E}_{x \sim \mu_{1/J}}[\mathbf{1}_{\{\mathcal{G}\}}(x)] \;\geq\; \mathbb{E}_{x \sim \mu_{1/J}}[\mathbf{1}_{\{\mathcal{G}\}}(x) \cdot p(x)] \\
&\geq p(\mathsf{f}) - \frac{1}{10(J+1)} \;\geq\; \left(1 - \frac{1}{J+1}\right)^J - \frac{1}{5(J+1)}
\end{aligned}$$

We now state our main claim, which, together with Proposition 3.2, completes the proof.

Claim 5.2. *If $\mu_{1/J}(\mathcal{G}) \geq \left(1 - \frac{1}{J+1}\right)^J - \frac{1}{5(J+1)}$, then there exists a set $\mathcal{J}$, $|\mathcal{J}| \leq J$, such that $\mathsf{Vr}_{\mathsf{f}}(\bar{\mathcal{J}}) \leq \frac{\epsilon}{2}$.*

The proof of Claim 5.2 requires the following lemmas.

Lemma 5.3. *For every $v \in \{0,1\}^n$, $\sum_{R:\, \chi_R(v)=-1} \hat{\mathsf{f}}^2(R) = 1 - s(v)$.*

Proof. Let $\mathsf{g}(x) \stackrel{\text{def}}{=} \mathsf{f}(x \oplus v)$. Then $\hat{\mathsf{g}}(R) = \hat{\mathsf{f}}(R) \cdot \chi_R(v)$ for every $R \subseteq [n]$. Consequently,

$$2s(v) - 1 = \langle \mathsf{f}, \mathsf{g} \rangle = \sum_{R \subseteq [n]} \hat{\mathsf{f}}(R)\hat{\mathsf{g}}(R) = \sum_{R:\, \chi_R(v)=1} \hat{\mathsf{f}}^2(R) - \sum_{R:\, \chi_R(v)=-1} \hat{\mathsf{f}}^2(R).$$

On the other hand, $\sum \hat{\mathsf{f}}^2(R) = 1$, so the above yields $1 - s(v) = \sum_{R:\, \chi_R(v)=-1} \hat{\mathsf{f}}^2(R)$. ■

Lemma 5.4. *If x is good, then $s(v) \geq 1 - \frac{\epsilon}{40(J+1)^2}$ for every $v \leq x$.*

Proof. Since by the definition of a good point x (Definition 10) more than half of the points $v \leq x$ satisfy $s(v) \geq 1 - \frac{\epsilon}{80(J+1)^2}$, every $v \leq x$ can be written as $v = v_1 \oplus v_2$ where v_1 and v_2 are two such points. Therefore

$$\begin{aligned} s(v) &= \Pr_z[\mathsf{f}(z) = \mathsf{f}(z \oplus v)] \;\geq\; \Pr_z\left[(\mathsf{f}(z) = \mathsf{f}(z \oplus v_1)) \wedge (\mathsf{f}(z \oplus v_1) = \mathsf{f}(z \oplus v_1 \oplus v_2))\right] \\ &\geq s(v_1) + s(v_2) - 1 \;\geq\; 1 - \frac{\epsilon}{40(J+1)^2}. \end{aligned}$$

■

Lemma 5.5. *Suppose x is good, and let $A_x \subseteq [n]$ be the set of coordinates i for which $x_i = 1$. Then*

$$\sum_{R \cap A_x \neq \emptyset} \hat{\mathsf{f}}^2(R) \leq \frac{\epsilon}{20(J+1)^2}$$

Proof. By combining Lemmas 5.4 and 5.3 we get that for all possible $w \leq x$ it holds that

$$\sum_{R:\, \chi_R(w)=-1} \hat{\mathsf{f}}^2(R) = 1 - s(w) \leq \frac{\epsilon}{40(J+1)^2}$$

Averaging this inequality over all $w \leq x$, and observing that for $R \cap A_x \neq \emptyset$, exactly half of the possible w satisfy $\chi_R(w) = -1$, we obtain

$$\sum_{R \cap A_x \neq \emptyset} \hat{\mathsf{f}}^2(R) = 2 \cdot \mathbb{E}_{w \leq x}\left[\sum_{R:\, \chi_R(w)=-1} \hat{\mathsf{f}}^2(R)\right] \leq \frac{\epsilon}{20(J+1)^2}.$$

■

Proof of Claim 5.2: Averaging the inequality of Lemma 5.5 over all x in $\mathcal{G}$, according to $\mu_{1/J}$ we obtain

$$\sum_R \hat{\mathsf{f}}^2(R) \cdot \frac{\mu_{1/J}\left(\{x \in \mathcal{G} \colon R \cap A_x \neq \emptyset\}\right)}{\mu_{1/J}(\mathcal{G})} \leq \frac{\epsilon}{20(J+1)^2} \,. \tag{1}$$

Now consider single coordinates i that satisfy

$$\frac{\mu_{1/J}\left(\{x \in \mathcal{G} \colon x_i = 1\}\right)}{\mu_{1/J}(\mathcal{G})} \leq \frac{1}{10(J+1)^2} \,. \tag{2}$$

We claim that there are at most J such coordinates. Indeed, if it were otherwise, let B be a set of $J+1$ such singletons, and for each $i \in B$ let $\mathcal{G}_i \stackrel{\text{def}}{=} \{x \in \mathcal{G} \colon\ x_i = 1\}$. Then $\mathcal{G} \setminus \cup_{i \in B} \mathcal{G}_i \subseteq \{x :\ x_i = 0 \colon \forall i \in B\}$, and therefore,

$$\begin{aligned}
\left(1 - \frac{1}{J+1}\right)^{J+1} &= \mu_{1/J}\left(\{x \colon\ \forall i \in B\ x_i = 0\}\right) \\
&\geq \mu_{1/J}\left(\mathcal{G} \setminus \cup_{i \in B} \mathcal{G}_i\right) \ \geq\ \mu_{1/J}\left(\mathcal{G}\right) \cdot \left(1 - \frac{1}{10(J+1)}\right) \\
&\geq \left(\left(1 - \frac{1}{J+1}\right)^J - \frac{1}{5(J+1)}\right) \cdot \left(1 - \frac{1}{10(J+1)}\right) \ >\ \left(1 - \frac{1}{J+1}\right)^{J+1}
\end{aligned}$$

and we reach a contradiction.

Let $\mathcal{J}$ be the family of coordinates that satisfy (2). We have shown $|\mathcal{J}| \leq J$. Now, for any $R \nsubseteq \mathcal{J}$,

$$\frac{\mu_{1/J}\left(\{x \in \mathcal{G} \colon R \cap A_x \neq \emptyset\}\right)}{\mu_{1/J}(\mathcal{G})} \geq \frac{1}{10(J+1)^2}$$

and therefore, by (1)

$$\sum_{R \cap \bar{\mathcal{J}} \neq \emptyset} \hat{\mathsf{f}}^2(R) \leq 10(J+1)^2 \cdot \frac{\epsilon}{20(J+1)^2} = \frac{\epsilon}{2},$$

completing the proof of the main claim. ∎

6 Lower bounds and a random walk on Z_2^q

To prove the lower bound we use Yao's principle, which states that to show a lower bound on the complexity of a randomized test, it is enough to present an input distribution for which any deterministic test with that complexity is likely to fail.

We define distributions D_P, D_N on positive (J-junta) and negative ($\frac{1}{2}$-far from any J-junta) input functions, respectively. Our input distribution first chooses D_P or D_N with equal probability and then draws an input according to the chosen distribution. We show that there exists a constant C such that every deterministic non-adaptive test with $q \leq C\sqrt{J}/\log(J)$

queries has an error probability larger than 1/3 (with respect to the induced probability on inputs). For this purpose we show that for any set of $q \leq C\sqrt{J}/\log(J)$ vertices of the hypercube, the distributions D_P and D_N induced on $\{-1,1\}^q$ by restricting the functions to these q vertices have a variation distance less than $\frac{1}{3}$.

The distributions D_P and D_N are simply uniform distributions over characters χ_S of size J and $J+2$ respectively. We will, however, work instead with two auxiliary distributions, $\tilde{D}_P$ and $\tilde{D}_N$, which are close to D_P and D_N, and which are easier to analyze. To choose a function from $\tilde{D}_P$, we first choose a random set $S \subseteq [n]$, $|S| \leq J$, in the following manner: We pick uniformly and independently J random elements in $[n]$ (with repetitions), and take S to be the set of elements that were selected an *odd* number of times. We then take the character χ_S to be our function. The distribution $\tilde{D}_N$ is defined in the same manner, only we start by picking $J+2$ elements in $[n]$.

Note that if $|S| > J$, then the character χ_S is $\frac{1}{2}$-far from any J-junta, and that both $\left|D_P - \tilde{D}_P\right|$ and $\left|D_N - \tilde{D}_N\right|$ are bounded by $O\left(\frac{J^2}{n}\right)$. Since Theorem 2 is stated for tests whose number of queries does not depend on n, we may and will assume in the following that n is large enough, i.e. that $J = o(\sqrt{n})$.

Now, consider the distributions induced by $\tilde{D}_P$ and $\tilde{D}_N$ on $\{-1,1\}^q$. Let $r_1, \ldots, r_q$ be the queries, and let M be a $q \times n$ boolean matrix, with rows $r_1, \ldots, r_q$. To choose an element x of $\{-1,1\}^q$ according to the first distribution, we choose at random, allowing repetitions, J columns of M and sum them up modulo 2. This gives us an element y of $\{0,1\}^q$. We take $x = (-1)^y$, where the power operation is performed coordinate-wise. The same holds for the second distribution, the only difference being that we choose $J+2$ columns.

For $x \in Z_2^q$, let $P(x)$ be the probability of choosing x when we pick a column of M at random. Consider a random walk on $Z_2^q \cong \{-1,1\}^q$, starting at 0, in which at every step we choose an element of the cube according to P and add it to the current location. Let P_t be the distribution induced by this walk after t steps. Note that P_J and P_{J+2} are precisely the distributions induced by $\tilde{D}_P$ and $\tilde{D}_N$. Note also that P_t is the distribution of $Y \oplus Y \oplus \ldots \oplus Y$, where we sum t independent copies of a Z_2^q-valued random variable Y, taking every value x with probability $P(x)$.

We want to show that for t sufficiently large compared to q, the distributions P_t, P_{t+2} are close in the variation distance. This is Theorem 3, presented in the introduction. Theorem 2 (see the introduction) now follows as an immediate corollary.

Theorem 3 is proven below. We first give a very brief overview of the proof. Every element x of Z_2^q defines a partition of the space into a subspace $V_0 = \{y : \langle y, x\rangle = 0\}$ and its complement V_1. We say that x is a *degenerate direction* if the probability of either of these sets according to P is at most $\tilde{O}\left(q^{-1}\right)$. The proof is inductive on the dimension q. We distinguish between two cases: if there are no degenerate directions, then the random walk is exponentially close to being stationary after $\tilde{O}\left(q^2\right)$ steps, and the claim holds. If, on the other hand, there is a degenerate direction x, then the walk 'splits' into two 'independent' walks, one on V_0 and one on V_1, each of which is isomorphic to Z_2^{q-1}, and we can use induction.

6.1 Proof of Theorem 3

Let us consider the distribution P_t of the walk at time t. Recall that the distribution of the sum of two independent random variables is the convolution of their distributions, $(P * Q)(x) = \sum_y P(y)Q(x \oplus y)$. This implies that P_t is the t-wise convolution of P, which we will denote by P^{*t}.

Now, for any $r \le t$ we have

$$|P_t - P_{t+2}| = \left|P^{*t} - P^{*(t+2)}\right| = \left|P^{*(t-r)} * \left(P^{*r} - P^{*(r+2)}\right)\right| = \left|P^{*(t-r)} * (P_r - P_{r+2})\right|.$$

The following fact is well-known and easy: for any two functions f, g on Z_2^q it holds that $\|\mathsf{f} * \mathsf{g}\|_1 \le 2^q \|\mathsf{f}\|_1 \|\mathsf{g}\|_1$. Taking into account that $P^{*(t-r)}$ is a distribution we deduce

$$\begin{aligned} |P_t - P_{t+2}| &= \left|P^{*(t-r)} * (P_r - P_{r+2})\right| = 2^{q-1} \cdot \|P^{*(t-r)} * (P_r - P_{r+2})\|_1 \\ &\le 2^{q-1} \cdot \|P_r - P_{r+2}\|_1 = |P_r - P_{r+2}|. \end{aligned}$$

Therefore, the distance $|P_t - P_{t+2}|$ is monotone non-increasing in t, and we are interested in the first time $t = t(q)$ for which P_t and P_{t+2} are δ-close. We show that $t(q) \le O\left(\frac{\log \frac{1}{\delta}}{\delta} \cdot b(q)\right)$, where we set $b(q) \stackrel{\text{def}}{=} q^2 \log^2(q+1)$.

In order to complete the proof of Theorem 3, we let S be the sum of the convergent series $\sum_{k=1}^{\infty} \frac{k}{b(k)}$, and show that there exists an absolute constant C, such that for any $t \ge C \frac{\log \frac{1}{\delta}}{\delta} \cdot b(q)$ and any distribution P on Z_2^q we have $|P_t - P_{t+2}| \le \frac{\delta}{S} \cdot \sum_{k=1}^{q} \frac{k}{b(k)}$.

The proof is by induction on q. We will assume, where needed, that C is sufficiently large. We set $t = C \frac{\log \frac{1}{\delta}}{\delta} \cdot b(q)$, assuming without loss of generality that this is an integer.

The case $q = 1$ is easy. It is possible to show that for a distribution P on Z_2 with $P(0) = p$ and $P(1) = 1 - p$, we have $|P_t - P_{t+2}| = \frac{1}{2} \cdot \left|(2p-1)^t - (2p-1)^{t+2}\right|$. A simple analysis shows that if $t \ge C \frac{\log \frac{1}{\delta}}{\delta}$, then the last expression is at most $\frac{\delta}{S}$.

Assume now that the claim holds for $q - 1$. We proceed with simple Fourier analysis, and show that our claim is true if all the non-zero Fourier coefficients of P are relatively small (a nice way to see this, though the actual proof is even simpler, is that this condition on the Fourier coefficients implies that P_t converges rapidly to the uniform distribution U, and $|P_t - P_{t+2}| \le |P_t - U| + |U - P_{t+2}|$). We have

$$\begin{aligned} |P_t - P_{t+2}|^2 &= 2^{2q-2} \cdot \|P_t - P_{t+2}\|_1^2 \le 2^{2q-2} \cdot \|P_t - P_{t+2}\|_2^2 \\ &= 2^{2q-2} \cdot \sum_R \left(\widehat{P}_t(R) - \widehat{P_{t+2}}(R)\right)^2 = \frac{1}{4} \cdot \sum_R \left(a^t(R) - a^{t+2}(R)\right)^2, \end{aligned} \tag{3}$$

where $a(R) \stackrel{\text{def}}{=} 2^q \hat{P}(R)$.

Clearly, $a(\emptyset) = \sum_x P(x) = 1$. Now consider the case in which, for all $R \ne \emptyset$ we have $|a(R)| \le 1 - \frac{\delta q}{\sqrt{C} b(q)}$. In this case, the right hand side of (3) is at most

$$\sum_{R \ne \emptyset} a^{2t}(R) \le 2^q \cdot \left(1 - \frac{\delta q}{\sqrt{C} b(q)}\right)^{2C \frac{\log \frac{1}{\delta}}{\delta} \cdot b(q)} \le 2^q \cdot exp\left\{-2\sqrt{C} \cdot \log \frac{1}{\delta} \cdot q\right\}.$$

This is smaller than $\frac{\delta}{S} \le \frac{\delta}{S} \cdot \sum_{k=1}^{q} \frac{k}{b(k)}$.

It remains to deal with the case where P has large Fourier coefficients. Let R be such that $|a(R)| \ge 1 - \frac{\delta q}{\sqrt{C} b(q)}$.

We now give a formal definition of a *degenerate direction*. We define x of Z_2^q to be degenerate if either $P\{V_0\}$ or $P\{V_1\}$ is at most $\frac{\delta q}{2\sqrt{C} b(q)}$. Here $V_0 = \{y : \langle y, x\rangle = 0\}$, and V_1 is the complement of V_0.

We claim that R is degenerate. Indeed $a(R) = P\{V_0\} - P\{V_1\}$. Therefore, if $a(R) \ge 0$ then $P\{V_1\} \le \frac{\delta q}{2\sqrt{C} b(q)}$. Otherwise, $P\{V_0\} \le \frac{\delta q}{2\sqrt{C} b(q)}$.

We make two assumptions for the sake of clarity: we assume that $R = e_1 \stackrel{\text{def}}{=} (10\cdots0)$, and that $a(R) \ge 0$. We omit the (straightforward) proof that both assumptions do not lead to loss of generality (for the second assumption it is indeed important that we compare P_t to P_{t+2}, and not to P_{t+1}).

Observe that the cube $\{0,1\}^q$ is now partitioned into two subcubes $V_0 = \{x : x_1 = 0\}$, and $V_1 = \{x : x_1 = 1\}$, both of which are isomorphic to Z_2^{q-1}. Because of the degeneracy of e_1, the walk will find it hard to leave the subcube it is in, and we will 'split' it into two walks, on V_0 and on V_1, and use the induction hypothesis for these walks.

For $i = 0, 1$ and for $r = t, t+2$ we set $P_r^i \stackrel{\text{def}}{=} (P_r | V_i)$. All four distributions so obtained can be viewed as distributions on Z_2^{q-1}.

We write P_t as a convex combination $P_t = P_t(V_0) \cdot P_t^0 + P_t(V_1) \cdot P_t^1$, and do the same for P_{t+2}. Note that $\big|P_t(V_0) - P_{t+2}(V_0)\big| \le \frac{\delta q}{\sqrt{C} b(q)}$. We will show, using the induction hypothesis, that for $i = 0, 1$ we have

$$\left|P_t^i - P_{t+2}^i\right| \le \frac{\delta}{S} \cdot \left(\sum_{k=1}^{q-1} \frac{k}{b(k)} + \frac{q}{2b(q)} \right).$$

This will conclude the proof, since

$$\begin{aligned} |P_t - P_{t+2}| &\le 2\big|P_t(V_0) - P_{t+2}(V_0)\big| + \big|P_t(V_0) \cdot \big(P_t^0 - P_{t+2}^0\big) + P_t(V_1) \cdot \big(P_t^1 - P_{t+2}^1\big)\big| \\ &\le \frac{\delta}{S} \cdot \left(\sum_{k=1}^{q-1} \frac{k}{b(k)} + \frac{q}{2b(q)} \right) + \frac{2\delta q}{\sqrt{C} b(q)} \le \frac{\delta}{S} \cdot \sum_{k=1}^{q} \frac{k}{b(k)}. \end{aligned}$$

Let $P^0 = (P|V_0)$ and $P^1 = (P|V_1)$. Let N_r be a random variable counting the number of times the walk makes a step in direction x with $x_1 = 1$ during the first r steps.

Let $i = 0$; the other case is treated similarly. The central (though simple) point of the argument is that for any r and for any even ℓ we have

$$(P_r^0 | N_r = \ell) = (P^1)^{*\ell} * (P^0)^{*(t-\ell)}.$$

This is true because the distribution on the left hand side is the distribution on Z_2^{q-1} given that the walk makes ℓ 'odd' steps, x with $x_1 = 1$, and $r - \ell$ 'even' steps, x with $x_1 = 0$. Since the addition in Z_2^q is commutative, we might as well assume that all the odd steps were made first, giving the right hand side.

Therefore, P_r^0 can be written as a convex combination

$$P_r^0 = \sum_{\ell \le r, \ell \text{ even}} \Pr(N_r = \ell) \cdot (P^1)^{*\ell} * (P^0)^{*(t-\ell)}.$$

Using this, we can bound $\left|P_t^0 - P_{t+2}^0\right|$:

$$\left|P_t^0 - P_{t+2}^0\right| \le \Pr(N_t \neq N_{t+2}) + Pr\left(N_t \ge \sqrt{C} \cdot \log\frac{1}{\delta} \cdot q\right) +$$

$$+ \sum_{\ell \le \sqrt{C} \cdot \log\frac{1}{\delta} \cdot q, \ell \text{ even}} \Pr(N_t = \ell) \cdot \left|(P^1)^{*\ell} * \left((P^0)^{*(t-\ell)} - (P^0)^{*(t+2-\ell)}\right)\right|. \quad (4)$$

The first summand in (4) is equal to the probability that an odd step was made in one of the times $t+1, t+2$, and this is at most $\frac{\delta q}{\sqrt{C} b(q)}$.

As to the second summand, observe that N_t is a binomial random variable with parameters $t = C^{\frac{\log \frac{1}{\delta}}{\delta}} \cdot b(q)$ and $p \le \frac{\delta q}{2\sqrt{C} b(q)}$. The probability of the second summand is that of $N_t \ge \sqrt{C} \cdot \log\frac{1}{\delta} \cdot q$, and this, using Chernoff bounds, is at most $exp\left\{-2\sqrt{C} \cdot \log\frac{1}{\delta} \cdot q/27\right\}$.

Thus, the sum of the two first summands is bounded from above by $\frac{\delta}{S} \cdot \frac{q}{2b(q)}$. It remains to deal with the third summand. For $\ell \le \sqrt{C} \cdot \log\frac{1}{\delta} \cdot q$ we have $t - \ell \ge C^{\frac{\log\frac{1}{\delta}}{\delta}} \cdot b(q) - \sqrt{C} \cdot \log\frac{1}{\delta} \cdot q \ge C^{\frac{\log\frac{1}{\delta}}{\delta}} \cdot b(q-1)$, and therefore we may use the induction hypothesis to conclude

$$\left|(P^1)^{*\ell} * \left((P^0)^{*(t-\ell)} - (P^0)^{*(t+2-\ell)}\right)\right| \le \left|(P^0)^{*(t-\ell)} - (P^0)^{*(t+2-\ell)}\right| \le \frac{\delta}{S} \cdot \sum_{k=1}^{q-1} \frac{k}{b(k)}.$$

Consequently, the third summand in (4) is bounded from above by $\frac{\delta}{S} \sum_{k=1}^{q-1} \frac{k}{b(k)}$, and

$$\left|P_t^0 - P_{t+2}^0\right| \le \frac{\delta}{S} \cdot \left(\sum_{k=1}^{q-1} \frac{k}{b(k)} + \frac{q}{2b(q)}\right),$$

concluding the proof of Theorem 3. ∎

7 Testing that f is a permutation of a given h

Given a boolean function $\mathsf{h} : \{0,1\}^n \to \{-1,1\}$, we say that a function f is a *permutation* of h if there exists a permutation $\sigma : [n] \to [n]$, such that for every $x = x_1 x_2 \ldots x_n \in \{0,1\}^n$ we have $\mathsf{f}(x) = \mathsf{h}(\sigma(x))$, where we define (with a slight abuse of notation) $\sigma(x) = x_{\sigma(1)} x_{\sigma(2)} \ldots x_{\sigma(n)}$. We show a test for this property for any h that is a J-junta. We first show a test with a linear dependence in ϵ^{-1} but an exponential dependence in J, and then show how to change it to a test with a polynomial dependence on both ϵ^{-1} and J. On the other hand, a closer look at the proof of Theorem 2 shows that it in fact proves something more than a lower bound on testing for being a J-junta: It also provides a lower bound, which depends on J, on testing that f is a permutation of $\mathsf{h}(x) = \chi_{[J]} = x_1 \oplus \ldots \oplus x_J$. This means that some dependence of the required number of queries in Theorem 4 on the junta size J is essential.

The tests constructed in the following are 2-sided. This is not a coincidence, since the following proposition shows that in some cases one needs a number of queries that is logarithmic in n to provide a non-adaptive 1-sided test for being a permutation of a given h. On the other hand, the results of [GTT99] can easily be used to construct a 1-sided *adaptive* test for this property, making a number of queries that is logarithmic in n (and depends on the junta size as well).

Proposition 7.1. *Any non-adaptive testing algorithm that makes less than* $\log(n/2)$ *queries on* $\mathsf{f}(x)$*, and accepts any permutation of* $\mathsf{h}(x) = x_1 \wedge x_2$ *with probability* 1*, will necessarily accept some permutation of* $\mathsf{h}'(x) = x_1$ *with probability at least* $\frac{1}{2}$*.*

Proof. Suppose that we are given a sequence of $l = \log(n/2)$ queries $q^{(1)}, \cdots, q^{(l)}$, where $q^{(i)}$ consists of querying the value of f at the point $(x_1^{(i)}, \ldots, x_n^{(i)})$. We define an equivalence relation over $\{1, \ldots, n\}$ by stating that $i \sim i'$ if for every $1 \leq j \leq l$ we have $x_i^{(j)} = x_{i'}^{(j)}$. We say that i is *isolated* if its equivalence class is $\{i\}$.

We observe that, by the choice of l, for every set of l queries there exists a set of at least $\frac{n}{2}$ coordinates that are not isolated. Thus, for every non-adaptive testing algorithm there exists a coordinate i, such that with probability at least $\frac{1}{2}$ it is not isolated with respect to the query sequence chosen by the algorithm (recall that a non-adaptive algorithm has to choose its query sequence in advance).

Now, for every query sequence $q^{(1)}, \ldots, q^{(l)}$ for which i is not isolated, and which is taken with positive probability by the algorithm, let i' be such that $i \sim i'$. Since the algorithm has to accept $\mathsf{f}(x) = x_i \wedge x_{i'}$ with probability 1, the algorithm must in particular accept this function when the sequence $q^{(1)}, \ldots, q^{(l)}$ is chosen. But this means that the algorithm must also accept the function $f'(x) = x_i$ when this query sequence is chosen, because these two functions are identical when restricted to the query sequence. Summing up over all query sequences for which i is not isolated, we conclude that the algorithm must accept $f'(x) = x_i$ with probability at least $\frac{1}{2}$, completing the proof. ∎

We now turn to the proof of Theorem 4. The constructed tests are adaptive, but they can be made non-adaptive with a penalty of an additional poly(J) factor. In addition, the second test can be made to work also for the case where the domain of f is Ω^n, for some finite Ω equipped with a (possibly biased) measure μ, only in this case the number of queries has to depend on $|\Omega|$ as well. These extensions are outlined at the end of the section. On a related issue, Appendix B contains an application of Theorem 1 to the question discussed in [PRS01] about testing that a function is a J-monomial.

7.1 A test with an exponential dependency on J

Before we continue, let us clarify first a notational convention used in the following. We would like to test f for the property of being a permutation of h, where h is a function with n variables that in fact depends only on a set $\mathcal{J}$ containing J (or fewer) of them. We now define g as the function on J variables defining the values of h, that is, the function for which $\mathsf{h}(x) = \mathsf{g}(x|_{\mathcal{J}})$ for every $x \in \{0,1\}^n$.

We assume without loss of generality that g depends on all its variables. In this case, it is not hard to see that the variation of g on every coordinate is at least 2^{1-J}. We begin by performing the J-junta test given by Theorem 1 on f, with $\min\{\frac{1}{4}\epsilon, 2^{-J}\}$ as the approximation parameter and $\frac{7}{8}$ as the detection probability (we go from $\frac{2}{3}$ to $\frac{7}{8}$ using the usual amplification techniques). If the test rejects then we reject the input. If the test accepts, we note that with high probability we have sets $I_{j_1}, \ldots, I_{j_l}$ of coordinates such that each of them contains exactly one member of a junta $\mathcal{J}$ of a function f' that is close to f (with $l \le J$), where $\mathcal{J}$ is the same set as the one defined in the proof of Theorem 1. If $l < J$ we reject the input (since g and hence h depend on exactly J coordinates), so from now on let us assume that $l = J$, and for convenience denote $V_k = I_{j_k}$ for $1 \le k \le J$.

For clarity, we first show how to test for the above property in the special case that g is symmetric with regards to permutations of its variables, and then show how to generalize this to every g. The idea is to check $\mathsf{f}(x)$ at a randomly chosen $x \in \{0,1\}^n$ for equality with $\mathsf{g}(x|_{\mathcal{J}})$, and repeat this $h = 12\epsilon^{-1}$ times so that any $\mathsf{f}(x)$ that is $\frac{1}{4}\epsilon$-far from $\mathsf{g}(x|_{\mathcal{J}})$ will be rejected with probability at least $\frac{7}{8}$. However, since we do not know $\mathcal{J}$ (but only $V_1, \ldots, V_J$), we perform the following procedure instead of a direct comparison.

The comparison procedure. Suppose that we are given a value $x \in \{0,1\}$, the sets of coordinates $V_1, \ldots, V_m$ (in the above context we have $m = J$), a function $\mathsf{g} : \{0,1\}^m \to \{-1,1\}$, and a parameter s. We denote by Z_x the set of the zero coordinates of x, namely $Z_x = \{i | x_i = 0\}$, and construct $y \in \{0,1\}^m$ as follows.

For every $1 \le k \le m$, we perform s iterations of the independence test for $V_k \cap Z_x$, and do the same for $V_k \backslash Z_x$. The idea is that if f is in fact a junta function dominated by $\mathcal{J}$, and each V_k holds exactly one coordinate of $\mathcal{J}$, then with sufficiently high probability we will know whether the only coordinate in $\mathcal{J} \cap V_k$ has received a value of 0 or 1.

For every k, if only $V_k \cap Z_x$ was found to have variation, we set $y_k = 0$. If only $V_k \backslash Z_x$ was found to have variation, we set $y_k = 1$. In the two other cases (where for the same k either both sets or none of them was found to have variation), we immediately reject the input f and terminate the entire algorithm. Having thus built $y \in \{0,1\}^m$, we now compare $\mathsf{f}(x)$ and $\mathsf{g}(y)$ and output the result.

A test for a symmetric g. We now show how to test that f is a permutation of h, where h is a junta function defined by a symmetric function g. After performing the junta test and constructing $V_1, \ldots, V_J$ as above, we perform $h = 12\epsilon^{-1}$ iterations of the comparison procedure. In every iteration we pick a uniformly random $x \in \{0,1\}^n$, and use the parameter $s = 3 \cdot 2^J(\log h + \log(2J) + 3)$.

We reject the input if any of the iterations of the comparison procedure has found a mismatch between f and g (or if any iteration of the comparison procedure rejected the input during the calculation of y from x), and otherwise we accept the input. Assuming that the junta test has succeeded in finding $V_1, \dots, V_J$ such that each of them contains exactly one member of the set $\mathcal{J}$ that includes all coordinates whose variation is at least 2^{-J}, it is not hard to see that with probability at least $\frac{7}{8}$ every comparison was in fact between $\mathsf{f}(x)$ and $\mathsf{g}(x|_{\mathcal{J}})$ (although we still do not know the identity of the members of $\mathcal{J}$) for the chosen x. Thus if f is ϵ-far from being a permutation of h then it was rejected in this stage with probability at least $\frac{3}{4}$, and on the other hand if f was in fact a permutation of h then it was accepted in this stage with probability at least $\frac{7}{8}$. To bound the success probability of the entire test we also have to subtract an additional $\frac{1}{8}$ for the possibility of failing to correctly construct $V_1, \dots, V_J$.

The general case. For a general (possibly asymmetric) g we need to consider all possible permutations of g for comparison with f. For every such permutation we perform $h = 12J\log(J+1)\epsilon^{-1}$ iterations of the comparison procedure, this time using $s = 3 \cdot 2^J(\log(J!) + \log h + \log(2J) + 3)$.

We use the same set of queries for every of the $J!$ (or less) possible permutations of g (noting that the way the comparison procedure chooses its queries is independent of the values of g). With probability at least $1 - J!(1/8J!) = \frac{7}{8}$ all instances of the comparison procedure will construct y correctly from x. Given this, with probability at least $1 - J!(1/8J!) = \frac{7}{8}$ we will detect the ϵ-farness of the input for any permutation of g for which it exists.

Our final testing algorithm accepts the input if there was any permutation of g for which a difference was not detected (unless at any time the comparison procedure itself rejected the input due to a failure in constructing y from x). Summing up, an input which is ϵ-far from being any permutation of h will be rejected with probability at least $\frac{5}{8}$, and an input which is a permutation of h will be accepted with probability at least $\frac{6}{8}$ (it could only be rejected if the J-junta test did not detect all the junta coordinates, or if in any of the constructions of y by the comparison procedure above, the dependence of f on $Z_x \cap V_k$ or on $V_k \setminus Z_x$ for the appropriate x was not detected correctly). It is not hard to amplify the first probability from $\frac{5}{8}$ to $\frac{2}{3}$. ■

7.2 Reducing the dependency on J

We construct here a test for f being a permutation of g using a polynomial number of queries. The running time itself is still exponential in J, however.

First, we perform the J-junta test with the approximation parameter $\frac{\epsilon}{6J}$ and detection probability $\frac{15}{16}$. We denote $I_{j_1}, \dots, I_{j_l}$ as before. However, after the size test we again use sufficiently many iterations of the independence test to distinguish between $\mathsf{Vr_f}(I_{j_k}) \geq \frac{\epsilon}{3J}$ and $\mathsf{Vr_f}(I_{j_k}) \leq \frac{\epsilon}{6J}$ with probability $\frac{15}{16J}$ for every k, and discard from $I_{j_1}, \dots, I_{j_l}$ also the sets whose variation is low.

Let us denote the remaining sets by $V_1, \dots, V_m$. Here we allow also for the possibility that $m < J$, as it could be the case that some sets containing junta coordinates (but with a small dependence of g on them) were not detected by the size test, or were discarded in the dependence rechecking phase. However, if m is smaller than the number of coordinates in g whose variation is at least $\frac{\epsilon}{3J}$, or larger than the number of coordinates in g whose variation is more than $\frac{\epsilon}{6J}$, then we reject the input, because such an outcome is inconsistent with the premise that f is indeed a permutation of h.

To accommodate our function g, we consider every permutation of any function $\tilde{\mathsf{g}}$ that can be constructed from g as follows: Let S be any subset of the coordinates of g of size m that contains all coordinates that have variation at least $\frac{\epsilon}{3J}$ and contains no coordinate that has variation at most $\frac{\epsilon}{6J}$ (with respect to g). We let $\tilde{\mathsf{g}} : \mathcal{P}(S) \to \{-1, 1\}$ be the following majority function, noting that it is no more than $\frac{1}{3}\epsilon$-far from g.

$$\tilde{\mathsf{g}}(y) = \begin{cases} 1 & \mathbb{E}_{z \in \mathcal{P}([J] \setminus S)}[\mathsf{g}(y \sqcup z)] \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

The total number of permutations of the functions that can be constructed from g as above is not more than $(J+1)!$. For every such function, noting that in particular each of its coordinates has variation at least $\frac{\epsilon}{6J}$ with respect to it, we use $h = 32J^2\epsilon^{-2}(4 + J\log(J+1))$ iterations of the comparison procedure, with the parameter $s = 36\epsilon^{-1}J\log(h(J+2)!)$. Assuming that there was no failure in the junta test or in the later picking of $V_1, \ldots, V_m$, every V_k contains exactly one coordinate whose variation with respect to f is at least $\frac{\epsilon}{6J}$. This implies (by the bound on the variation of the coordinates) that the construction of y from x will be correct in all iterations with probability at least $\frac{7}{8}$. As before, we use the same set of queries for every function that was derived from g.

Given that all comparisons were made correctly, for every function that was compared with f we can with probability at least $1 - 1/8(J+1)!$ distinguish between the case that the probability of $\mathsf{f}(x) = \tilde{\mathsf{g}}(y)$ is at least $1 - \frac{1}{3}\epsilon$, and the case that it is at most $1 - \frac{2}{3}\epsilon$ (this is done using a standard large deviation inequality, see for example [AS00, Appendix A]). We accept the input if at least one permutation of one of the possible $\tilde{\mathsf{g}}$ has the high probability for $\mathsf{f}(x) = \tilde{\mathsf{g}}(\sigma(y))$, assuming that it was not rejected by an iteration of the comparison procedure failing to produce y from x in any stage (but we may still accept the input if after constructing y some of the comparisons themselves discovered a mismatch). The correctness probabilities of this algorithm can be amplified to $\frac{2}{3}$ as usual. ∎

Variations on the permutation test. The following explains how to make the test non-adaptive. We note that the only place where information from previous queries is used in determining new queries is where the information concerning the identity of $V_1, \ldots, V_m$ is used for testing the independence of $V_k \cap Z_x$ and $V_k \setminus Z_x$ for every $1 \leq k \leq m$. If instead we make queries for testing the independence of $I_j \cap Z_x$ and $I_j \setminus Z_x$ for every $1 \leq j \leq r$ (whenever we use the comparison procedure), then we can place all those queries in advance, and later discard the ones corresponding to any I_j which is not one of $V_1, \ldots, V_m$. Similarly, when choosing which of the I_{j_k} to discard for determining $V_1, \ldots, V_m$ (right after the junta test), we can place in advance queries for every I_j and later discard the irrelevant ones. The above makes for a polynomial penalty in the total number of queries of the test.

As for making the test work also for non-boolean domains, this is done by changing the comparison test to check the independence of $V_k \cap Z_{\rho,x}$ for every $\rho \in \Omega$, where we define $Z_{\rho,x} = \{i | x_i = \rho\}$ (and where each x is now randomly chosen from Ω^n using μ^n). The change in the number of queries is a factor of $\tilde{O}(|\Omega|)$.

8 Open problems and remarks

Relaxing the soundness requirement. Other than making the test two-sided, it is also possible to obtain quadratic dependency on J by somewhat relaxing the soundness requirement. This is obtained if we only require that the test accepts every J-junta, and rejects inputs which are, say, not even $(\epsilon, 2J)$-juntas.

To achieve the quadratic dependency on J, note that in the original size test we have chosen the number of elements in the partition to be quadratic in J, so that any $J + 1$ influential coordinates would go into distinct subsets in the partition with high probability. If we allow juntas of size up to $2J$ to be accepted, it is enough to take a partition of size only linear in J. This reduces the number of queries by a factor of J. But since the subsets in the partition are now larger, we can take the 'junta threshold' t to be linear in $1/J$, and reduce by a factor of J the number of independence test applied to each subset.

A lower bound conjecture. We believe that J^2/ϵ is a lower bound for the query complexity of both the one-sided and the two-sided non-adaptive tests. In light of the two-sided test presented in Section 4, if proven this would be a tight lower bound, up to logarithmic factors, for the two-sided test. As far as we know, it is possible that J^2/ϵ is a lower bound even for the relaxed test proposed in the previous remark.

Reading juntas consistently. There are also interesting questions related to hardness of approximations. In particular, it would be interesting to see what is the best error probability (with regards to ϵ) that can be achieved from tests that query f in a constant number of points that is *independent* of ϵ. It would also be interesting to construct list decoders (also known as consistent readers) for juntas, in the spirit of the consistent readers for low degree polynomials used for constructing Probabilistically Checkable Proofs ([ALM+98, AS98, RS97], see also [DFK+99]). List decoders for long codes with a possible bias, which can be viewed as functions dominated by a junta of one variable, were constructed and applied with good results in [DS02].

Characterizing testable properties. Another open problem goes back to the primal question of characterizing the testable properties. This question is known to be extremely hard even to formulate well, but partial results in the sense of proving the testability of large classes of properties go back to [GGR98]. Now that Fourier transforms are also known to play a part in property testing, the question arises as to whether harmonic analysis can be used in identifying large classes of testable properties of functions.

Random walk convergence. There is also an open problem arising from the proof of the lower bound: For what groups G (other than Z_2^q) can one prove a convergence result similar to Theorem 3? In addition, it would be interesting to improve the lower bound on the convergence rate (remove a factor of q from the bound on t), or to give an example for which the current lower bound is tight.

Testing permutations for non-juntas. Finally, with regards to testing that f is a permutation of a given function h, we can pose the following question: Is there a full characterization of the functions h for which this is easy to test? A simple example of a non-junta function h for which there exists an easy test is the majority function of n boolean variables. On the other hand, it is tempting to conjecture that if n is large enough with respect to J, and h is a J-junta function that is η-far from all $(J-1)$-juntas for some fixed η, then the number of queries that the test requires has to depend on J. The proof of Theorem 2 already implies such a bound for *some* functions h, namely, those that are characters of size J.

A Proof of Proposition 3.5

For $0 \leq x \leq t$, $\quad e^{-x/t} \leq 1 - \frac{x}{et}$. This holds since $e^{-x/t}$ is convex as a function of x, and since the inequality holds at the ends of the segment $[0,t]$. It follows that for all i,

$$\mathbb{E}\left[e^{-X_i/t}\right] \leq \mathbb{E}\left[1 - \frac{X_i}{et}\right] = 1 - \frac{\mathbb{E}[X_i]}{et}$$

Since the expectation is multiplicative for independent variables, we have

$$\mathbb{E}\left[e^{-X/t}\right] = \prod_{i=1}^{l} \mathbb{E}\left[e^{-X_i/t}\right] \leq \prod_{i=1}^{l} \left(1 - \frac{\mathbb{E}[X_i]}{et}\right)$$

We use the convexity of the above expression, together with the fact that $\sum_i \mathbb{E}[X_i] = \alpha$, and obtain

$$\mathbb{E}\left[e^{-X/t}\right] \leq \left(1 - \frac{\alpha}{elt}\right)^l \leq e^{-\alpha/et}$$

The Markov inequality now yields

$$\Pr[X \leq \eta\alpha] = \Pr\left[e^{-X/t} \geq e^{-\eta\alpha/t}\right] \leq \frac{e^{-\alpha/et}}{e^{-\eta\alpha/t}} = e^{\frac{\alpha}{et}(\eta e - 1)}$$

B A new test for being a J-monomial

As a corollary of our testing algorithms for juntas, we present an algorithm that tests whether the function is a J-monomial, namely an AND of J boolean variables and/or negations of variables. This algorithm asks $O(\epsilon^{-1}\text{polylog}(1/\epsilon))$ queries. This is slightly worse than the algorithm in [PRS01], which is linear in $\frac{1}{\epsilon}$. However, the resulting new algorithm is simpler.

Let J, ϵ, and an input function f be given. First, as observed in [PRS01], if $\epsilon \geq 2^{-J+2}$ then the test just needs to approximate $\Pr[\mathsf{f} = 1]$ up to an additive factor of $\frac{\epsilon}{4}$ (because for these parameters every J-monomial is $\frac{1}{2}\epsilon$-close to the zero function), with sufficiently high probability. By the multiplicative Chernoff bound, this costs a number of queries which is linear in $\frac{1}{\epsilon}$.

Assuming that $\epsilon < 2^{-J+2}$, we first test whether f is a J-junta, or is $\epsilon/8$-far from any J-junta. We use sufficiently many queries so that the test succeeds with probability at least $\frac{5}{6}$. If the function passes the junta test, we approximate $\alpha = \Pr[\mathsf{f} = 1]$, asking $O\left(2^J\right)$ queries, so that $\Pr\left[\left|\hat{\alpha} - \alpha\right| \geq 2^{-J-2}\right] \leq \frac{1}{6}$, where $\hat{\alpha}$ is the approximation. We return "ACCEPT" if $\frac{1}{2} \cdot 2^{-J} < \hat{\alpha} < \frac{3}{2} \cdot 2^{-J}$, and "REJECT" otherwise. It is easy to see that this is a J-monomial test with success probability at least $\frac{2}{3}$.

At this stage, it is also possible to check for the number of coordinates that appear with a negation sign in the monomial. This is done by approximating the probability that $\mathsf{f}(x) = 0$ where each coordinate of x is independently chosen to be 0 with probability $\frac{1}{3}$, and 1 with probability $\frac{2}{3}$. This stage of the test is not linear in ϵ^{-1}, but polynomial in it (assuming that $\epsilon < 2^{-J+2}$, as otherwise this question has little meaning due to the observations above).

Acknowledgments

We wish to thank Michal Parnas for the discussions concerning some of the questions that led to the writing of this paper, and wish to thank Avi Wigderson for his comments. We also wish to thank two anonymous referees for their invaluable comments, which led among other things to a noticeable simplification of the proof of Theorem 1.

References

[AD86] D. Aldous and P. Diaconis. Shuffling cards and stopping times. *American Mathematical Monthly*, 93(5):333–348, 1986.

[AS00] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley-Interscience (John Wiley & Sons), New York, 1992 (1st ed), 2000 (2nd ed).

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, May 1998.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: a new characterization of NP. *Journal of the ACM*, 45(1):70–122, January 1998.

[BHL95] Avrim Blum, Lisa Hellerstein, and Nick Littlestone. Learning in the presence of finitely or infinitely many irrelevant attributes. *Journal of Computer and System Sciences*, 50(1):32–40, February 1995.

[BL89] M. Ben-Or and N. Linial. Collective coin flipping. *ADVCR: Advances in Computing Research*, 5, 1989.

[CG02] H. Chockler and D. Gutfreund. Property testing: Worst case vs. average case. manuscript, 2002.

[DFK+99] I. Dinur, E. Fischer, G. Kindler, R. Raz, and S. Safra. PCP characterizations of NP: Towards a polynomially-small error-probability. In *Proc. 31th ACM Symp. on Theory of Computing*, 1999.

[DS02] I. Dinur and S. Safra, On the importance of being biased, In *Proc. 34th ACM Symp. on Theory of Computing*, 2002.

[DGL+99] Yevgeniy Dodis, Oded Goldreich, Eric Lehman, Sofya Raskhodnikova, Dana Ron, and Alex Samorodnitsky. Improved testing algorithms for monotonicity. *3rd International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM)*, August 1999.

[Fis01] E. Fischer. The art of uninformed decisions: A primer to property testing. *The Bulletin of the European Association for Theoretical Computer Science*, 75:97–126, 2001.

[FLN+02] E. Fischer, E. Lehman, I. Newman, S. Raskhodnikova, R. Rubinfeld, and A. Samorodnitsky. Monotonicity testing over general poset domains. In *Proc. 34th ACM Symp. on Theory of Computing*, pages 474–483, 2002.

[GGR98] S. Goldwasser O. Goldreich and D. Ron. Property testing and its connections to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.

[GGL+00] Oded Goldreich, Shafi Goldwasser, Eric Lehman, Dana Ron, and Alex Samorodnitsky. Testing monotonicity. *Combinatorica*, 20(3):301–337, 2000.

[GTT99] David Guijarro, Jun Tarui and Tatsuie Tsukiji. Finding relevant variables in PAC model with membership queries. In *Proc. Algorithmic Learning Theory, 10th International Conference, ALT '99, Tokyo, Japan, December 1999*, volume 1720 of *Lecture Notes in Artificial Intelligence*, pages 313–322. Springer, 1999.

[Lit87] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285, 1987.

[Mar06] A. A. Markov. Extension of the law of large numbers to dependent events. *Bull. Soc. Phys. Math.*, 15(2):135–156, 1906.

[MOS02] E. Mossel, R. O'Donnell, and R. A. Servedio. Learning juntas. *Proceedings of the 35th Annual symposium on the theory of computing (STOC)*, pages 206–212, 2003.

[PRS01] M. Parnas, D. Ron, and A. Samorodnitsky. Testing Basic Boolean Formulae. *SIAM Journal on Discrete Math*, 16(1):20–46, 2002.

[Ron01] D. Ron. Property testing (a tutorial), In *Handbook on Randomization*, Vol. II, pages 597–649, 2001.

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

[RS97] R. Raz and S. Safra. A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In *Proc. 29th ACM Symp. on Theory of Computing*, pages 475–484, 1997.

[UTW97] R. Uehara, K. Tsuchida, and I. Wegener. Optimal attribute-efficient learning of disjunction, parity and threshold functions. In Shai Ben-David, editor, *Proceedings of the 3rd European Conference on Computational Learning Theory, Berlin, March 17–19 1997*, volume 1208 of *LNAI*, pages 171–184. Springer.