Lower bounds for testing function isomorphism

Eric Blais   Ryan O'Donnell[*]

Computer Science Department
Carnegie Mellon University
{eblais,odonnell}@cs.cmu.edu

December 15, 2009

Abstract

We prove new lower bounds in the area of property testing of boolean functions. Specifically, we study the problem of testing whether a boolean function f is isomorphic to a fixed function g (i.e., is equal to g up to permutation of the input variables). The analogous problem for testing graphs was solved by Fischer in 2005. The setting of boolean functions, however, appears to be more difficult, and no progress has been made since the initial study of the problem by Fischer et al. in 2004.

Our first result shows that any non-adaptive algorithm for testing isomorphism to a function that "strongly" depends on k variables requires $\log k - O(1)$ queries (assuming k/n is bounded away from 1). This lower bound affirms and strengthens a conjecture appearing in the 2004 work of Fischer et al. Its proof relies on total variation bounds between hypergeometric distributions which may be of independent interest.

Our second result concerns the simplest interesting case not covered by our first result: non-adaptively testing isomorphism to the Majority function on k variables. Here we show that $\Omega(k^{1/12})$ queries are necessary (again assuming k/n is bounded away from 1); this exponentially improves on a related lower bound of Matulef et al. from 2009. The proof of this result relies on recently developed multidimensional "invariance principle" tools.

[*] Supported by NSF grants CCF-0747250 and CCF-0915893, BSF grant 2008477, and Sloan and Okawa fellowships.

1 Introduction

This paper is concerned with the field of property testing for boolean functions. Let us recall the standard framework, as originally introduced by Rubinfeld and Sudan [RS96].

Definition 1.1. Let $\mathcal{P}$ be a class of functions $\{0,1\}^n \to \{0,1\}$. We say that a randomized query algorithm $\mathcal{T}$ with black-box access to an unknown function $f : \{0,1\}^n \to \{0,1\}$ is an *(ϵ, q)-tester for* $\mathcal{P}$ if it makes at most q queries to $f : \{0,1\}^n \to \{0,1\}$ and then:

- Accepts with probability at least $2/3$ when f is in $\mathcal{P}$; and,
- Rejects with probability at least $2/3$ when f is ϵ-far from every function $f' \in \mathcal{P}$.

Here we say that f and f' are *ϵ-far* if they differ on at least an ϵ fraction of the inputs in $\{0,1\}^n$, and are ϵ-close otherwise. When the algorithm chooses all of its queries in advance it is *non-adaptive*; otherwise we say it is *adaptive*.

Definition 1.2. For a fixed $\epsilon > 0$ and choice of adaptivity, the *query complexity* of $\mathcal{P}$ is the minimum value of q for which there is an (ϵ, q)-tester. Following standard conventions, when the query complexity q is independent of n for every $\epsilon > 0$, we say that $\mathcal{P}$ is *easy to test*; otherwise we say that it is *hard to test*.[1] This notion is independent of the choice of adaptivity, since non-adaptive query complexity can be (exponentially) bounded in terms of adaptive query complexity.

The last five years have seen great strides in understanding the testability of *graph properties*. This is the special case in which $f : \{0,1\}^{\binom{V}{2}} \to \{0,1\}$ encodes the adjacency matrix of a graph on vertex set V, and $\mathcal{P}$ is a property that is closed under graph symmetries; i.e., permutations of V. Indeed, the works [AS05a, AS05b, AFNS06, AT08] have to a large extent characterized the testability of graph properties.

However the characterization problem for general boolean functions is very far from understood and remains a longstanding open problem. In this paper we revisit a major subproblem, introduced in the early work of Fischer, Kindler, Ron, Safra, and Samorodnitsky [FKR$^+$04]: the difficulty of testing isomorphism to a function given in advance.

Testing g-isomorphism. Two boolean functions $f : \{0,1\}^n \to \{0,1\}$ and $g : \{0,1\}^n \to \{0,1\}$ are said to be *isomorphic* to each other if they are identical up to reordering input variables. More precisely, we say that f and g are isomorphic to each other if there is a permutation σ on $[n]$ such that for every $x = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$, $f(x_1, x_2, \ldots, x_n) = g(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)})$.

For each function $g : \{0,1\}^n \to \{0,1\}$, we let $\mathcal{P}_g$ denote the class of all functions isomorphic to g. This gives a natural testing problem: testing whether an unknown function f is isomorphic to the known function g. This is called the problem of *testing g-isomorphism*. Fischer et al. [FKR$^+$04] proposed the following question, a significant component of the research program to characterize testability of boolean functions:

Research goal: Classify all boolean functions g according to whether testing g-isomorphism is easy or hard.

We remark that Fischer [Fis05] solved the analogous problem for *graph properties* soon after; see also [FM08]. But he called the general case of boolean functions "rather hard", and indeed the authors are not aware of nay additional progress specifically on this problem.

In this paper we make progress towards a characterization by showing hardness of testing g-isomorphism for a large class of functions g.

Prior work. When the problem of testing function isomorphism was first raised by Fischer et al. [FKR$^+$04], a few simple cases were already well-understood. First, it is easy to see that for any totally symmetric function $g : \{0,1\}^n \to \{0,1\}$, testing g-isomorphism is easy – no other functions are isomorphic to g, and testing function *identity* requires only $O(1/\epsilon)$ queries.

Another instance of the g-isomorphism testing problem that is well understood is one where $g(x) = x_i$ for some $i \in \{1, \ldots, n\}$. Then the g-isomorphism problem is equivalent to the well-studied *dictatorship* testing problem at the heart of PCP constructions (first studied in [BGS98]). The query complexity of the dictatorship testing problem is $O(1/\epsilon)$, so this special case of the function isomorphism problem is also easy. Parnas, Ron,

[1] Formally speaking, this makes sense only for a family of properties $(\mathcal{P}_n)_n$, one for each input length.

and Samorodnitsky [PRS02] also showed that testing g-isomorphism is easy when g is an AND function on any number of variables.

The paper of Fischer et al. [FKR$^+$04] introduced a strong new upper bound: they showed that for any k, if g is a *k-junta* – meaning that g depends on at most k variables – then it is possible to ϵ-test g-isomorphism (non-adaptively, even) with $\mathrm{poly}(k/\epsilon)$ queries. Therefore, testing isomorphism to any $O(1)$-junta is easy.

Regarding hardness results, prior to this work the only known lower bound for testing g-isomorphism was for the case of g being the Parity_k function, where $\mathrm{Parity}_k : \{0,1\}^n \to \{0,1\}$ is defined by $\mathrm{Parity}_k(x) = x_1 \oplus x_2 \oplus \cdots \oplus x_k$. Fischer et al. [FKR$^+$04] showed that when $k \le o(\sqrt{n})$, ϵ-testing Parity_k-isomorphism non-adaptively requires $\tilde{\Omega}(\sqrt{k}/\epsilon)$ queries. This result implies that testing Parity_k-isomorphism is hard for $\omega(1) \le k \le o(\sqrt{n})$.

Our results. It seems clear that more work needs to be done on the hardness side of testing g-isomorphism. A first direction would be to investigate the following conjecture, stated in [FKR$^+$04]: "If n is sufficiently large compared to k, and g is a k-junta which is ϵ-far from all $(k-1)$-juntas, then ϵ-testing g-isomorphism requires $\omega_k(1)$ queries."

Our first main result affirms and significantly strengthens this conjecture:

Theorem 1.3. *Let $g : \{0,1\}^n \to \{0,1\}$ be a k-junta which is ϵ-far from being a $(k-e)$-junta for some $e \ge 1$. Then any non-adaptive ϵ-tester for g-isomorphism must make at least $\log_2(k'/e^2) - O(1)$ queries, where $k' = \min(k, n-k)$.*

Qualitatively, Theorem 1.3 implies that testing g-isomorphism for k-juntas is hard whenever $\omega(1) \le k \le n - \omega(1)$, provided the k-junta g is far from all juntas on $k - o(\sqrt{k})$ variables. We discuss the possibility of improving this theorem in Section 5.

As one very special case, Theorem 1.3 extends the Fischer et al. hardness result for testing Parity_k-isomorphism to all $\omega(1) \le k \le n - \omega(1)$ (albeit with a worse query complexity lower bound). It is important to note that the restriction $k \le n - \omega(1)$ is not an artifact of our proof, but rather is inherent. This is because, e.g., testing Parity_k-isomorphism when $k = n - O(1)$ is *easy*: if the tester XORs each query response with the parity of all bits in the query, it reduces testing Parity_k-isomorphism to testing Parity_{n-k}-isomorphism. Thus testing $\mathrm{Parity}_{n-O(1)}$-isomorphism is easy, by the junta isomorphism-testing result of Fischer et al.

Our Theorem 1.3 is only useful for k-juntas g that are far from being $(k - o(\sqrt{k}))$-juntas. Perhaps the most natural case not covered by this theorem is that of the majority function on k variables, Maj_k. In Proposition 4.1 we do a straightforward calculation to show that for all $\delta > 0$, the function Maj_k *is* $o_\delta(1)$-close to being a junta on $k - \delta k$ variables: namely $\mathrm{Maj}_{k-\delta k}$. Nevertheless, the second main result in our paper proves a strong hardness result for testing Maj_k-isomorphism:

Theorem 1.4. *For every constant $\delta > 0$, there exists a constant $\epsilon > 0$ such that the following holds: Assuming $1/\epsilon \le k \le (1-\delta)n$, any non-adaptive algorithm for ϵ-testing Maj_k-isomorphism must make at least $\Omega((\delta k)^{1/12})$ queries.*

Qualitatively, Theorem 1.4 implies that testing Maj_k-isomorphism is hard for every $\omega(1) \le k \le n - \Omega(n)$. Again, this range is optimal: the upper bound cannot be improved because as mentioned, $\mathrm{Maj}_{n-o(n)}$ is $o(1)$-close to Maj_n; thus we can test $\mathrm{Maj}_{n-o(n)}$-isomorphism using the easy Maj_n-isomorphism tester that follows from Maj_n being totally symmetric.

In this extended abstract, we prove Theorem 1.4 only in the case where $\delta = 1/4$ (and assuming k is divisible by 3). The few tedious technical modifications needed to handle smaller values of δ are deferred to the full version of the article. I.e., we prove:

Theorem 1.5. *There is a universal $\epsilon_0 > 0$ such that whenever $k \le (3/4)n$ (and is divisible by 3), any non-adaptive algorithm for ϵ_0-testing Maj_k-isomorphism must make at least $\Omega(k^{1/12})$ queries.*

We remark that the bound in Theorem 1.5 is exponentially stronger than the one in Theorem 1.3: it shows that any non-adaptive algorithm for testing Maj_k-isomorphism must make a number of queries *polynomial* in k. This bound is optimal (up to the right value of the exponent), since Fischer et al.[FKR$^+$04]'s upper bound on the query complexity of testing isomorphism to k-juntas implies that testing isomorphism to the Maj_k function can be done with $\mathrm{poly}(k/\epsilon)$ queries.

The result of Theorem 1.4 is also interesting in light of the recent results of Matulef et al. on testing halfspaces: they showed that testing the class of halfspaces is easy [MORS09a], but testing a natural subclass of halfspaces – the class of ±1-weight halfspaces – is hard [MORS09b]. More precisely, they showed a non-adaptive query lower bound of $\Omega(\log n)$ for this class. Our Theorem 1.4 gives an exponentially improved lower bound for a similar subclass of halfspaces: $\Omega(n^{1/12})$ queries for the class of majority functions on, say, $n/2$ variables.

In light of Fischer [Fis05]'s solution to the isomorphism testing problem for graph properties – i.e., boolean functions with a certain high degree of symmetry – we believe that characterizing testability of g-isomorphism for *symmetric* k-juntas is an approachable first step. Theorem 1.5 represents progress in this direction. We remark that our method of proving Theorem 1.5 can be extended to handle certain other symmetric k-juntas. However the general case of symmetric k-juntas g has some unexpected tricky aspects to it, which we discuss in Section 5.

Our techniques. The proofs of Theorems 1.3 and 1.5 both use the standard approach for proving lower bounds in property testing: Yao's Minimax Principle [Yao77]. That is, we prove both theorems by introducing distributions $\mathcal{F}_{\text{yes}}$ and $\mathcal{F}_{\text{no}}$ on functions that should be accepted and rejected, respectively, by algorithms testing g-isomorphism, and show a lower bound on the number of queries required by any *deterministic* testing algorithm. The main technical contribution of this research is in the design and the analysis of the distributions $\mathcal{F}_{\text{yes}}$ and $\mathcal{F}_{\text{no}}$.

The main challenge in proving Theorem 1.3 is that the lower bound applies to a very general class of functions g. To prove the theorem we need to design distributions that work *without using any structural properties of the function g being tested.* The key to doing this involves analyzing the statistical (total variation) distance between two *multivariate hypergeometric distributions.* What follows is the main lemma we need; it may be of independent interest:

Lemma 1.6. *Suppose $X \sim \text{Hyp}(n, r, k)$ and $Y \sim \text{Hyp}(n, r, \ell+e)$, where $\text{Hyp}(n, r, \ell)$ denotes the (univariate) hypergeometric distribution: i.e., the number of red balls drawn when selecting ℓ balls randomly without replacement from an urn containing n balls, r of which are red. Then*

$$d_{\text{TV}}(X, Y) \leq .01$$

provided $(1 - \frac{r}{n})\min(\ell, n-\ell) \geq Ce^2$, where C is a universal constant. Here $d_{\text{TV}}(\cdot, \cdot)$ denotes total variation distance.

In Section B.2 we comment on why the somewhat complicated hypothesis on r, n, ℓ, and e is necessary. The proof of Theorem 1.3, as well as a more complete discussion of the techniques it requires, is presented in Section 3.

The challenge in proving Theorem 1.5 is fundamentally different. For Theorem 1.3, we know that the k-junta g is far from all $(k-1)$-juntas, say, which means it is okay for our $\mathcal{F}_{\text{no}}$ functions to be "small tweaks" to g. However when $g = \text{Maj}_k$, small tweaks result in functions that are still close to Maj_k. Thus our $\mathcal{F}_{\text{no}}$ functions must be somewhat drastic changed from Maj_k, yet still "look like" Maj_k. We arrange for this by making the $\mathcal{F}_{\text{no}}$ functions very carefully constructed *weighted* majority functions on $(4/3)k$ coordinates. To show that such functions still "look like" Maj_k, we use recently developed multidimensional invariance principles [Mos08, GOWZ09]. However these need to be adapted to the case of sums of random vectors which are *not independent*, but rather are drawn without replacement from a fixed pool of random vectors.

2 Preliminaries and definitions

Given two random variables X, Y defined on a common discrete sample space Ω, let $d_{\text{TV}}(X, Y)$ denote the *total variation* distance between X and Y, where $d_{\text{TV}}(X, Y) = \frac{1}{2}\sum_{\omega \in \Omega} |\mathbf{Pr}[X = \omega] - \mathbf{Pr}[Y = \omega]|$.

Let $\text{Hyp}(n, m, k)$ denote the hypergeometric distribution – the distribution on the number t of red balls drawn when k balls are drawn without replacement from a set of n balls, m of which are red. Our results rely on the following anti-concentration property of hypergeometric distributions.

Lemma 2.1. *Let $X \sim \text{Hyp}(n, m, k)$ and let $\sigma^2 = \frac{km}{n}\left(1 - \frac{m}{n}\right)\left(1 - \frac{k}{n}\right)$. Then for every $t \geq 0$,*

$$\mathbf{Pr}[X = t] \leq \frac{C}{\sigma}$$

where C is an absolute constant.

We include a proof of Lemma 2.1 in Appendix A. Our results also use a special case of the Berry-Esseen Central Limit Theorem, which we recall below.

Berry-Esseen Theorem (Special case)**.** *Let $S = \sum_{i \in [n]} \alpha_i X_i$ where $X_1, \dots, X_n$ are i.i.d. random variables drawn uniformly at random from $\{-1, 1\}$, and where $\sum_{i \in [n]} \alpha_i^2 = 1$ and $\max_{i \in [n]} |\alpha_i| = \gamma < \infty$. Then for any $t \in \mathbb{R}$,*

$$|\mathbf{Pr}[S \le t] - \Phi(t)| \le C \cdot \gamma,$$

where C is a universal constant (one can take $C = 0.7056$ by a recent result of Shevtsova [She07]).

Influence. The *influence* of the ith variable in the function $f : \{0,1\}^n \to \{0,1\}$ is $\mathrm{Inf}_f(i) = \mathbf{Pr}_x[f(x) \neq f(x^{(i)})]$, where the probability is taken over the uniform distribution of $x \in \{0,1\}^n$ and $x^{(i)}$ is the input formed by flipping the value of the ith bit in x. The influence of the set $S \subseteq [n]$ of variables in f is

$$\mathrm{Inf}_f(S) = 2 \mathop{\mathbf{Pr}}_{x,y}[f(x) \neq f(y)],$$

where x is generated uniformly at random and y is generated by taking a copy of x and re-randomizing the value of the variables in S.

When $\mathrm{Inf}_f(i) > 0$, we say that the ith variable is *relevant* in f. A function that contains at most k relevant variables is called a *k-junta*. There is a close relation between the distance of a function to being a junta and the influence of sets of variables:

Proposition 2.2. *Let $f : \{0,1\}^n \to \{0,1\}$ be a k-junta, where $J \subseteq [n]$ is the set of relevant variables. Define $\tau = \min_{I \subseteq J, |I| = e} \mathrm{Inf}_I(f)$. Then f is 2τ-far and 4τ-close to being a $(k-e)$-junta.*

The proof of Proposition 2.2 is presented in Appendix A.

3 Our general isomorphism-testing lower bound

In this section, we prove the following theorem:

Theorem 1.3 (Restated)**.** *Let $g : \{0,1\}^n \to \{0,1\}$ be a k-junta which is ϵ-far from being a $(k-e)$-junta for some $e \ge 1$. Then any non-adaptive ϵ-tester for g-isomorphism must make at least $\log_2(k'/e^2) - O(1)$ queries, where $k' = \min(k, n-k)$.*

On first reading, the reader is encouraged to focus on the simplest case, where $e = 1$. In this case, Theorem 1.3 affirms a conjecture stated in [FKR+04], and we have the following easy-to-apply corollary:

Corollary 3.1. *For $\omega(1) \le k \le n - \omega(1)$, if $g : \{0,1\}^n \to \{0,1\}$ is a k-junta in which every relevant variable has influence at least ϵ, then g-isomorphism is not ϵ-testable.*

Using the theorem with general e, we see that the corollary holds even when we only assume that every collection of $o(\sqrt{\min(k, n-k)})$ relevant variables has influence at least ϵ.

We now begin the proof of Theorem 1.3. Let g, n, k, k', e, and ϵ be as in the statement of the theorem. Without loss of generality, we may assume that the k relevant coordinates for g are $[k] = \{1, 2, \dots, k\}$. We write $g_{\mathrm{core}} : \{0,1\}^k \to \{0,1\}$ for the restriction of g to these coordinates.

As is standard in property testing lower bounds, the proof of Theorem 1.3 uses Yao's Minimax Principle [Yao77]. Specifically, we construct two probability distributions $\mathcal{F}_{\mathrm{yes}}$ and $\mathcal{F}_{\mathrm{no}}$ over functions isomorphic to g and functions ϵ-far from being isomorphic to g, respectively. We then show that any deterministic non-adaptive algorithm making $\ll \log_2(k'/e^2)$ queries cannot distinguish with probability at least $1/3$ between functions drawn from $\mathcal{F}_{\mathrm{yes}}$ or from $\mathcal{F}_{\mathrm{no}}$.

The distributions $\mathcal{F}_{\mathrm{yes}}$ and $\mathcal{F}_{\mathrm{no}}$. We define $\mathcal{F}_{\mathrm{yes}}$ in the most natural way, by randomly embedding g_{core} into $[n]$. More precisely, to obtain a draw $f_{\mathrm{yes}} \sim \mathcal{F}_{\mathrm{yes}}$, we first choose a uniformly random subset $J \subseteq [n]$ of cardinality k. Next, we choose a uniformly random bijection $\sigma : [k] \to J$. Finally, we define $f_{\mathrm{yes}}(x) = g_{\mathrm{core}}(x_{\sigma(1)}, \dots, x_{\sigma(k)})$. It is clear that every such f_{yes} is isomorphic to g.

As for $\mathcal{F}_{\mathrm{no}}$, we define a draw $f_{\mathrm{no}} \sim \mathcal{F}_{\mathrm{no}}$ as follows: First, we choose a uniformly random subset $J \subseteq [n]$ of cardinality $k - e$. Next, we choose a uniformly random map $\sigma : [k] \to J$ from among those satisfying the following property: there is one $j_1 \in J$ with $e+1$ preimages under σ, and the remaining $j \in J \setminus \{j_1\}$ have a unique preimage. Finally, we define $f_{\mathrm{no}}(x) = g_{\mathrm{core}}(x_{\sigma(1)}, \dots, x_{\sigma(k)})$. Each such f_{no} only depends on the coordinates J and hence is a $(k-e)$-junta. Thus by the assumption in Theorem 1.3, each such f_{no} is indeed ϵ-far from being isomorphic to g.

To prove Theorem 1.3, it suffices to prove the following:

Theorem 3.2. *Let $\mathcal{T}$ be any* deterministic *non-adaptive q-query testing algorithm for functions* $\{0,1\}^n \to \{0,1\}$. *Then*

$$\left|\underset{f_{\text{yes}}\sim\mathcal{F}_{\text{yes}}}{\mathbf{Pr}}[\mathcal{T} \textit{ accepts } f_{\text{yes}}] - \underset{f_{\text{no}}\sim\mathcal{F}_{\text{no}}}{\mathbf{Pr}}[\mathcal{T} \textit{ accepts } f_{\text{no}}]\right| \le 2^q \cdot \frac{O(e^2)}{k'} + .01.$$

Note that if $q < \log_2(k'/e^2) - c_0$ for a sufficiently large constant c_0, then the upper bound in this theorem is at most 1/3. From this we deduce Theorem 1.3 immediately using Yao's Minimax Principle.

3.1 Bounding the distance between two multivariate hypergeometrics

The typical way to prove a property testing bound such as Theorem 3.2 is as follows. First, we write the q queries of tester $\mathcal{T}$ as $x^1, \dots, x^q \in \{0,1\}^n$. We then introduce the *Response Vector* random variables R_{yes} and R_{no}. Here $R_{\text{yes}} \in \{0,1\}^q$ is defined by drawing $f_{\text{yes}} \sim \mathcal{F}_{\text{yes}}$ and letting $R_{\text{yes}} = \langle f_{\text{yes}}(x^1), \dots, f_{\text{yes}}(x^q)\rangle$, and R_{no} is defined analogously. Finally, we show that

$$d_{\text{TV}}(R_{\text{yes}}, R_{\text{no}}) \le 2^q \cdot \frac{O(e^2)}{k'} + .01, \tag{1}$$

where $d_{\text{TV}}(\cdot,\cdot)$ denotes the total variation distance between two discrete random variables.

We will in fact prove a stronger statement. To understand it, let's reconsider the complete random processes $\mathcal{P}_{\text{yes}}$ and $\mathcal{P}_{\text{no}}$ by which the Response Vectors R_{yes} and R_{no} are generated. We begin by focusing on the "yes" process, $\mathcal{P}_{\text{yes}}$.

Given the tester $\mathcal{T}$'s queries $x^1, \dots, x^q \in \{0,1\}^n$, we think of them as row vectors and arrange them into a $q \times n$ *Query Matrix* Q. We will be especially interested in the *Columns* of this matrix Q, the jth column consisting of the jth bits of all the query strings. Abstractly, we define the set of all possible Column (types)

$$\mathfrak{C} = \{0,1\}^q$$

Since $|\mathfrak{C}| = 2^q \ll n$, at least some Columns will occur many times in the matrix Q. In fact, we will think of the Query Matrix Q as being an ordered *multiset* of Columns from $\mathfrak{C}$.

Recalling the definition of $\mathcal{F}_{\text{yes}}$, we think of the first step of $\mathcal{P}_{\text{yes}}$ as choosing k column indices $j_1, \dots, j_k$ randomly and without replacement from $[n]$. We next extract Columns $j_1, \dots, j_k$ from Q. We view this as a *multiset* of Columns, and call it the *Argument Multiset* S_{yes}. Next, we *randomly order* the Columns in S_{yes}, forming a $q \times k$ *Argument Matrix* A_{yes}. Finally, we produce the Response Vector R_{yes} by applying g_{core} to the Argument Matrix, row-wise. (A diagram of $\mathcal{P}_{\text{yes}}$ is included in Figure 1 of the Appendix.)

The reader can easily verify this process $\mathcal{P}_{\text{yes}}$ generates the correct distribution on the Response Vector random variable R_{yes}.

The "no" process $\mathcal{P}_{\text{no}}$ is very similar, differing only in the way it generates the Argument Multiset from the Query Matrix. Recalling the definition of $\mathcal{F}_{\text{no}}$, we think of $\mathcal{P}_{\text{no}}$ as forming the Argument Multiset S_{no} by choosing $\ell = k - e$ random Columns from Q without replacement, and including an *additional* e copies of the first-chosen Column. The process $\mathcal{P}_{\text{no}}$ then forms the Argument Matrix A_{no} by again randomly ordering the Columns in the Argument Multiset, and finally produces the Response Vector R_{no} again by applying g_{core} to A_{no}, row-wise. The reader can again easily verify that $\mathcal{P}_{\text{no}}$ generates the correct distribution on R_{no}.

Because the processes are identical after the Argument Multiset is formed, a coupling argument immediately implies that

$$d_{\text{TV}}(R_{\text{yes}}, R_{\text{no}}) \le d_{\text{TV}}(S_{\text{yes}}, S_{\text{no}}). \tag{2}$$

This inequality can be extremely lossy, depending on the function g_{core}. However, since Theorem 1.3 applies for an extremely broad range of functions, we are almost forced to design a proof of Theorem 3.2 that *uses no properties of the function g_{core}*. That is, in the absence of additional restrictions on the class of functions considered, there is no obvious way to bound $d_{\text{TV}}(R_{\text{yes}}, R_{\text{no}})$ except by $d_{\text{TV}}(S_{\text{yes}}, S_{\text{no}})$.

Letting $\mathcal{S}_{\text{yes}}$ denote the subprocess of $\mathcal{P}_{\text{yes}}$ generating S_{yes}, and similarly for $\mathcal{S}_{\text{no}}$, we have reduced proving (1), and hence Theorem 1.3, to the following:

Theorem 3.3. *For $S_{\text{yes}} \sim \mathcal{S}_{\text{yes}}, S_{\text{no}} \sim \mathcal{S}_{\text{no}}$, we have* $d_{\text{TV}}(S_{\text{yes}}, S_{\text{no}}) \le |\mathfrak{C}| \cdot \dfrac{O(e^2)}{\min(k, n-k)} + .01.$

The reader can see now why our query complexity lower bound in Theorem 1.3 is only logarithmic; we have $|\mathfrak{C}| = 2^q$ competing against $\frac{1}{k}$ in the above bound. Indeed, we can never prove a better-than-logarithmic lower bound if our proof only involves showing statistical closeness of the Argument Multisets S_{yes} and S_{no}. To see this, suppose $k = n/2$, so $n - k = n/2$ as well. Then if $2^q \gg n/2$, it is possible that every Column in the Query Matrix is unique. In this case, the total variation distance between Argument Multisets S_{yes} and S_{no} will be 1 even in the case $e = 1$, because S_{yes} will always consist of unique Columns, whereas S_{no} will always have one Column duplicated.

Notice that the ordering of the Columns in the Query Matrix Q has proven to be unimportant; we can think of Q simply as an unordered multiset of Columns from $\mathfrak{C}$. Thus Theorem 3.3 is really a statement about the total variation distance between certain multivariate hypergeometric random variables. Specifically, for each Column $\mathfrak{c} \in \mathfrak{C}$, let $m(\mathfrak{c})$ denote the number of copies of $\mathfrak{c}$ in Q. In process $\mathcal{S}_{\text{yes}}$, we choose k random Columns from Q without replacement and count the number of copies of each Column (type) in the draw. Process $\mathcal{S}_{\text{no}}$ is similar, except we choose ℓ random Columns from Q without replacement, and count an extra e copies of the first-drawn Column.

3.2 Reduction of Theorem 3.3 to two lemmas

This preceding discussion motivates the following notation:

Definition 3.4. Given integers $N, e \geq 1$, $M, L \geq 0$, with $M, L + e \leq N$, we define $\lambda(N, M, L, e) = d_{\text{TV}}(X, Y)$, where $X \sim \text{Hyp}(N, M, L + e)$ and $Y \sim \text{Hyp}(N, M, L) + e$.

The proof of Theorem 3.3 relies on the following two lemmas. The first lemma is relatively straightforward, and relates the distance between $\mathcal{S}_{\text{yes}}$ and $\mathcal{S}_{\text{no}}$ to the total variation distance between hypergeometric distributions.

Lemma 3.5. $d_{\text{TV}}(S_{\text{yes}}, S_{\text{no}}) \leq \displaystyle\sum_{\mathfrak{c} \in \mathfrak{C} : m(\mathfrak{c}) \neq 0} \frac{m(\mathfrak{c})}{n} \cdot \lambda(n-1, m(\mathfrak{c}) - 1, \ell - 1, e).$

The second lemma is a total variation distance bound between (univariate) hypergeometric random variables which may be of independent interest.

Lemma 3.6. *Write $L' = \min(L, N - L)$ and assume $\frac{ML'}{N} \geq \kappa e^2$, where $2 \leq \kappa < \infty$ is a certain universal constant. Then $\lambda(N, M, L, e) \leq .01$.*

This lemma is in fact identical to the key Lemma 1.6: to see this, one only needs to replace M with $r = N - M$ and use the obvious fact that $\text{Hyp}(N, N - M, L)$ is the same distribution as $L - \text{Hyp}(N, M, L)$.

We briefly comment on why a hypothesis like $\frac{ML'}{N} \gg e^2$ is necessary to show that $\text{Hyp}(N, M, L + e)$ and $\text{Hyp}(N, M, L) + e$ are close in total variation distance. For simplicity, first suppose that $e = 1$. It is necessary that that $\frac{ML}{N} \gg 1$; this quantity is the mean of $\text{Hyp}(N, M, L)$, and if it were $\ll 1$ and then $X \sim \text{Hyp}(N, M, L + 1)$ is likely to be 0 whereas $Y \sim \text{Hyp}(N, M, L) + 1$ is at least 1. Second, it is also necessary that $\frac{M(N-L)}{N} = M(1 - \frac{L}{N}) \gg 1$. To see this, note that if by way of contrast $1 - \frac{L}{N} \ll \frac{1}{M}$, then X will be concentrated at M and Y be concentrated at $M + 1$. Finally, to understand the hypothesis's dependence on e, suppose $M = N/2$ and L is quite small. Then $\text{Hyp}(N, M, L)$ is distributed very much like $\text{Binomial}(L, e)$; hence we require $L \gg e^2$ or else the extra $+e$ in Y will dominate the standard deviation of $\text{Binomial}(L, e)$.

Due to space constraints, we defer the proofs of Lemmas 3.5 and 3.6 to Appendix B. We now show that Theorem 3.3 follows from the lemmas:

Proof of Theorem 3.3. Note that we may freely assume $k \geq 2e + 2$, as otherwise the bound we are trying to prove exceeds 1 (assuming the constant in the $O(\cdot)$ is large enough). Let us introduce the notation $N = n - 1$, $M(\mathfrak{c}) = m(\mathfrak{c}) - 1$, $L = \ell - 1$, $L' = \min(L, N - L)$. Then by Lemma 3.5,

$$
\begin{aligned}
d_{\text{TV}}(\mathcal{S}_{\text{yes}}, \mathcal{S}_{\text{no}}) &\leq \sum_{\mathfrak{c} \in \mathfrak{C} : m(\mathfrak{c}) \neq 0} \frac{m(\mathfrak{c})}{n} \cdot \lambda(n-1, m(\mathfrak{c}) - 1, \ell - 1, e) \\
&= \sum_{0 \leq \frac{M(\mathfrak{c})}{N} L' < \kappa e^2} \frac{m(\mathfrak{c})}{n} \cdot \lambda(N, M(\mathfrak{c}), L, e) + \sum_{\frac{M(\mathfrak{c})}{N} L' \geq \kappa e^2} \frac{m(\mathfrak{c})}{n} \cdot \lambda(N, M(\mathfrak{c}), L, e) \\
&\leq \sum_{0 \leq \frac{M(\mathfrak{c})}{N} L' < \kappa e^2} \frac{m(\mathfrak{c})}{n} + \sum_{\frac{M(\mathfrak{c})}{N} L' \geq \kappa e^2} \frac{m(\mathfrak{c})}{n} \cdot .01,
\end{aligned}
$$

where the last inequality uses Lemma 3.6. Since $\sum_{\mathfrak{c}\in\mathfrak{C}} m(\mathfrak{c}) = n$, the second sum above is at most .01. Thus it remains to bound the first sum by $|\mathfrak{C}|\frac{O(e^2)}{\min(k,n-k)}$. There are at most $|\mathfrak{C}|$ summands in this first sum, and for each we have

$$\frac{m(\mathfrak{c})}{n} = \frac{M(\mathfrak{c})+1}{N+1} \leq \frac{M(\mathfrak{c})}{N} + 1 \leq \frac{\kappa e^2}{L'}$$

by the condition of the sum. It thus remains to show $L' \geq \Omega(\min(k, n-k))$. But

$$\begin{aligned} L' = \min(\ell-1, n-\ell) &= \min(k-e-1, n-k+e) \\ &\geq \min(k-e-1, n-k) \geq \min(k/2, n-k) = \Omega(\min(k, n-k)), \end{aligned}$$

where we used the inequality $k \geq 2e+2$, completing the proof. □

4 Majority functions

Recall that $\mathrm{Maj}_k : \{-1,1\}^n \to \{-1,1\}$ is defined by $\mathrm{Maj}_k(x) = \mathrm{sgn}(\sum_{i=1}^k x_i)$, where we define sgn(0) (arbitrarily) to be 1. We sometimes abuse notation by thinking of Maj_k also as a function on $\{-1,1\}^k$. In this section we prove our lower bound for testing Maj_k-isomorphism, Theorem 1.5, which we restate here for convenience:

Theorem 1.5 (Restated)**.** *There is a universal $\epsilon_0 > 0$ such that whenever $k \leq (3/4)n$ (and is divisible by 3), any non-adaptive algorithm for ϵ_0-testing Maj_k-isomorphism must make at least $\Omega(k^{1/12})$ queries.*

Note that the result of Theorem 1.5 cannot be handled by Theorem 1.3 since the Maj_k function is $o(1)$-close to being a $(k-e)$-junta whenever $e = o(k)$. This is shown in the the following proposition, which is very similar to the problem of computing the noise sensitivity of majority [Per04].

Proposition 4.1. *For $0 < e < k/2$, the Maj_k function is ϵ-close to the Maj_{k-e} function, where $\epsilon = 6(\epsilon/k)^{1/3}$.*

The proof of Proposition 4.1 is included in Appendix C.

4.1 Proof of Theorem 1.5

Our proof of Theorem 1.5 reuses much of the framework introduced in the previous section in our testing lower bound for general g. As before, our goal is to construct probability distributions $\mathcal{F}_{\mathrm{yes}}$ and $\mathcal{F}_{\mathrm{no}}$ over functions isomorphic to Maj_k and functions ϵ_0-far from being isomorphic to Maj_k (respectively) such that any deterministic non-adaptive testing algorithm making $o(k^{1/12})$ queries cannot distinguish with probability at least 1/3 between functions drawn from $\mathcal{F}_{\mathrm{yes}}$ or from $\mathcal{F}_{\mathrm{no}}$.

We define $\mathcal{F}_{\mathrm{yes}}$ just as we did in Section 3: a function $f_{\mathrm{yes}} \sim \mathcal{F}_{\mathrm{yes}}$ is obtained by choosing $j_1, \ldots, j_k$ randomly and without replacement from $[n]$ and defining $f_{\mathrm{yes}}(x) = \mathrm{Maj}_k(x_{j_1}, \ldots, x_{j_k})$. The definition $\mathcal{F}_{\mathrm{no}}$, however, is very different from the definition we used in that section to ensure that it is supported on functions far from Maj_k.

The distribution $\mathcal{F}_{\mathrm{no}}$. To define $\mathcal{F}_{\mathrm{no}}$, we first introduce a certain *weighted* majority function WgtMaj_k on $(4/3)k$ bits (note that $(4/3)k \leq n$):

$$\mathrm{WgtMaj}_k(x_1, \ldots, x_{(4/3)k}) = \mathrm{sgn}\Big(\sum_{i=1}^{k/3}\big(\tfrac{1}{2}x_{4i-3} + \tfrac{1}{2}x_{4i-2} + \tfrac{1}{2}x_{4i-1} + \tfrac{3}{2}x_{4i}\big)\Big). \tag{3}$$

I.e., WgtMaj_k gives k variables weight $\frac{1}{2}$ and $k/3$ variables weight $\frac{3}{2}$. This weight pattern is chosen very carefully; see the proof of Lemma 4.7 below. We then define $f_{\mathrm{no}} \sim \mathcal{F}_{\mathrm{no}}$ is obtained by choosing $j_1, \ldots, j_{(4/3)k}$ randomly and without replacement from $[n]$ and taking $f_{\mathrm{no}}(x) = \mathrm{WgtMaj}_k(x_{j_1}, \ldots, x_{j_{(4/3)k}})$. We have the following key lemma:

Proposition 4.2. *There exist universal constants $\epsilon_0 > 0$ and $k_0 \in \mathbb{N}$ such that when $k \geq k_0$, every function f_{no} in the support of $\mathcal{F}_{\mathrm{no}}$ is ϵ_0-far from being a k-junta.*

Note that we may always assume $k \geq k_0$ as otherwise Theorem 1.5 is trivial. The proof of Proposition 4.2 is presented in Appendix C. To complete the proof of Theorem 1.5, it suffices to prove the following:

Theorem 4.3. *Let $\mathcal{T}$ be any* deterministic *non-adaptive q-query algorithm for testing isomorphism to* Maj_k. *Then*

$$\left|\Pr_{f_{\mathrm{yes}}\sim\mathcal{F}_{\mathrm{yes}}}[\mathcal{T} \textit{ accepts } f_{\mathrm{yes}}] - \Pr_{f_{\mathrm{no}}\sim\mathcal{F}_{\mathrm{no}}}[\mathcal{T} \textit{ accepts } f_{\mathrm{no}}]\right| \le O(q^{3/2}/k^{1/8}).$$

To prove Theorem 4.3, we continue to recall the framework developed in Section 3. Given a deterministic q-query tester $\mathcal{T}$, we arrange its q queries $x^1, \dots, x^q \in \{-1,1\}^n$ into a $q \times n$ Query Matrix Q. We again think of two processes $\mathcal{S}_{\mathrm{yes}}$ and $\mathcal{S}_{\mathrm{no}}$ for generating Argument Multisets S_{yes} and S_{no}. However in the present case we simply have that $\mathcal{S}_{\mathrm{yes}}$ chooses k Columns at random from Q without replacement, and $\mathcal{S}_{\mathrm{no}}$ chooses $(4/3)k$ Columns at random from Q without replacement. Again, we imagine that the Argument Multisets are randomly ordered to form Argument Matrices: A_{yes} which is $q \times k$, and A_{no} which is $q \times (4/3)k$. Finally, we obtain the Response Vector random variable $R_{\mathrm{yes}} \in \{-1,1\}^q$ by applying Maj_k to A_{yes} row-wise, and the Response Vector $R_{\mathrm{no}} \in \{-1,1\}^q$ by applying WgtMaj_k to A_{no} row-wise. It is clear that this distribution on R_{yes} is equivalent to the one given by drawing $f_{\mathrm{yes}} \sim \mathcal{F}_{\mathrm{yes}}$ and letting $R_{\mathrm{yes}} = \langle f_{\mathrm{yes}}(x^1), \dots, f_{\mathrm{yes}}(x^q)\rangle$. The analogous statement is true for R_{no}. Hence we can prove Theorem 4.3 by showing

$$d_{\mathrm{TV}}(R_{\mathrm{yes}}, R_{\mathrm{no}}) \le O(q^{3/2}/k^{1/8}). \tag{4}$$

We now come to the main difference between our Maj_k lower bound and the general lower bound from Section 3. Obviously, we cannot proceed as in Section 3 by bounding the total variation distance between S_{yes} and S_{no}: this total variation distance is 1, since $\mathcal{S}_{\mathrm{yes}}$ and $\mathcal{S}_{\mathrm{no}}$ have disjoint support! (Specifically, the multiset S_{yes} has cardinality k whereas S_{no} has cardinality $(4/3)k$.) Instead, we exploit the fact that applying Maj_k or WgtMaj_k involves *adding up* the Columns in the Argument Matrix (in WgtMaj_k's case, with certain weights), and this addition "loses a lot of information".

More precisely, suppose that we write $X_1, \dots, X_k$ for the (randomly chosen) Columns in Argument Matrix A_{yes} and let

$$S = X_1 + \cdots + X_k.$$

Then the Response Vector R_{yes} is given by taking the sgn of each entry of S; i.e., it is determined by the orthant of $\mathbb{R}^q$ in which S lies. Similarly, if we write $Y_1, \dots, Y_{(4/3)k}$ for the Columns in Argument Matrix A_{no}, and let

$$T = \tfrac{1}{2}Y_1 + \tfrac{1}{2}Y_2 + \tfrac{1}{2}Y_3 + \tfrac{3}{2}Y_4 + \cdots + \tfrac{1}{2}Y_{(4/3)k-3} + \tfrac{1}{2}Y_{(4/3)k-2} + \tfrac{1}{2}Y_{(4/3)k-1} + \tfrac{3}{2}Y_{(4/3)k},$$

then R_{no} is determined by the orthant in which T lies. Hence we can establish (4) and thus Theorem 1.5 by proving the following:

Theorem 4.4. *Let S and T be defined as above. Then for any union $\mathcal{O}$ of orthants in $\mathbb{R}^d$,*

$$|\Pr[S \in \mathcal{O}] - \Pr[T \in \mathcal{O}]| \le O(q^{3/2}/k^{1/8}).$$

4.2 Multidimensional invariance

To prove Theorem 4.4 we will use recently developed invariance principle tools [MOO05, Mos08, GOWZ09]. In particular, we quote the following multidimensional results which essentially appear in [Mos08, GOWZ09].

Lemma 4.5. *(Essentially Theorem 4.1 in [Mos08]; cf. [GOWZ09].) Let $\psi : \mathbb{R}^q \to \mathbb{R}$ be a thrice continuously differentiable function with uniformly bounded third partial derivatives: $|\psi^{(J)}| \le \beta$ for all multi-indices $J = (j_1, \dots, j_q)$ with $|J| = j_1 + \cdots + j_q = 3$. Let $S = S_1 + \cdots + S_m$, where the S_i's are independent $\mathbb{R}^q$-valued random variables, and let $T = T_1 + \cdots + T_m$ similarly. Assume that for each $i \in [m]$, S_i and T_i have matching means and covariance matrices: $\mathbf{E}[S_i] = \mathbf{E}[T_i]$ and $\mathbf{Cov}[S_i] = \mathbf{Cov}[T_i]$. Then*

$$|\mathbf{E}[\psi(S)] - \mathbf{E}[\psi(T)]| \le O(\beta) \cdot \sum_{i=1}^{m} \sum_{|J|=3} \left(\mathbf{E}[|S_i^J|] + \mathbf{E}[|T_i^J|]\right),$$

where U^J denotes $U_1^{j_1} \cdots U_q^{j_q}$ when $U \in \mathbb{R}^q$.

Lemma 4.6. *(Essentially appears in [GOWZ09].) Let $\mathcal{O}$ be any union of orthants in $\mathbb{R}^q$ and let S, T be any $\mathbb{R}^q$-valued random variables. Let $r > 0$. Then there is a certain smooth function ψ satisfying $|\psi^{(J)}| \le O(1/r^3)$ for all multi-indices J with $|J| = 3$ and such that*

$$|\Pr[S \in \mathcal{O}] - \Pr[T \in \mathcal{O}]| \le \Pr[S \in W_r] + |\mathbf{E}[\psi(S)] - \mathbf{E}[\psi(T)]|,$$

where

$$W_r = \{x \in \mathbb{R}^q : |x_i| \le r/2 \text{ for some } i \in [q]\}.$$

At first, it does not appear as though these tools are of any help to us, because Lemma 4.5 very crucially uses the fact that the random vectors being summed are independent. Whereas, in our Theorem 4.4 the random vectors $X_1, \dots, X_k$ are certainly not independent, being drawn randomly *without replacement* from the fixed population Q. The same goes for $Y_1, \dots, Y_{(4/3)k}$. Nevertheless, we can still reduce to Lemma 4.5 using a trick: finding random vectors which are *conditionally independent*.

4.3 How to handle drawing without replacement

Let us recap the scenario in Theorem 4.4. We have a fixed multiset Q of n Columns (vectors) from $\{-1,1\}^q$. We draw k Columns randomly from Q without replacement, yielding the vector-valued random variables $X_1, \dots, X_k$; we also define

$$S = X_1 + \cdots + X_k.$$

Similarly, the vector-valued random variables $Y_1, \dots, Y_{(4/3)k}$ are drawn randomly from Q without replacement, and

$$T = \tfrac{1}{2}Y_1 + \tfrac{1}{2}Y_2 + \tfrac{1}{2}Y_3 + \tfrac{3}{2}Y_4 + \cdots + \tfrac{1}{2}Y_{(4/3)k-3} + \tfrac{1}{2}Y_{(4/3)k-2} + \tfrac{1}{2}Y_{(4/3)k-1} + \tfrac{3}{2}Y_{(4/3)k}.$$

To introduce conditional independence, we reimagine how the X_i's and Y_i's are drawn. Specifically, we can couple the drawing of these random vectors as follows:

1. Define $m = k/3$ (an integer).
2. Randomly partition the Columns of Q into m parts $Q_1, \dots, Q_m$, each of cardinality 4, along with a leftover set of cardinality $n - 4m \ge 0$.
3. Independently for each $i \in [m]$, choose X_{3i-2}, X_{3i-1}, X_{3i} randomly without replacement from Q_i. Define also $S_i = X_{3i-2} + X_{3i-1} + X_{3i}$.
4. Independently for each $i \in [m]$, choose Y_{4i-3}, Y_{4i-2}, Y_{4i-1}, Y_{4i}, randomly without replacement from Q_i (i.e., choose them by randomly ordering the vectors in Q_i). Define also $T_i = \frac{1}{2}Y_{4i-3} + \frac{1}{2}Y_{4i-1} + \frac{1}{2}Y_{4i-2} + \frac{3}{2}Y_{4i}$.

It is easy to see that this coupling gives the correct marginal distributions on $X_1, \dots, X_k$ and $Y_1, \dots, Y_{(4/3)k}$. We also have $S = S_1 + \cdots + S_m$ and $T = T_1 + \cdots + T_m$. And crucially, $S_1, \cdots, S_m$ are independent *conditioned on any choice of the partition* $(Q_1, \dots, Q_m)$, and similarly for $T_1, \dots, T_m$. The following lemma will allow us to apply Lemma 4.5; it also explains the choice of the weight pattern $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{3}{2})$:

Lemma 4.7. *Conditioned on any choice of the partition* $(Q_1, \dots, Q_m)$*, we have* $\mathbf{E}[S_i] = \mathbf{E}[T_i]$ *and* $\mathbf{Cov}[S_i] = \mathbf{Cov}[T_i]$ *for each* $i \in [m]$.

Let $W_r = \{x \in \mathbb{R}^d : \exists j \in [d] \ s.t. |x_j| \le r\}$ represent the region around the orthant boundaries. The following Lemma gives an upper bound on the probability that our random vector S lands near any orthant boundary:

Lemma 4.8. *For* $r \ge 1$ *it holds that* $\mathbf{Pr}[S \in W_r] \le O(qr/\sqrt{m})$.

We present the proofs of Lemmas 4.7 and 4.8 in Appendix C. Let us now combine all these results to complete the proof of Theorem 4.4 and hence Theorem 1.5:

Proof of Theorem 4.4. Let us first condition on a particular partition $(Q_1, \dots, Q_m)$. Having done so, $S_1, \dots, S_m$ become independent, as do $T_1, \dots, T_m$. By Lemma 4.7, we may apply Lemma 4.5. Doing so with the function ψ from Lemma 4.6 (with $r \ge 1$ to be chosen later) yields

$$|\,\mathbf{E}[\psi(S)] - \mathbf{E}[\psi(T)]| \le O(1/r^3) \cdot \sum_{i=1}^{m} \sum_{|J|=3} \left(\mathbf{E}[|S_i^J|] + \mathbf{E}[|T_i^J|]\right). \tag{5}$$

We emphasize that (5) is conditional on a particular $(Q_1, \dots, Q_m)$. However, note that we can bound the quantities $\mathbf{E}[|S_i^J|]$ and $\mathbf{E}[|T_i^J|]$ uniformly in $(Q_1, \dots, Q_m)$: Each coordinate of S_i is at most $1+1+1=3$ in absolute value, and similarly each coordinate of T_i is at most $\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{3}{2}=3$ in absolute value. Hence each expectation is at most 27, and we can therefore upper-bound the right-hand side of (5) by $O(mq^3/r^3)$, since there are at most q^3 many J's.

Now averaging over the choice of partition $(Q_1, \ldots, Q_m)$, the triangle inequality implies

$$|\mathbf{E}[\psi(S)] - \mathbf{E}[\psi(T)]| \le O(mq^3/r^3).$$

Here the expectation is over the whole definition of S and T. Substituting this into Lemma 4.6 gives

$$|\mathbf{Pr}[S \in \mathcal{O}] - \mathbf{Pr}[T \in \mathcal{O}]| \le \mathbf{Pr}[S \in W_r] + O(mq^3/r^3).$$

Applying Lemma 4.8 we can bound this by $O(qr/\sqrt{m}) + O(mq^3/r^3)$. We optimize by taking $r = m^{3/8}d^{1/2}$, yielding a final upper bound of $O(q^{3/2}/m^{1/8}) = O(q^{3/2}/k^{1/8})$ and completing the proof. □

5 Discussion

We conclude this work by discussing what we feel are promising directions towards closing the basic research problem of characterizing the functions g for which g-isomorphism is testable.

It is possible that Fischer et al. [FKR⁺04]'s positive result, that testing g-isomorphism is easy when g is an $O(1)$-junta is mostly best possible. We pose the following question:

Question: Suppose $g : \{0,1\}^n \to \{0,1\}$ is an $(n-\ell)$-junta which is ϵ-far from being an ℓ-junta. Is it true that ϵ-testing g-isomorphism requires $\omega_\ell(1)$ queries?

We are not quite bold enough to conjecture that this is true, but we do not know any g which rules it out. Proving the result seems like it might be difficult, but we believe the problem is approachable for the special case when g is a *symmetric* k-junta. Our Theorem 1.4 establishes the result for the simplest symmetric function, Maj_k. It is not too hard to extend our methods to deal with similar symmetric functions; for example, we are able to show (proof omitted) that testing g-isomorphism is hard for a function such as

$$g : \{0,1\}^n \to \{0,1\}, \qquad g(x) = \begin{cases} 1 & \text{if } k/2 - \sqrt{k} \le \sum_{i=1}^k x_i \le k/2 + \sqrt{k}, \\ 0 & \text{else.} \end{cases}$$

Roughly speaking, we can handle this case because it has only a constant number of "jumps" (just 2, in fact) between 0 and 1 on the main range of $\sum_{i=1}^k x_i$, namely $k/2 \pm O(\sqrt{k})$. If, on the other hand, g is a symmetric k-junta with "many" jumps between 0 and 1 on the main range of $\sum_{i=1}^k x_i$ (e.g., if g is Parity_k), then the techniques we used for our general lower bound Theorem 1.3 may begin to apply.

However, we wish to close by drawing attention to a peculiar intermediate case. Suppose g is the following symmetric k-junta:

$$g(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^k x_i = 1 \text{ or } 3 \le \sum_{i=1}^k x_i < k/2, \\ 1 & \text{if } \sum_{i=1}^k x_i \in \{0,2\} \text{ or } k/2 \le \sum_{i=1}^k x_i \le k, \end{cases}$$

In this case g is $o_k(1)$-close to Maj_k. Hence it would seem at first blush that the few jumps between 0 and 1 when $\sum_{i=1}^k \le 3$ are irrelevant, and we should have an $\omega_k(1)$ lower bound for testing isomorphism to this g.

But oddly, this is not clear. Because g is so close to Maj_k, it seems we would need to use "$\mathcal{F}_{\text{no}}$ functions" which are fairly different from Maj_k, like the weighted majority function WgtMaj_k introduced in Section 4. However there is a *one-query* test that distinguishes between a function isomorphic to the above g and a function isomorphic to WgtMaj_k: simply query the string $(0,0,\ldots,0)$, which has value 1 under g and value 0 under WgtMaj_k! We could fix this by changing $\text{WgtMaj}_k(0,\ldots,0)$ to 0, but there are still problems. For example, the tester could query random strings having a $1/k$ fraction of 1's. For such strings x, $\mathbf{Pr}[g(x) = 1]$ will be noticeably higher than $\mathbf{Pr}[\text{Maj}_k(x) = 1]$, because there is a good chance that the string x will contain exactly two 1's among the k coordinates on which g depends.

So strangely, even though strings in $\{0,1\}^k$ with zero, one, or two 1's constitute only an $o_k(1) \ll \epsilon$ probability mass, a clever tester can exploit them for its advantage. This makes proving untestability for functions isomorphic to the above g somewhat tricky, and we leave it as a problem for future research.

Acknowledgments

The second author would like to thank Adi Akavia and Guy Kindler for several helpful discussions.

References

[AFNS06] Noga Alon, Eldar Fischer, Ilan Newman, and Asaf Shapira. A combinatorial characterization of the testable graph properties: It's all about regularity. In *Proceedings of the 38th annual ACM Symposium on the Theory of Computing*, pages 251–260, 2006.

[AS05a] Noga Alon and Asaf Shapira. A characterization of the (natural) graph properties testable with one-sided error. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science*, pages 429–438, 2005.

[AS05b] Noga Alon and Asaf Shapira. Every monotone graph property is testable. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science*, pages 128–137, 2005.

[AT08] Tim Austin and Terence Tao. On the testability and repair of hereditary hypergraph properties, 2008.

[BGS98] Mihir Bellare, Oded Goldreich, and Madhu Sudan. Free bits, PCPs and non-approximability – towards tight results. *SIAM J. Comput.*, 27(3):804–915, 1998.

[Fis05] Eldar Fischer. The difficulty of testing for isomorphism against a graph that is given in advance. *SIAM J. Comput.*, 34(5):1147–1158, 2005.

[FKR$^+$04] Eldar Fischer, Guy Kindler, Dana Ron, Shmuel Safra, and Alex Samorodnitsky. Testing juntas. *J. Comput. Syst. Sci.*, 68(4):753–787, 2004.

[FM08] Eldar Fischer and Arie Matsliah. Testing graph isomorphism. *SIAM J. Comput.*, 38(1):207–225, 2008.

[GOWZ09] Parikshit Gopalan, Ryan O'Donnell, Yi Wu, and David Zuckerman. Fooling functions of halfspaces under product distributions, 2009. Submitted.

[Hög78] Thomas Höglund. Sampling from a finite population. A remainder term estimate. *Scandinavian Journal of Statistics*, 5:69–71, 1978.

[MOO05] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science*, pages 21–30, 2005. To appear, *Annals of Mathematics*.

[MORS09a] Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco A. Servedio. Testing halfspaces. In *SODA '09: Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms*, pages 256–264, 2009.

[MORS09b] Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco A. Servedio. Testing ±1-weight halfspace. In *RANDOM '09*, pages 646–657, 2009.

[Mos08] Elchanan Mossel. Gaussian bounds for noise correlation of functions and tight analysis of long codes. In *Proceedings of the 49th IEEE Symposium on Foundations of Computer Science*, pages 156–165, 2008.

[Per04] Yuval Peres. Noise stability of weighted majority, 2004.

[PRS02] Michal Parnas, Dana Ron, and Alex Samorodnitsky. Testing basic boolean formulae. *SIAM J. Discrete Math.*, 16(1):20–46, 2002.

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

[She07] Irina Shevtsova. Sharpening of the upper bound of the absolute constant in the Berry–Esseen inequality. *Theory of Probability and its Applications*, 51(3):549–553, 2007.

[Yao77] Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the 28th IEEE Symposium on Foundations of Computer Science*, pages 222–227, 1977.

A Proofs for Section 2

We include in this section the proofs for Lemma 2.1 and Proposition 2.2.

A.1 Proof of Lemma 2.1

Lemma 2.1 is implied by Höglund's Theorem on the normal approximation of hypergeometric distributions.

Höglund Theorem ([Hög78])**.** *Let F be the cdf of the random variable $X = X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are chosen uniformly at random without replacement from $A = \{x_1, \ldots, x_N\}$. Then for all $t \in \mathbb{R}$,*

$$\left| F(t) - \Phi\left(\frac{t - n\mu}{\sigma\sqrt{n(1-n/N)}} \right) \right| \leq C \cdot \frac{\sum_{i=1}^N |x_i - \mu|^3 / N}{\sigma^3 \sqrt{n(1-n/N)}},$$

where $\mu = \sum_{i=1}^N x_i/N$, $\sigma^2 = \sum_{i=1}^N (x_i - \mu)^2/N$, and C is an absolute constant.

Corollary 2.1 (Restated)**.** *Let $X \sim \mathrm{Hyp}(n, m, k)$ and let $\sigma^2 = \frac{km}{n}\left(1 - \frac{m}{n}\right)\left(1 - \frac{k}{n}\right)$. Then for every $t \geq 0$,*

$$\mathbf{Pr}[X = t] \leq \frac{C}{\sigma}$$

where C is an absolute constant.

Proof. Let $X = X_1 + \cdots X_k$ where $X_1, \ldots, X_k$ are chosen uniformly at random without replacement from the set A containing m ones and $n - m$ zeros. Then $X \sim \mathrm{Hyp}(n, m, k)$ and if we let F represent the cdf of X and $\mu = km/n$, Höglund's Theorem implies that for every $t \in \mathbb{R}$

$$\left| F(t) - \Phi\left(\frac{t - \mu}{\sigma} \right) \right| \leq c_0 \cdot \frac{\sum |x_i - m/n|^3}{\sum (x_i - m/n)^2} \cdot \frac{1}{\sigma}$$

for some absolute constant c_0. Furthermore, for every $i = 1, \ldots, n$, $|x_i - m/n| \in [0, 1]$ so $|x_i - m/n|^3 \leq (x_i - m/n)^2$ and

$$\left| F(t) - \Phi\left(\frac{t - \mu}{\sigma} \right) \right| \leq \frac{c_0}{\sigma}.$$

Finally,

$$\mathbf{Pr}[X = t] = F(t) - F(t-1) \leq \Phi\left(\frac{t - \mu}{\sigma} \right) - \Phi\left(\frac{(t-1) - \mu}{\sigma} \right) + 2\frac{c_0}{\sigma} \leq \frac{C}{\sigma},$$

where $C = 2c_0 + \sqrt{2/\pi}$. □

A.2 Proof of Proposition 2.2

Proposition 2.2 (Restated)**.** *Let $f : \{0,1\}^n \to \{0,1\}$ be a k-junta, where $J \subseteq [n]$ is the set of relevant variables. Define $\tau = \min_{I \subseteq J, |I| = e} \mathrm{Inf}_I(f)$. Then f is 2τ-far and 4τ-close to being a $(k-e)$-junta.*

In order to prove Proposition 2.2, we introduce one more definition and fact. For any function $f : \{0,1\}^n \to \{0,1\}$, we say that the function $g : \{0,1\}^n \to \{0,1\}$ is a *projection of f onto I* if the relevant variables in g are contained in I and for every $x \in \{0,1\}^n$, $\mathbf{Pr}_y[f(x_I y_{[n]\setminus I}) = g(x)] \geq 1/2$.

Proposition A.1. *Let g be a projection of f onto I, and let h be any other junta on I. Then $\partial(f, g) \leq \partial(f, h)$.*

Proof. We prove the proposition by contradiction. Let h be a junta on I that minimizes the distance $\partial(f, h)$. Assume that h is not a projection of f on I. Then there is an element $z \in \{0,1\}^n$ such that $\mathbf{Pr}_y[f(z_I y_{[n]\setminus I}) = h(z)] < 1/2$. Define h' by setting

$$h'(x) = \begin{cases} 1 - h(x) & \text{if } x_I = z_I \\ h(x) & \text{otherwise.} \end{cases}$$

The function h' is also a junta on I. Let $\bar{I} = [n] \setminus I$. Then

$$\begin{aligned} \partial(f, h') &= \mathop{\mathbf{Pr}}_x[f(x) \neq h'(x)] \\ &= \mathop{\mathbf{E}}_x[\mathop{\mathbf{Pr}}_y[f(x_I y_{\bar{I}}) \neq h'(x_I y_{\bar{I}})]] \\ &= \frac{1}{2^{|I|}} \mathop{\mathbf{Pr}}_y[f(z_I y_{\bar{I}}) \neq h'(z_I y_{\bar{I}})] + \left(1 - \frac{1}{2^{|I|}}\right) \mathop{\mathbf{Pr}}_y[f(x_I y_{\bar{I}}) \neq h'(x_I y_{\bar{I}}) \mid x_I \neq z_I]. \end{aligned}$$

By definition of h',

$$\Pr_y[f(z_I y_{\bar{I}}) \neq h'(z_I y_{\bar{I}})] = \Pr_y[f(z_I y_{\bar{I}}) = h(z_I y_{\bar{I}})] < 1/2 < \Pr_y[f(z_I y_{\bar{I}}) \neq h(z_I y_{\bar{I}})]$$

and

$$\Pr_y[f(x_I y_{\bar{I}}) \neq h'(x_I y_{\bar{I}}) \mid x_I \neq z_I] = \Pr_y[f(x_I y_{\bar{I}}) \neq h(x_I y_{\bar{I}}) \mid x_I \neq z_I].$$

So $\partial(f, h') < \partial(f, h)$, which contradicts our assumption that h is a junta on I with minimal distance to f. □

We now complete the proof of Proposition 2.2.

Proof of Proposition 2.2. Let $I = \operatorname{argmin}_{I \subseteq J, |I|=e} \operatorname{Inf}_I(f)$, define $J' = J \setminus I$, and let g be a projection of f on J'. By Proposition A.1, it suffices to show that $\partial(f, g)/2 \leq \tau \leq \partial(f, g)$.

For $x \in \{0, 1\}^n$, let us write $\kappa(x) = \mathbf{Pr}_y[f(x_{J'} y_{[n] \setminus J'} \neq g(x)]$. Then

$$\partial(f, g) = \mathop{\mathbf{E}}_x[\mathop{\mathbf{Pr}}_y[f(x_{J'} y_{[n] \setminus J'}) \neq g(x)]] = \mathop{\mathbf{E}}_x[\kappa(x)].$$

Also, letting $\bar{I} = [n] \setminus I$, we have that $\operatorname{Inf}_I(f) = 2\,\mathbf{E}_x[\mathbf{Pr}_{y,z}[f(x_I y_{\bar{I}}) \neq f(x_I z_{\bar{I}})]]$. The event $f(x_I y_{\bar{I}}) \neq f(x_I z_{\bar{I}})$ occurs only when $f(x_I y_{\bar{I}}) = g(x)$ and $g(x) \neq f(x_I z_{\bar{I}})$, or when $f(x_I y_{\bar{I}}) \neq g(x)$ and $g(x) = f(x_I z_{\bar{I}})$, so

$$\begin{aligned}
\operatorname{Inf}_I(f) &= 2 \mathop{\mathbf{E}}_x\left[\mathop{\mathbf{Pr}}_{y,z}[f(x_I y_{\bar{I}}) = g(x) \cap g(x) \neq f(x_I z_{\bar{I}})] + \mathop{\mathbf{Pr}}_{y,z}[f(x_I y_{\bar{I}}) \neq g(x) \cap g(x) = f(x_I y_{\bar{I}})]\right] \\
&= 4 \mathop{\mathbf{E}}_x\left[\mathop{\mathbf{Pr}}_y[f(x_I y_{\bar{I}}) = g(x)] \mathop{\mathbf{Pr}}_z[g(x) \neq f(x_I z_{\bar{I}})]\right] \\
&= 4 \mathop{\mathbf{E}}_x[\kappa(x)(1 - \kappa(x))].
\end{aligned}$$

Since g is a projection of f onto J', $1/2 \leq \kappa(x) \leq 1$ for every $x \in \{0, 1\}^n$, and the Proposition follows. □

B Proofs for Section 3

B.1 Proof of Lemma 3.5

This section is devoted to the proof of Lemma 3.5. Recall that experiments $\mathcal{S}_{\text{yes}}$ and $\mathcal{S}_{\text{no}}$ generate random Argument Multisets S_{yes} and S_{no} (respectively) as follow. We have a fixed Query Matrix Q, thought of as a multiset of n Columns from $\mathfrak{C}$. In $\mathcal{S}_{\text{yes}}$, we form the multiset S_{yes} by drawing $k = \ell + e$ Columns from Q without replacement. In $\mathcal{S}_{\text{no}}$, we form the multiset S_{no} by drawing ℓ Columns from Q without replacement, and adding an additional e copies of the first-drawn Column. Our goal is to bound $d_{\text{TV}}(S_{\text{yes}}, S_{\text{no}})$.

Let us think of the experiment $\mathcal{S}_{\text{yes}}$ in an alternate way. We begin by choosing a first Column from Q for S_{yes} — call it C_1. We next decide how many additional copies of C_1 to include into S_{yes}. Call this quantity T. We have

$$T \mid (C_1 = \mathfrak{c}) \sim \operatorname{Hyp}(n - 1, m(\mathfrak{c}) - 1, k - 1).$$

(Note that $m(\mathfrak{c}) - 1 \geq 0$ always, because $\mathfrak{c}$ won't be chosen if $m(\mathfrak{c}) = 0$.) So far, S_{yes} consists of $T + 1$ copies of C_1. Finally, we complete the draw of S_{yes} by choosing $k - (T + 1)$ Columns without replacement from "$Q \setminus C_1$", meaning the multiset of Columns formed from Q by removing all copies of C_1.

We think of the experiment $\mathcal{S}_{\text{no}}$ in a similar way. Again, we begin by choosing a first column C_1 from Q for S_{no}. We next determine how many additional copies of C_1 there will be from among the remaining $\ell - 1$ choices. Calling this quantity U, we have

$$U \mid (C_1 = \mathfrak{c}) \sim \operatorname{Hyp}(n - 1, m(\mathfrak{c}) - 1, \ell - 1).$$

Recall, however, that in $\mathcal{S}_{\text{no}}$, we include an additional e copies of C_1 into S_{no}. Hence S_{no} ends up with $U + e + 1$ copies of C_1. Finally, we complete S_{no} by adding $\ell - (U + 1)$ Columns drawn without replacement from $Q \setminus C_1$.

Let $V = U + e$. We claim that by coupling the random variables $T \mid (C_1 = \mathfrak{c})$ and $V \mid (C_1 = \mathfrak{c})$, we couple S_{yes} and S_{no}. This follows immediately from the two descriptions, as then $T + 1 = V + 1 = U + e + 1$, and $k - (T + 1) = \ell + e - (V + 1) = \ell - (U + 1)$. Hence

$$d_{\text{TV}}(S_{\text{yes}}, S_{\text{no}}) \le \sum_{\mathfrak{c} \in \mathfrak{C}} \mathbf{Pr}[C_1 = \mathfrak{c}] \cdot d_{\text{TV}}(T \mid (C_1 = \mathfrak{c}), V \mid (C_1 = \mathfrak{c})).$$

On one hand, $\mathbf{Pr}[C_1 = \mathfrak{c}]$ is simply $\frac{m(\mathfrak{c})}{n}$. On the other hand, we have

$$T \mid (C_1 = \mathfrak{c}) \sim \text{Hyp}(n-1, m(\mathfrak{c}) - 1, \ell + e - 1), \qquad V \mid (C_1 = \mathfrak{c}) \sim \text{Hyp}(n-1, m(\mathfrak{c}) - 1, \ell - 1) + e.$$

So by definition, $d_{\text{TV}}(T \mid (C_1 = \mathfrak{c}), V \mid (C_1 = \mathfrak{c})) = \lambda(n-1, m(\mathfrak{c}) - 1, \ell - 1, e)$, and hence

$$d_{\text{TV}}(S_{\text{yes}}, S_{\text{no}}) \le \sum_{\mathfrak{c} \in \mathfrak{C} : m(\mathfrak{c}) \ne 0} \frac{m(\mathfrak{c})}{n} \cdot \lambda(n-1, m(\mathfrak{c}) - 1, \ell - 1, e),$$

as claimed.

B.2 Total variation between hypergeometrics — the proof of Lemma 3.6

This section is devoted to the proof of Lemma 3.6. Recall that $L' = \min(L, N - L)$,

$$\frac{M}{N} L' \ge \kappa e^2, \tag{6}$$

and our goal is to bound $\lambda(N, M, L, e) = d_{\text{TV}}(X, Y) \le .01$, where $X \sim \text{Hyp}(N, M, L{+}e)$ and $Y \sim \text{Hyp}(N, M, L) + e$.

We begin by coupling X and Y, as follows. Imagine drawing balls randomly and without replacement from an urn containing N balls, M of which are white. We draw $L + e$ balls from the urn. We let X be the number of white balls among all balls drawn; we let Y be the number of white balls among the first L balls drawn, plus e. Note that $X \le Y$ always under this coupling.

Let us now compare the probability mass functions of X and Y. The integers $u < e$ can be in X's range but not Y's; the integers $u > \min(M, L + e)$ can be in Y's range but not X's. The remaining integers are in the range of both X and Y, and we have

$$\mathbf{Pr}[X = u] \Big/ \mathbf{Pr}[Y = u] = \frac{\binom{M}{u}\binom{N-M}{L+e-u}}{\binom{N}{L+e}} \Big/ \frac{\binom{M}{u-e}\binom{N-M}{L+e-u}}{\binom{N}{L}} = \frac{\binom{M}{u}}{\binom{M}{u-e}} \cdot \frac{\binom{N}{L}}{\binom{N}{L+e}}$$
$$= \frac{(M-u+e)(M-u+e-1)\cdots(M-u+1)}{u(u-1)\cdots(u-e+1)} \cdot \frac{\binom{N}{L}}{\binom{N}{L+e}}.$$

Evidently (and unsurprisingly), this ratio is a decreasing function of u. Letting t be the largest integer for which the ratio is at least 1, we conclude that

$$\mathbf{Pr}[X = u] \ge \mathbf{Pr}[Y = u] \text{ iff } u \le t.$$

It follows immediately that

$$d_{\text{TV}}(X, Y) = \mathbf{Pr}[X \le t] - \mathbf{Pr}[Y \le t].$$

But by our coupling,

$$\mathbf{Pr}[X \le t] - \mathbf{Pr}[Y \le t] = \mathbf{Pr}[X \le t \cap Y > t] - \mathbf{Pr}[X > t \cap Y \le t] = \mathbf{Pr}[X \le t \cap Y > t],$$

since $X \le Y$ always. Our goal, then, is to bound

$$d_{\text{TV}}(X, Y) = \mathbf{Pr}[X \le t \cap Y > t]. \tag{7}$$

We will in fact prove something slightly stronger: we will show that for *any* value of t, the right-hand side of (7) is small.

To analyze (7) we recall the ball and urn process defining X and Y. Having drawn $L+e$ balls, let W be the number of white balls among the *last* e balls drawn, and let Z be the number of white balls among the first L. Thus $X = W + Z$ and $Y = e + Z$. As a first observation, we may note that if $W = e$ then $X = Y$ and hence the event in (7) does not occur. I.e.,

$$d_{\mathrm{TV}}(X, Y) \le \mathbf{Pr}[W \neq e] \le e(1 - \tfrac{M}{N}), \tag{8}$$

where we used a union bound over each of the last e balls being non-white. Now by (6),

$$e \le \sqrt{\frac{1}{\kappa}\frac{M}{N}L'} \le \sqrt{\frac{1}{\kappa}}\sqrt{L'} \le .001\sqrt{N}, \tag{9}$$

if we assume κ large enough. It follows that we may additionally assume

$$M \le N - .01\sqrt{N} \qquad \Leftrightarrow \qquad 1 - \frac{M}{N} \ge \frac{.01}{\sqrt{N}} \tag{10}$$

because otherwise the bound in (8) is at most $.001\sqrt{N} \cdot \frac{.01}{\sqrt{N}} = .00001$, which establishes the theorem with room to spare. We also use this opportunity to mention that

$$M \ge 2e, \qquad L' \ge 2e \qquad \text{(and hence certainly } N \ge 2e\text{)} \tag{11}$$

follow easily from (6).

We next give a more refined upper bound on (7). By conditioning on W we have

$$d_{\mathrm{TV}}(X, Y) = \mathbf{Pr}[X \le t \cap Y > t] = \sum_{i=0}^{e-1} \mathbf{Pr}[W = i]\, \mathbf{Pr}[Z \in \{t-e+1, t-e+2, \dots, t-i\} \mid W = i].$$

Now $Z \mid (W = i)$ has distribution $\mathrm{Hyp}(N-e, M-i, L)$ (and note that $M - i \ge M - e \ge 0$ by (11)). Let us write $\sigma^2 = L(1 - \frac{L}{N-e})\frac{M-i}{N-e}(1 - \frac{M-i}{N-e})$. Applying Lemma 2.1 and a union bound we get

$$\begin{aligned} d_{\mathrm{TV}}(X, Y) &\le \sum_{i=0}^{e-1} \mathbf{Pr}[W = i] \cdot (e - i)\frac{C}{\sigma} \\ &\le \max_{0 \le i < e}\left\{\frac{C}{\sigma}\right\} \cdot \sum_{i=0}^{e-1} \mathbf{Pr}[W = i](e - i) \\ &= \max_{0 \le i < e}\left\{\frac{C}{\sigma}\right\} \cdot \mathbf{E}[e - W]. \end{aligned}$$

We have $W \sim \mathrm{Hyp}(N, M, e)$, and thus $\mathbf{E}[e - W] = e(1 - \frac{M}{N})$. And by definition,

$$\max_{0 \le i < e}\left\{\frac{C}{\sigma}\right\} = \max_{0 \le i < e}\left\{\frac{C}{\sqrt{L(1 - \frac{L}{N-e})(\frac{M-i}{N-e})(1 - \frac{M-i}{N-e})}}\right\} \le \frac{C}{\sqrt{L(1 - \frac{L}{N-e})(\frac{M-e}{N-e})(1 - \frac{M}{N-e})}}.$$

Thus we have established

$$d_{\mathrm{TV}}(X, Y) \le \frac{Ce}{\sqrt{L(1 - \frac{L}{N-e})}} \cdot \frac{1 - \frac{M}{N}}{\sqrt{1 - \frac{M}{N-e}}} \cdot \frac{1}{\sqrt{\frac{M-e}{N-e}}}. \tag{12}$$

We will bound the three fractions in (12) one at a time. We begin with the middle one.

$$\frac{d}{dM}\left(\frac{1 - \frac{M}{N}}{\sqrt{1 - \frac{M}{N-e}}}\right) = -\frac{N - 2e - M}{2N\sqrt{1 - \frac{M}{N-e}}(N - e - M)}.$$

By combining (9) and (10) we get $M \le N - 10e < N - 2e$. Hence the derivative above is always negative, implying that

$$\frac{1 - \frac{M}{N}}{\sqrt{1 - \frac{M}{N-e}}} \quad \text{is a decreasing function of } M \text{ on } M\text{'s range.}$$

Hence we may upper-bound this fraction by taking $M = 0$, giving an upper bound of 1. Substituting this into (12) gives

$$d_{\mathrm{TV}}(X, Y) \le \frac{Ce}{\sqrt{L(1 - \frac{L}{N-e})}} \cdot \frac{1}{\sqrt{\frac{M-e}{N-e}}}. \tag{13}$$

We next examine the fraction on the right. It is at most

$$\frac{1}{\sqrt{\frac{M-e}{N}}} \le \frac{1}{\sqrt{\frac{M/2}{N}}} = \sqrt{\frac{2N}{M}},$$

where we used (11). By virtue of (6), we can upper-bound this by $\sqrt{\frac{2}{\kappa}} \cdot \frac{\sqrt{L'}}{e}$. Substituting this upper bound into (13) yields

$$d_{\mathrm{TV}}(X, Y) \le C\sqrt{\frac{2}{\kappa}} \cdot \sqrt{\frac{L'}{L(1 - \frac{L}{N-e})}} \le .001\sqrt{\frac{L'}{L(1 - \frac{L}{N-e})}}, \tag{14}$$

assuming κ is sufficiently large compared with C.

Finally, we split into two cases, depending on whether $L \le N/2$. If indeed $L \le N/2$, then $L' = L$ and we have

$$.001\sqrt{\frac{L'}{L(1 - \frac{L}{N-e})}} = \frac{.001}{\sqrt{1 - \frac{L}{N-e}}} \le \frac{.001}{\sqrt{1 - \frac{N/2}{N-e}}}.$$

But $N - e \ge N - .001\sqrt{N} \ge (2/3)N$ (using (9) and $N \ge 2$ from (11)), so we upper-bound

$$d_{\mathrm{TV}}(X, Y) \le \frac{.001}{\sqrt{1 - \frac{N/2}{(2/3)N}}} = .002 \le .01,$$

as needed. The second case is that $L \ge N/2$, in which case $L = N - L'$ and the bound in (14) is

$$.001\sqrt{\frac{L'}{(N - L')(1 - \frac{N-L'}{N-e})}} = .001\sqrt{\frac{L'}{(N - L')\frac{L'-e}{N-e}}} = .001\sqrt{\frac{L'}{L' - e}}\sqrt{\frac{N-e}{N - L'}}. \tag{15}$$

But using (11),

$$\sqrt{\frac{L'}{L' - e}} \le \sqrt{\frac{L'}{L'/2}} = \sqrt{2},$$

and using $L' \le N/2$,

$$\sqrt{\frac{N-e}{N-L'}} \le \sqrt{\frac{N}{N-L'}} \le \sqrt{\frac{N}{N/2}} = \sqrt{2}.$$

Hence the upper bound (15) on $d_{\mathrm{TV}}(X, Y)$ is at most $.001\sqrt{2}\sqrt{2} \le .01$, as needed.

This completes the proof of Lemma 3.6.

C Proofs for Section 4

C.1 Proof of Proposition 4.1

Proposition 4.1 (Restated)**.** *For $0 < e < k/2$, the* Maj_k *function is ϵ-close to the* Maj_{k-e} *function, where $\epsilon = 6(\epsilon/k)^{1/3}$.*

Proof. We need to upper-bound

$$\mathop{\mathbf{Pr}}_{x}[\mathrm{Maj}_k(x) \neq \mathrm{Maj}_{k-e}(x)].$$

The event $\mathrm{Maj}_k(x) \neq \mathrm{Maj}_{k-e}(x)$ can only occur when $\left|\sum_{i=1}^{k-e} x_i\right| \leq \left|\sum_{i=k-e+1}^{k} x_i\right|$, so it suffices to upper-bound

$$\mathop{\mathbf{Pr}}_{x}\left[\left|\sum_{i=1}^{k-e} x_i\right| \leq \left|\sum_{i=k-e+1}^{k} x_i\right|\right].$$

For any $t > 0$ we may use the crude bound

$$\mathop{\mathbf{Pr}}_{x}\left[\left|\sum_{i=1}^{k-e} x_i\right| \leq \left|\sum_{i=k-e+1}^{k} x_i\right|\right] \leq \mathop{\mathbf{Pr}}_{x}\left[\left|\sum_{i=1}^{k-e} x_i\right| \leq t\right] + \mathop{\mathbf{Pr}}_{x}\left[\left|\sum_{i=k-e+1}^{k} x_i\right| > t\right], \tag{16}$$

Let us now examine both terms on the right-hand side of (16).

For the first term, let $\sigma = \sqrt{k-e}$. The Berry-Esseen Theorem implies that for any $t \in \mathbb{R}$,

$$\left|\mathop{\mathbf{Pr}}_{x}\left[\frac{1}{\sigma}\sum_{i=1}^{k-e} x_i \leq t\right] - \Phi(t)\right| \leq \frac{1}{\sigma}.$$

Hence for $t \geq 1$, say, we have

$$\mathop{\mathbf{Pr}}_{x}\left[\left|\sum_{i=1}^{k-e} x_i\right| \leq t\right] \leq \Phi(t/\sigma) - \Phi(-t/\sigma) + \frac{2}{\sigma} \leq \frac{2t+2}{\sqrt{k-e}} \leq \frac{3t}{\sqrt{k-e}}. \tag{17}$$

For the second term, note that $\mathbf{E}_x[\sum_{i=k-e+1}^{k} x_i] = 0$ and $\mathbf{Var}_x[\sum_{i=k-e+1}^{k} x_i] = e$. So given any $t \geq 1$, Chebyshev's Inequality implies that

$$\mathop{\mathbf{Pr}}_{x}\left[\left|\sum_{i=k-e+1}^{k} x_i\right| > t\right] \leq \frac{e}{t^2}. \tag{18}$$

Substituting the inequalities (17) and (18) in (16) and setting $t = e^{1/3}(k-e)^{1/6} \geq 1$, we get

$$\mathop{\mathbf{Pr}}_{x}\left[\left|\sum_{i=1}^{k-e} x_i\right| \leq t\right] + \mathop{\mathbf{Pr}}_{x}\left[\left|\sum_{i=k-e+1}^{k} x_i\right| > t\right] \leq \frac{4e^{1/3}}{(k-e)^{1/3}} \leq 6(e/k)^{1/3},$$

where the last inequality uses $e < k/2$. □

C.2 Proof of Proposition 4.2

Proposition 4.2 (Restated)**.** *There exist universal constants $\epsilon_0 > 0$ and $k_0 \in \mathbb{N}$ such that when $k \geq k_0$, every function f_{no} in the support of $\mathcal{F}_{\mathrm{no}}$ is ϵ_0-far from being a k-junta.*

Proof. It suffices to show that the function WgtMaj_k is ϵ_0-far from being a k-junta. And by Proposition 2.2, it suffices to show that the influence of any set $I \subseteq [4k/3]$ of size $|I| \geq k/3$ satisfies $\mathrm{Inf}_I(\mathrm{WgtMaj}_k) \geq 2\epsilon_0$.

For $i = 1, \ldots, (4/3)k$, let α_i represent the weight of the ith term in the sum in (3) (i.e., $\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{2}$, $\alpha_4 = \frac{3}{2}, \ldots$). The influence of I in WgtMaj_k is

$$\begin{aligned}
\mathrm{Inf}_I(\mathrm{WgtMaj}_k) &= \Pr_{x,y}\left[\mathrm{sgn}\left(\sum_{i\in[4k/3]}\alpha_i x_i\right) \neq \mathrm{sgn}\left(\sum_{i\in[4k/3]\setminus I}\alpha_i x_i + \sum_{i\in I}\alpha_i y_i\right)\right] \\
&= \Pr_{x,y}\left[\mathrm{sgn}\left(\sum_{i\in[4k/3]}\alpha_i x_i\right) \neq \mathrm{sgn}\left(\sum_{i\in I}\alpha_i y_i\right) \cap \left|\sum_{i\in[4k/3]\setminus I}\alpha_i x_i\right| \leq \left|\sum_{i\in I}\alpha_i y_i\right|\right] \\
&= \frac{1}{2}\Pr_{x,y}\left[\left|\sum_{i\in[4k/3]\setminus I}\alpha_i x_i\right| < \left|\sum_{i\in I}\alpha_i y_i\right|\right].
\end{aligned}$$

As a crude lower bound, we have that for any $t > 0$,

$$\mathrm{Inf}_I(\mathrm{WgtMaj}_k) \geq \frac{1}{2}\Pr_x\left[\left|\sum_{i\in[4k/3]\setminus I}\alpha_i x_i\right| \leq t\right]\Pr_y\left[\left|\sum_{i\in I}\alpha_i y_i\right| \geq t\right]. \tag{19}$$

Let's examine the two terms on the right-hand side of the equation individually.

First, let $\sigma_1^2 = \sum_{i\in[4k/3]\setminus I}\alpha_i^2$. By the Berry-Esseen Theorem, for every $t \in \mathbb{R}$

$$\left|\Pr_x\left[\frac{1}{\sigma_1}\sum_{i\in[4k/3]\setminus I}\alpha_i x_i \leq t\right] - \Phi(t)\right| \leq \frac{3}{2\sigma_1}.$$

Therefore,

$$\Pr_x\left[\left|\sum_{i\in[4k/3]\setminus I}\alpha_i x_i\right| \leq t\right] \geq \Phi(t/\sigma_1) - \Phi(-t/\sigma_1) - 3/\sigma_1 = \mathrm{erf}(t/\sqrt{2}\sigma_1) - 3/\sigma_1.$$

Since $k/4 = k(1/2)^2 \leq \sigma_1^2 \leq k/3 \cdot (3/2)^2 + 2k/3 \cdot (1/2)^2 < k$,

$$\Pr_x\left[\left|\sum_{i\in[4k/3]\setminus I}\alpha_i x_i\right| \leq t\right] \geq \mathrm{erf}(t/\sqrt{2k}) - 6/\sqrt{k}. \tag{20}$$

Let's now examine the other term in (19). Letting $\sigma_2^2 = \sum_{i\in I}\alpha_i^2$, the Berry-Esseen Theorem now implies that for $t \in \mathbb{R}$,

$$\left|\Pr_y\left[\frac{1}{\sigma_2}\sum_{i\in I}\alpha_i y_i \leq t\right] - \Phi(t)\right| \leq \frac{3}{2\sigma_2}$$

so

$$\Pr_y\left[\left|\sum_{i\in I}\alpha_i y_i\right| > t\right] \geq \Phi(-t/\sigma_2) + (1 - \Phi(t/\sigma_2)) - \frac{3}{2\sigma_2} = 2\Phi(-t/\sigma_2) - \frac{3}{2\sigma_2}.$$

Since $\sigma_2^2 \geq k/3 \cdot (1/2)^2 = k/12$,

$$\Pr_y\left[\left|\sum_{i\in I}\alpha_i y_i\right| > t\right] \geq 2\,\Phi\left(-\frac{2\sqrt{3}}{\sqrt{k}}t\right) - \frac{3\sqrt{3}}{\sqrt{k}}. \tag{21}$$

Substituting (20) and (21) into (19) and setting $t = \sqrt{k}/4$, we get that

$$\mathrm{Inf}_I(\mathrm{WgtMaj}_k) \geq \frac{1}{2}\left(\mathrm{erf}(1/4\sqrt{2}) - 6/\sqrt{k}\right)\left(2\,\Phi(-\sqrt{3}/2) - 3\sqrt{3}/\sqrt{k}\right).$$

Setting $k_0 = 10^4$ and $\epsilon_0 = 0.01$ (for instance) completes the proof of the Proposition. □

C.3 Proof of Lemma 4.7

Lemma 4.7 (Restated)**.** *Conditioned on any choice of the partition $(Q_1, \dots, Q_m)$, we have*

$$\mathbf{E}[S_i] = \mathbf{E}[T_i], \qquad \mathbf{Cov}[S_i] = \mathbf{Cov}[T_i]$$

for each $i \in [m]$.

Proof. It suffices to check the claim for $i = 1$. Let $\mu \in \mathbb{R}^q$ denote the average of the four vectors in Q_1. Then

$$\mathbf{E}[S_1] = \mathbf{E}[X_1 + X_2 + X_3] = 3\,\mathbf{E}[X_1] = 3\mu,$$

using the fact that X_1, X_2, X_3 are identically distributed, and similarly

$$\mathbf{E}[T_1] = \mathbf{E}[\tfrac{1}{2}Y_1 + \tfrac{1}{2}Y_2 + \tfrac{1}{2}Y_3 + \tfrac{3}{2}Y_4] = (\tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{3}{2})\,\mathbf{E}[Y_1] = 3\mu,$$

verifying the first claim. As for the covariance matrices, fix $j, j' \in [q]$ and let us write x_1 for the j-coordinate of X_1, x_1' for the j'-coordinate of X_1, and similarly $x_2, x_2', \dots, y_4, y_4'$. Then

$$\mathbf{Cov}[S_1]_{j,j'} = \mathbf{E}[(x_1 + x_2 + x_3)(x_1' + x_2' + x_3')] = 3\,\mathbf{E}[x_1x_1'] + 6\,\mathbf{E}[x_1x_2'],$$

using the fact that (x_1, x_2') has the same distribution as (x_1, x_3') and similar facts. And

$$\begin{aligned}
\mathbf{Cov}[T_1]_{j,j'} &= \mathbf{E}[(\tfrac{1}{2}y_1 + \tfrac{1}{2}y_2 + \tfrac{1}{2}y_3 + \tfrac{3}{2}y_4)(\tfrac{1}{2}y_1' + \tfrac{1}{2}y_2' + \tfrac{1}{2}y_3' + \tfrac{3}{2}y_4')] \\
&= 3 \cdot (\tfrac{1}{2})^2\,\mathbf{E}[y_1y_1'] + (\tfrac{3}{2})^2\,\mathbf{E}[y_4y_4'] + 6 \cdot (\tfrac{1}{2})^2\,\mathbf{E}[y_1y_2'] + 6 \cdot \tfrac{1}{2} \cdot \tfrac{3}{2}\,\mathbf{E}[y_1y_4'] \\
&= 3\,\mathbf{E}[y_1y_1'] + 6\,\mathbf{E}[y_1y_2'] \\
&= 3\,\mathbf{E}[x_1x_1'] + 6\,\mathbf{E}[x_1x_2'],
\end{aligned}$$

using the fact that (Y_1, Y_2) has the same distribution as (X_1, X_2) and similar facts. The proof is complete. □

C.4 Proof of Lemma 4.8

Lemma 4.8 (Restated)**.** *For $r \geq 1$ it holds that $\mathbf{Pr}[S \in W_r] \leq O(qr/\sqrt{m})$.*

By union-bounding over the q coordinates, Lemma 4.8 reduces to proving the following statement:

Lemma C.1. *Let $r \geq 1$. Suppose we fix any query row $(x_1, \dots, x_n) \in \{-1, 1\}^n$ from Q and form the random variable*

$$s = x_{i_1} + \cdots + x_{i_{3m}},$$

where the sequence $i_1, \dots, i_{3m}$ is drawn randomly without replacement from $[n]$. Then

$$\mathbf{Pr}[|s| \leq r/2] \leq O(r/\sqrt{m}).$$

Proof. Let us recall that $k = 3m \leq (3/4)n$. Let u denote the number of 1's among $x_1, \dots, x_n$. The statement to be proved is precisely equivalent to the following: Let $Z \sim \mathrm{Hyp}(n, u, k)$. Then

$$\mathbf{Pr}[Z \in [k/2 - r/4, k/2 + r/4]] \leq O(r/\sqrt{k}). \tag{22}$$

We divide into two cases.

Case 1: $1/4 \leq u/n \leq 3/4$. In this case we use Lemma 2.1 and a union bound over the at most $r/2 + 1$ integers in the range $[k/2 - r/4, k/2 + r/4]$ to deduce

$$\mathbf{Pr}[Z \in [k/2 - r/4, k/2 + r/4]] \leq O(r)/\sigma_{n,u,k},$$

where $\sigma_{n,u,k} = \sqrt{k(1 - k/n)(u/n)(1 - u/n)}$. We have $1 - k/n \geq 1/4$ and also $u/n, 1 - u/n \geq 1/4$. Thus $\sigma_{n,u,k} = \Omega(\sqrt{k})$, establishing (22).

Case 2: $u/n \notin [1/4, 3/4]$. By symmetry, it suffices to treat just one of the cases $u/n < 1/4$ or $u/n > 3/4$; say, the former. In this case we have $\mathbf{E}[Z] = k(u/n) \leq k/4$, and we have $\mathbf{Var}[Z] = ku/n(1 - u/n)(1 - k/n) \leq k$. Finally, we may assume that $r \leq k/2$, as otherwise (22) is trivial. Thus by Chebyshev's Inequality,

$$\mathbf{Pr}[Z \geq k/2 - r/4] \leq \mathbf{Pr}[Z \geq (3/8)k] \leq \frac{k}{(k/8)^2} = O(1/k),$$

establishing (22) with room to spare. □

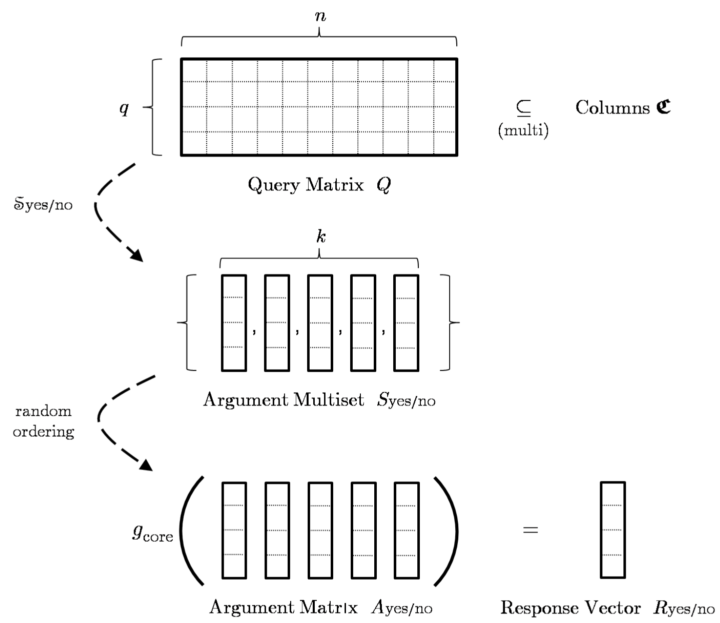
n
q
$\subseteq$
(multi)
Columns $\mathfrak{C}$
Query Matrix Q
$\mathfrak{S}_{\text{yes/no}}$
k
Argument Multiset $S_{\text{yes/no}}$
random ordering
g_{core}
$=$
Argument Matrix $A_{\text{yes/no}}$
Response Vector $R_{\text{yes/no}}$

Figure 1: The complete random processes $\mathcal{P}_{\text{yes}}$ and $\mathcal{P}_{\text{no}}$