Boolean function monotonicity testing requires (almost) $n^{1/2}$ non-adaptive queries

Xi Chen[*]
Columbia University
xichen@cs.columbia.edu

Anindya De[†]
IAS
anindya@math.ias.edu

Rocco A. Servedio[‡]
Columbia University
rocco@cs.columbia.edu

Li-Yang Tan[§]
Simons Institute, UC Berkeley
liyang@cs.columbia.edu

December 19, 2014

Abstract

We prove a lower bound of $\Omega(n^{1/2-c})$, for all $c > 0$, on the query complexity of (two-sided error) non-adaptive algorithms for testing whether an n-variable Boolean function is monotone versus constant-far from monotone. This improves a $\tilde{\Omega}(n^{1/5})$ lower bound for the same problem that was recently given in [CST14] and is very close to $\Omega(n^{1/2})$, which we conjecture is the optimal lower bound for this model.

1 Introduction

1.1 Motivation and background

Monotonicity testing of Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ is one of the most natural and well-studied problems in Property Testing. Introduced by Goldreich, Goldwasser, Lehman, and Ron in 1998 [GGLR98], this problem is concerned with the query complexity of determining whether a Boolean function f is *monotone* or *far from monotone*. Recall that f is monotone if $f(X) \le f(Y)$ for all $X \prec Y$, where $\prec$ denotes the bitwise partial order on the hypercube. We say that f is ε-close to monotone if $\mathbf{Pr}[f(\mathbf{X}) \neq g(\mathbf{X})] \le \varepsilon$ for some monotone Boolean function g, where the probability is over a uniform draw of $\mathbf{X}$ from $\{-1,1\}^n$, and that f is ε-far from monotone otherwise. We are interested in query-efficient randomized algorithms for the following task:

> *Given as input a distance parameter $\varepsilon > 0$ and oracle access to an unknown Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, output* Yes *with probability at least $2/3$ if f is monotone, and* No *with probability at least $2/3$ if f is ε-far from monotone.*

[*] Supported by NSF grant CCF-1149257 and a Sloan research fellowship.

[†] Supported by NSF under agreements Princeton University Prime Award No. CCF-0832797 and Sub-contract No. 00001583.

[‡] Supported by NSF grants CCF-1115703 and CCF-1319788.

[§] This work was done while the author was at Columbia University, supported by NSF grants CCF-1115703 and CCF-1319788.

The work of Goldreich et al. [GGLR98] proposed a simple "edge tester" for this task and proved an $O(n^2 \log(1/\varepsilon)/\varepsilon)$ upper bound on its query complexity, subsequently improved to $O(n/\varepsilon)$ in the journal version [GGL+00]. Fischer et al. [FLN+02] established the first lower bounds shortly after, showing that there exists a constant distance parameter $\varepsilon_0 > 0$ such that $\Omega(\log n)$ queries are necessary for any *non-adaptive* tester (one whose queries do not depend on the oracle's responses to prior queries). This directly implies an $\Omega(\log\log n)$ lower bound for adaptive testers, since any q-query adaptive tester can be simulated by a non-adaptive one that simply carries out all 2^q possible executions. (Via a simple argument [FLN+02] also gave an $\Omega(n^{1/2})$ lower bound for non-adaptive *one-sided* testers, which must output Yes with probability 1 if f is monotone. Throughout this work we consider only general two-sided testers, for which lower bounds are more difficult to prove.)

In spite of considerable work on this problem and its variants [GGLR98, DGL+99, GGL+00, FLN+02, AC06, HK08, BCGSM12], these were the best known results for the basic problem for more than a decade, until Chakrabarty and Seshadhri [CS13] improved on the linear upper bound of Goldreich et al. with an $\tilde{O}(n^{7/8}\varepsilon^{-3/2})$-query tester. More recently, Chen et al. [CST14] closed the gap between upper and lower bounds on the query complexity of non-adaptive testers to within a polynomial factor by giving a lower bound of $\tilde{\Omega}(n^{1/5})$ (an exponential improvement of the [FLN+02] lower bound). [CST14] also gave an upper bound of $\tilde{O}(n^{5/6}\varepsilon^{-4})$ queries (a polynomial improvement of the [CS13] upper bound in terms of the dependence on n).

In this paper we make further progress towards a complete resolution of the problem with a lower bound of (almost) $\Omega(n^{1/2})$ against non-adaptive testers, which we conjecture is optimal. In more detail, our main result is the following:

Theorem 1. *For all $c > 0$ there is a $\kappa = \kappa(c) > 0$ such that any non-adaptive algorithm for testing whether $f : \{-1,1\}^n \to \{-1,1\}$ is monotone versus κ-far from monotone must use $\Omega(n^{1/2-c})$ queries.*

The paper of Chen et al. [CST14] also considered the problem of testing monotonicity of Booelan-valued functions over general hypergrid domains $\{1,\dots,m\}^n$ for $m \geq 2$, and showed that it reduces to that of testing monotonicity Boolean functions as defined above (i.e. the case when $m = 2$) with essentially no loss in parameters. More precisely, they proved that any lower bound for κ-testing monotonicity of $f : \{-1,1\}^n \to \{-1,1\}$ translates into a lower bound for $\Omega(\kappa)$-testing monotonicity of $F : \{1,\dots,m\}^n \to \{-1,1\}$ with only a logarithmic loss in terms of n in the query lower bound. Therefore Theorem 1 along with this reduction yields our most general result:

Theorem 2. *For all $c > 0$ there is a $\kappa = \kappa(c) > 0$ such that for all $m \geq 2$, any non-adaptive algorithm for testing whether $F : \{1,\dots,m\}^n \to \{-1,1\}$ is monotone versus κ-far from monotone must use $\Omega(n^{1/2-c})$ queries.*

1.2 Previous work: the [CST14] lower bound

In order to explain our approach in the current paper we first briefly recall the key elements of the [CST14] lower bound. That paper uses Yao's method, i.e. it exhibits two distributions $\mathcal{D}_{yes}, \mathcal{D}_{no}$ over Boolean functions, where each $\boldsymbol{f} \sim \mathcal{D}_{yes}$ is monotone and almost every $\boldsymbol{f} \sim \mathcal{D}_{no}$ is constant-far from monotone. The main conceptual novelty of the [CST14] lower bound was to use linear threshold functions (LTFs) as both the yes- and no- functions, thereby enabling the application of sophisticated multidimensional central limit theorems to establish the closeness in distribution that is required by Yao's method. In more detail, a function drawn from the "yes-distribution" $\mathcal{D}_{yes}$ of [CST14] is

$$\boldsymbol{f}(X) = \mathrm{sign}(\boldsymbol{u}_1 X_1 + \cdots + \boldsymbol{u}_n X_n) \tag{1}$$

where each $\boldsymbol{u}_i$ is independently uniform over $\{1, 3\}$, and a function drawn from the "no-distribution" $\mathcal{D}_{no}$ is

$$\boldsymbol{f}(X) = \mathrm{sign}(\boldsymbol{v}_1 X_1 + \cdots + \boldsymbol{v}_n X_n) \tag{2}$$

where each $\boldsymbol{v}_i$ is independently -1 with probability $1/10$ and is $7/3$ with probability $9/10$.

Fix an arbitrary (adversarially chosen) $d \times n$ query matrix $\mathcal{X}$ whose elements all are $\pm 1/\sqrt{n}$, and let $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(n)} \in \{\pm 1/\sqrt{n}\}^d$ be the columns of this matrix. The d rows of this matrix correspond to an arbitrary d-element set of n-bit query strings scaled by a factor of $1/\sqrt{n}$. (Note that scaling the input does not change the value of a zero-threshold linear threshold function such as (1) or (2) above.) Define the $\mathbb{R}^d$-valued random variables

$$\mathbf{S} = \sum_{i=1}^{n} \boldsymbol{u}_i \mathcal{X}^{(i)} \quad \text{and} \quad \mathbf{T} = \sum_{i=1}^{n} \boldsymbol{v}_i \mathcal{X}^{(i)}. \tag{3}$$

Recalling (1) and (2) and Yao's minimax lemma, to prove a d-query monotonicity testing lower bound for non-adaptive algorithms, it suffices to upper bound

$$d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) \leq 0.1 \tag{4}$$

(here the "0.1" constant is arbitrary, any constant in $(0, 1)$ would do) for all possible choices of $\mathcal{X}$, where d_{UO} is the "union-of-orthants" distance:

$$d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) := \max\left\{ |\,\mathbf{Pr}[\mathbf{S} \in \mathcal{O}] - \mathbf{Pr}[\mathbf{T} \in \mathcal{O}]| \colon \mathcal{O} \text{ is a union of orthants in } \mathbb{R}^d \right\}.$$

Thus, in this approach, the goal is to make d be as large as possible (as a function of n) while keeping $d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T})$ at most 0.1.

To obtain their main $\tilde{\Omega}(n^{1/5})$ lower bound, [CST14] use a multidimensional central limit theorem (CLT) of Valiant and Valiant [VV11], which is proved using Stein's method and which bounds the earthmover (Wasserstein) distance between sums of independent vector-valued random variables. [CST14] adapts this earthmover CLT to obtain a CLT for the "union-of-orthants" distance d_{UO}, and shows that using this CLT the value of d can be taken as large as $\tilde{\Omega}(n^{1/5})$.

The key properties of the random variables $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ used in [CST14] are that

1. Their first and second moments match, i.e. $\mathbf{E}[\boldsymbol{u}_i] = \mathbf{E}[\boldsymbol{v}_i]$ and $\mathbf{E}[\boldsymbol{u}_i^2] = \mathbf{E}[\boldsymbol{v}_i^2]$. (This ensures that $\mathbf{S}$ and $\mathbf{T}$ have matching means and covariance matrices, which makes it possible to apply the [VV11] CLT.)

2. The random variable $\boldsymbol{u}_i$ is supported entirely on non-negative values, while $\boldsymbol{v}_i$ has nonzero weight on negative values. (The first condition ensures that $\boldsymbol{f} \sim \mathcal{D}_{yes}$ will be monotone, and the second ensures that a random $\boldsymbol{f} \sim \mathcal{D}_{no}$ will w.h.p. be constant-far from monotone.)

1.3 Our approach and techniques

In light of the above, it is natural to ask whether imposing stronger requirements on the $\boldsymbol{u}_i, \boldsymbol{v}_i$ random variables can lead to stronger results: in particular, can matching higher moments than just the first two lead to an improved lower bound? Pursuing such an approach, one quickly discovers that extending the [VV11] CLT for earthmover distance (which, as mentioned above, is proved using Stein's method) to exploit matching higher moments is a nontrivial technical challenge. Instead, in this work we return to a much older proof method for CLTs, namely Lindeberg's "replacement method" (discussed in detail in Section 4.2), which is well suited for higher moments. Our arguments show that by combining a careful construction of the random variables (the coefficients of the LTFs) with a careful analysis of all possible query matrices, the Lindeberg method can be used to obtain an $\Omega(n^{1/2-c})$ lower bound for monotonicity testing.

We observe that a high-level difference between our paper and that of [VV11] is that [VV11] proves that a sum of independent d-dimensional random variables converges to a multi-dimensional Gaussian with matching first two moments (mean and covariance). In contrast, we work with two different but carefully constructed sums of independent d-dimensional random variables which have many matching moments, namely the $\mathbf{S}$ and $\mathbf{T}$ random variables defined in (3). Our goal is *not* to establish smaller distance to a multi-dimensional Gaussian (indeed our arguments do not establish this); rather, as described above, having $d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) \leq 0.1$ is sufficient for our purposes, and our goal is to achieve such "rough" closeness for d-dimensional random variables where d is as large as possible (i.e. as close as possible to $n^{1/2}$).

As a warmup, in Section 4.2 we first prove an $\Omega(n^{1/4-c})$ lower bound via a fairly straightforward application of the Lindeberg method. This argument essentially requires only matching moments of order $1, 2, \ldots, 1/c$ for the $\boldsymbol{u}_i, \boldsymbol{v}_i$ random variables without other special properties — in particular, it does not matter just what those moments are as long as they match each other — and the analysis proceeds in the usual way for the Lindeberg method. However, improving this lower bound to $\Omega(n^{1/2-c})$ requires many new ideas and significantly more care in the construction and analysis. We discuss several of the necessary ingredients, and in so doing give an overview of our proof approach, below.

(1): Suitable choice of distributions. We show that given any positive integer ℓ, there is a non-negative value $\mu = \mu(\ell)$ and a non-negative random variable $\boldsymbol{u}$ such that the first ℓ moments of $\boldsymbol{u}$ match those of the mean-μ, variance-1 Gaussian $\mathcal{N}(\mu, 1)$. (This non-negative support of $\boldsymbol{u}$ ensures that the $\mathcal{D}_{yes}$ functions defined by (1) are monotone as required.) For the $\mathcal{D}_{no}$ functions, we show that there is a random variable $\boldsymbol{v}$ (see (2)) that has first ℓ moments matching those of $\mathcal{N}(\mu, 1)$, has finite support, and takes negative values with nonzero probability. The finite support and negativity conditions enable us to argue that almost all functions drawn from $\mathcal{D}_{no}$ are indeed constant-far from monotone, and the fact that $\boldsymbol{v}$'s moments match those of a Gaussian plays a crucial role in enabling step (4) to go through, as described below.

(2): Careful choice and analysis of mollifier. The Lindeberg method uses smooth "mollifiers" with useful analytic properties (bounded derivatives and the like) to approximate discontinuous indicator functions. We give a careful construction of a particular mollifier which exploits some of the "nice structure" of the sets (unions of orthants) that we must deal with, and show how this mollifier's special properties can be used to obtain a significant savings in bounding the error terms that arise in Lindeberg's method. Our analysis based on this particular mollifier shows that to bound the error terms in Lindeberg's method, it is enough to give an *anticoncentration* bound. In more detail, we identify a family of (roughly) d^{h+1} random variables $\mathbf{R}_{-i}|_J$ (corresponding to the different possible outcomes of the multi-index J in (19); see Section 4.3), and show that it is enough to establish that for almost all of these random variables (outcomes of J), there is a strong upper bound on the probability that $\mathbf{R}_{-i}|_J$ (which is a sum of $n-1$ independent $(h+1)$-dimensional vector-valued random variables) lands in a small origin-centered rectangular box, which we denote $\mathcal{B}_J$, in $\mathbb{R}^{h+1}$. Here h is a value which is chosen to be significantly less than ℓ, but still "large enough" that it suffices for steps (2) and (3) being described here; we will use the remaining $\ell - h$ matching moments later in the argument, in step (4).

(3): Pruning arbitrary query sets. We may associate each multi-index J that has $|J| = h+1$ with a multiset $\mathcal{A}$ of size $h+1$ drawn from the d-element query set. For simplicity, in the following informal discussion let us assume that every element in $\mathcal{A}$ occurs with multiplicity exactly 1 (this is indeed the case for most multisets of $[d]$ of size $h+1$; recall that h is a fixed integer whereas d should be thought of as $n^{\Theta(1)}$).

A major difficulty is that for some query sets, it may be the case that for many outcomes of $\mathcal{A}$ (equivalently, J) it is simply impossible to give a strong upper bound on the probability that $\mathbf{R}_{-i}|_J$ lands in the small rectangular box $\mathcal{B}_J$. For example, this can be the case if many query strings lie very close to each other (see the discussion in the last two paragraphs of Section 4 for an extreme instance of this phenomenon). However, if there are two query strings which are very close to each other, then with very high probability over the outcomes of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ the responses to the two queries for $\boldsymbol{f} \sim \mathcal{D}_{yes}$ will be the same, and likewise for $\boldsymbol{f} \sim \mathcal{D}_{no}$. This should effectively allow us to "prune" the query set and reduce its size by 1. On the other hand, there is a non-zero probability that two close but distinct query strings have different answers, and it is intuitively clear that this probability increases with the distance between the query strings; thus any such pruning must be done with care.

There is indeed a delicate balance between these two competing demands (pruning queries to eliminate cases where the desired anti-concentration probability cannot be effectively bounded, and introducing errors by pruning queries). In Section 5, we perform a careful tradeoff between these demands, and show that *any* query set can be pruned (at the cost of a small acceptable increase in error) in a useful way. The exact condition we require of our pruned query sets is rather involved so we defer a precise statement of it to Section 5, but roughly speaking, it involves having only a small fraction of all queries lie too close to the linear span of any small set of query strings (see Definition 8 for a precise definition). We show in later sections that this condition, which we refer to as a query set being "scattered," lets us establish the desired anti-concentration mentioned above. We note that our pruning procedure heavily uses the fact that query strings are elements of the (scaled) Boolean hypercube; this enables us to establish and employ some useful facts which, roughly speaking, exploit some geometrical incompatibility between linear subspaces of $\mathbb{R}^n$ and the Boolean hypercube.

(4): Handling scattered query sets. A careful analysis of scattered query sets lets us show that if $\mathbf{G}$ is a $(h+1)$-dimensional *Gaussian* whose mean and covariance matrix match those of $\mathbf{R}_{-i}|_J$, then $\mathbf{G}$ satisfies the desired anti-concentration bound. To show that the above-mentioned random variable $\mathbf{R}_{-i}|_J$ — which is *not* a Gaussian — also satisfies this anti-concentration bound, we exploit the fact that $\boldsymbol{u}$'s first ℓ moments match those of $\boldsymbol{v}$, *which in turn match the first ℓ moments of a variance-1 Gaussian* (recall ingredient (1), "Suitable choice of distributions," above). (This is where we use the "remaining" $\ell - h$ matching moments for $\boldsymbol{u}$ and $\boldsymbol{v}$ alluded to earlier.) This lets us adapt the simple argument that was employed for the "warm-up" result to establish that the two distributions $\mathbf{R}_{-i}|_J$ and $\mathbf{G}$ must both put almost the same amount of weight as each other on the box $\mathcal{B}_J$ mentioned above; since $\mathbf{G}$ is anti-concentrated on this box, it follows that $\mathbf{R}_{-i}|_J$ must have similar anti-concentration.

The fact that $\boldsymbol{v}$ matches the first ℓ moments of a Gaussian is crucial here, since otherwise the penalty incurred for the "smoothing" term in Lindeberg's method (the final term on the RHS of the inequality of Proposition 4.2) would be prohibitively large. By having $\boldsymbol{v}$ match the moments of a Gaussian, though, we can use the aforementioned analysis (showing that the Gaussian $\mathbf{G}$ satisfies the desired anti-concentration bound) in order to give a strong upper bound on this smoothing penalty, and thereby obtain our overall desired result.

1.4 Organization.

In Section 3 we establish the existence of real random variables $\boldsymbol{u}, \boldsymbol{v}$ with the "matching moments" property that we require. Section 4 proves an $\Omega(n^{1/4-c})$ lower bound for monotonicity testing via a "vanilla" application of Lindeberg's method using higher-order matching moments, and outlines our approach for going beyond $n^{1/4-c}$. Section 5 describes our pruning procedure that transforms an arbitrary query set into a "scattered" query set. In Sections 6 we give our lower bound for scattered query sets, and finally in Section 7 we put together the pieces and complete the proof of Theorem 1.

2 Preliminaries

Given $n \in \mathbb{N}$, we let $[n]$ denote $\{1, \ldots, n\}$, and given $a \le b \in \mathbb{N}$ we let $[a : b]$ denote $\{a, \ldots, b\}$. We use lowercase letters to denote real numbers, uppercase letters to denote vectors of real numbers, and boldface (e.g. $\boldsymbol{x}$ and $\mathbf{X}$) to denote random variables. We will also use calligraphic letters like $\mathcal{X}$ to denote sets or multisets of vectors.

For $X \in \mathbb{R}^n$ we use $B_{\ell_2}(X, r)$ to denote $\{Y \in \mathbb{R}^n : \|X - Y\|_2 \le r\}$, the Euclidean ball of radius r centered at X. For $Y, Z \in \{\pm 1/\sqrt{n}\}^n$, the Hamming distance $d_{\mathrm{Ham}}(Y, Z)$ is defined as the number of coordinates where Y and Z differ.

Recall that a k-variable Boolean function f is a *linear threshold function* (LTF) if there exist real values $w_1, \ldots, w_k, \theta$ such that $f(x) = \mathrm{sign}(\sum_{i=1}^k w_i x_i - \theta)$.

We will require the following useful fact on the number of distinct LTFs over the k-dimensional Boolean hypercube (where we view two LTFs as distinct if they differ as Boolean functions):

Fact 2.1. [Sch50] *The total number of distinct LTFs over $\{-1, 1\}^k$ is upper bounded by 2^{k^2}.*

Given a d-dimensional multi-index $J = (J_1, \ldots, J_d) \in \mathbb{N}^d$, we write $|J|$ to denote $J_1 + \cdots + J_d$ and $J!$ to denote $J_1!J_2! \cdots J_d!$. We write $\mathrm{supp}(J)$ to denote the set $\{i \in [d] \colon J_i \ne 0\}$, and $\#J$ to denote $|\mathrm{supp}(J)|$. (Note that $\#J \le |J|$.) Given $X \in \mathbb{R}^d$ we write X^J to denote $\prod_{i=1}^d (X_i)^{J_i}$, and $X|_J \in \mathbb{R}^{\#J}$ to denote the projection of X onto the coordinates in $\mathrm{supp}(J)$. For $f : \mathbb{R}^d \to \mathbb{R}$, we write $f^{(J)}$ to denote the J-th derivative, i.e.

$$f^{(J)} = \frac{\partial^{J_1 + \cdots + J_d} f}{\partial x_1^{J_1} \cdots \partial x_d^{J_d}}.$$

We will use the standard multivariate Taylor expansion:

Fact 2.2 (Multivariate Taylor expansion)**.** *Given a smooth function $f : \mathbb{R}^d \to \mathbb{R}$ and $k \in \mathbb{N}$,*

$$f(X+\Delta) = \sum_{|J|\le k} \frac{f^{(J)}(X)}{J!} \cdot \Delta^J + (k+1) \sum_{|J|=k+1} \left(\frac{\Delta^J}{J!} \mathbf{E}\left[(1-\boldsymbol{\tau})^k f^{(J)}(X+\boldsymbol{\tau}\Delta)\right] \right),$$

for $X, \Delta \in \mathbb{R}^d$, where $\boldsymbol{\tau}$ is a random variable uniformly distributed on the interval $[0,1]$.

We recall the standard Berry–Esséen theorem (see for example, [Fel68]) for sums of independent real random variables:

Theorem 3 (Berry–Esséen)**.** *Let $\boldsymbol{s} = \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_n$, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent real-valued random variables with $\mathbf{E}[\boldsymbol{x}_j] = \mu_j$ and $\mathbf{Var}[\boldsymbol{x}_j] = \sigma_j^2$, and suppose that $|\boldsymbol{x}_j - \mathbf{E}[\boldsymbol{x}_j]| \le \tau$ with probability 1 for all $j \in [n]$. Let $\boldsymbol{g}$ denote a Gaussian random variable with mean $\sum_{j=1}^n \mu_j$ and variance $\sum_{j=1}^n \sigma_j^2$, matching those of $\boldsymbol{s}$. Then for all $\theta \in \mathbb{R}$, we have*

$$\left| \mathbf{Pr}[\boldsymbol{s} \le \theta] - \mathbf{Pr}[\boldsymbol{g} \le \theta] \right| \le \frac{O(\tau)}{\sqrt{\sum_{j=1}^n \sigma_j^2}}.$$

3 The $\mathcal{D}_{yes}$ and $\mathcal{D}_{no}$ distributions

The main results of this section are the following:

Proposition 3.1 (The "yes" random variable)**.** *Given an odd $\ell \in \mathbb{N}$, there exists a value $\mu = \mu(\ell) > 0$ and a real random variable $\boldsymbol{u}$ such that*

1. *$\boldsymbol{u}$ is supported on at most ℓ nonnegative real values; and*
2. *$\mathbf{E}[\boldsymbol{u}^k] = \mathbf{E}[\mathcal{N}(\mu,1)^k]$ for all $k \in [\ell]$.*

Proposition 3.2 (The "no" random variable)**.** *Given $\mu > 0$ and $\ell \in \mathbb{N}$, there exists a real random variable $\boldsymbol{v}$ such that*

1. *$\boldsymbol{v}$ is supported on at most $\ell + 1$ real values, with $\mathbf{Pr}[\boldsymbol{v} < 0] > 0$; and*
2. *$\mathbf{E}[\boldsymbol{v}^k] = \mathbf{E}[\mathcal{N}(\mu,1)^k]$ for all $k \in [\ell]$.*

Note the difference between these two propositions: the first requires $\boldsymbol{u}$ to be supported entirely on nonnegative values, while the second requires $\boldsymbol{v}$ to put nonzero weight on some negative value.

Let $c > 0$ (this should be viewed as the "c" of Theorem 1), and let $h = h(c) \in \mathbb{N}$ denote an odd constant that depends on c only. Let $\boldsymbol{u}$ and $\boldsymbol{v}$ denote random variables given in Proposition 3.1 and 3.2, respectively, with $\ell = h^3$ and $\mu = \mu(\ell)$. As discussed in Section 1.2 the "yes" distribution $\mathcal{D}_{yes}$ of Boolean functions is given by (1) and the "no" distribution by (2), where each $\boldsymbol{u}_i$ is i.i.d. distributed according to $\boldsymbol{u}$ and likewise for the $\boldsymbol{v}_i$'s and $\boldsymbol{v}$. It is clear that $\boldsymbol{u}$ and $\boldsymbol{v}$ have matching first ℓ-th moments, and Proposition 3.1 ensures that every function in the support of $\mathcal{D}_{yes}$ is monotone. In Appendix B we show that with probability $1 - o_n(1)$, a random LTF drawn from $\mathcal{D}_{no}$ is κ-far from all monotone Boolean functions, where $\kappa > 0$ depends on the values of μ and ℓ and hence on c only.

Thus the above two Propositions 3.1 and 3.2 are enough for the basic framework of Yao's method to go through and establish our lower bound, once we show that equation (4) holds. We do this in the rest of the paper, but first in the remainder of this section we prove Propositions 3.1 and 3.2. We start with the easier Proposition 3.2.

3.1 Proof of Proposition 3.2

For each $x \in \mathbb{R}$, let $A(x)$ denote the $(\ell+1)$-dimensional real vector defined by

$$A(x)_k = \begin{cases} x^k & \text{for } k \in [\ell], \\ \mathbf{1}[x < 0] & \text{for } k = \ell + 1. \end{cases}$$

Consider the vector $P \in \mathbb{R}^{\ell+1}$ defined by

$$P_k = \begin{cases} \mathbf{E}[\mathcal{N}(\mu, 1)^k] & \text{for } k \in [\ell], \\ \mathbf{Pr}[\mathcal{N}(\mu, 1) < 0] & \text{for } k = \ell + 1. \end{cases}$$

Since $P = \mathbf{E}_{\boldsymbol{x} \sim \mathcal{N}(\mu,1)}[A(\boldsymbol{x})]$ the point P is in the convex hull of the point set $V := \{A(x) : x \in \mathbb{R}\} \subset \mathbb{R}^{\ell+1}$. Hence Carathéodory's theorem implies that P lies in the convex hull of some $(\ell+1)$-point subset of V, i.e. there exist $x_1, \ldots, x_{\ell+1} \in \mathbb{R}$ and $0 \leq \mu_1, \ldots, \mu_{\ell+1}$ with $\sum_j \mu_j = 1$ such that

$$P = \sum_{j=1}^{\ell+1} \mu_j A(x_j).$$

The desired random variable $\boldsymbol{v}$ is defined by $\mathbf{Pr}[\boldsymbol{v} = x_j] = \mu_j$. It is clear that $\boldsymbol{v}$ is supported on at most $\ell + 1$ real values, and $\boldsymbol{v}$ satisfies the desired moment condition since

$$\mathbf{E}[\mathcal{N}(\mu, 1)^k] = P_k = \sum_{j=1}^{\ell+1} \mu_j x_j^k = \mathbf{E}[\boldsymbol{v}^k], \quad \text{for all } k \in [\ell].$$

Finally, since

$$0 < \mathbf{Pr}[\mathcal{N}(\mu, 1) < 0] = P_{\ell+1} = \sum_{j=1}^{\ell+1} \mu_j \mathbf{1}[x_j < 0],$$

it must be the case that $\mathbf{Pr}[\boldsymbol{v} < 0] > 0$, and the proposition is proved. □

3.2 Proof of Proposition 3.1

Given a sequence of real numbers $(m_1, \ldots, m_\ell)$, consider the problem of deciding whether there exists a real random variable $\boldsymbol{x}$ such that $\mathbf{E}[\boldsymbol{x}^i] = m_i$ for $i = 1, \ldots, \ell$. This is a form of the well-studied classical moment problem, and a complete solution has been given in terms of the moment vector lying in a particular well-specified cone. More precisely, the following can be found in [Akh65].

Theorem 4. *Let $\overline{m} = (m_1, m_2, \ldots, m_{2n})$. There is a random variable $\boldsymbol{x}$ supported on $\mathbb{R}$ such that $\mathbf{E}[\boldsymbol{x}^i] = m_i$ for $i = 1, \ldots, 2n$ if and only if*

$$A_{\mathbb{R}}(\overline{m}) = \begin{pmatrix} 1 & m_1 & \ldots & m_n \\ m_1 & m_2 & \ldots & m_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_n & m_{n+1} & \ldots & m_{2n} \end{pmatrix} \succeq 0.$$

The corresponding problem when the support of the desired random variable $\boldsymbol{x}$ is restricted to non-negative reals is also completely solved, by the following result:

Theorem 5. *Let $\overline{m} = (m_1, m_2, \ldots, m_{2n+1})$. There is a random variable $\boldsymbol{x}$ supported on $[0, \infty)$ such that $\mathbf{E}[\boldsymbol{x}^i] = m_i$ for $i = 1, \ldots, 2n+1$ if and only if*

$$A_{\mathbb{R}}(\overline{m}) = \begin{pmatrix} 1 & m_1 & \ldots & m_n \\ m_1 & m_2 & \ldots & m_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_n & m_{n+1} & \ldots & m_{2n} \end{pmatrix} \succeq 0 \quad \textit{and} \quad A^{+}_{\mathbb{R}}(\overline{m}) = \begin{pmatrix} m_1 & m_2 & \ldots & m_{n+1} \\ m_2 & m_3 & \ldots & m_{n+2} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n+1} & m_{n+2} & \ldots & m_{2n+1} \end{pmatrix} \succeq 0.$$

We use the above results by taking each m_ℓ to equal $\mathbf{E}[\boldsymbol{z}_\mu^\ell]$, where $\boldsymbol{z}_\mu$ is distributed according to $\mathcal{N}(\mu, 1)$ (so $m_\ell = m_\ell(\mu)$ is a function of μ). Our aim is to show that $\mu = \mu(\ell)$ can be taken to be a sufficiently large integer (in terms of ℓ) such that

$$A_{\mathbb{R}}(\overline{m}(\mu)) = \begin{pmatrix} 1 & m_1 & \ldots & m_\ell \\ m_1 & m_2 & \ldots & m_{\ell+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_\ell & m_{\ell+1} & \ldots & m_{2\ell} \end{pmatrix} \succeq 0 \tag{5}$$

$$\text{and} \qquad A^{+}_{\mathbb{R}}(\overline{m}(\mu)) = \begin{pmatrix} m_1 & m_2 & \ldots & m_{\ell+1} \\ m_2 & m_3 & \ldots & m_{\ell+2} \\ \vdots & \vdots & \ddots & \vdots \\ m_{\ell+1} & m_{\ell+2} & \ldots & m_{2\ell+1} \end{pmatrix} \succeq 0. \tag{6}$$

If these two conditions hold, then we may take $\boldsymbol{u}'$ to be the nonnegative random variable $\boldsymbol{x}$ whose existence is asserted by Theorem 5. Applying Carathéodory's theorem, an argument similar to the proof of Proposition 3.2 allows us to obtain from $\boldsymbol{u}'$ a nonnegative random variable $\boldsymbol{u}$ with support size at most ℓ and the same moments. This will finish the proof of Proposition 3.1.

As the Gaussian $\boldsymbol{z}_\mu = \mathcal{N}(\mu, 1)$ is itself a random variable such that $\mathbf{E}[\boldsymbol{z}_\mu^i] = m_i$, for $i = 1, \ldots, 2\ell$, Theorem 4 implies that (5) holds; thus, it remains to prove (6).

Observe that $m_k = m_k(\mu)$ is a degree-k polynomial in μ. We define

$$P_{\det}(\mu) = \det(A^{+}_{\mathbb{R}}(\overline{m})).$$

Our argument requires the following four technical claims.

Claim 3.3. *There exists μ_0 such that $A^{+}_{\mathbb{R}}(\overline{m}(\mu))$ is non-singular for all $\mu \in \mathbb{R} \setminus [-\mu_0, \mu_0]$.*

Proof. Observe that $P_{\det}(\mu)$ is a degree-T polynomial in μ with $T < (\ell + 1)(2\ell + 1)$. Thus, given that $P_{\det}(\mu)$ is not the identically-0 polynomial, if μ_0 is set to be the largest magnitude of the zero of this polynomial, we get the claim.

To see that $P_{\det}(\mu)$ is not identically zero, we consider the matrix $A^{+}_{\mathbb{R}}(\overline{m}(0))$ obtained by taking $\mu = 0$. This matrix has (i, j)th entry $\mathbf{E}[\mathcal{N}(0, 1)^{i+j-1}]$, which is 0 if $i + j$ is even. By inspection of this matrix we see that for odd $\ell > 1$, we have

$$A^{+}_{\mathbb{R}}(\overline{m}(0)) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \otimes B^{(\ell)},$$

where $B^{(\ell)}$ is the square matrix of dimension $(\ell+1)/2$ that has (i,j)th entry $\mathbf{E}[\mathcal{N}(0,1)^{2(i+j-1)}]$. It follows that $P_{\det}(0) = (-1)^{(\ell+1)/2}\det(B^{(\ell)})^2$. Recalling that for k even we have

$$\mathbf{E}[\mathcal{N}(0,1)^k] = (k-1)!! = 1 \cdot 3 \cdot 5 \cdots (k-1),$$

the product of the odd numbers from 1 to $k-1$, it can be shown that

$$\det(B^{(\ell)}) = \prod_{j \text{ odd}, 1 \le j \le \ell} j!.$$

We include a proof of this fact in Appendix C. The lemma then follows. □

Claim 3.4. *If $\mu > \mu_0$ is an integer, then $|\det(A^+_{\mathbb{R}}(\overline{m}(\mu)))| \ge 1$.*

Proof. The fact that each raw moment $\mathbf{E}[\mathcal{N}(0,1)^k]$ of the Gaussian is an integer easily implies that $m_k(\mu) = \mathbf{E}[(\mathcal{N}(0,1)+\mu)^k]$ is a polynomial with integer coefficients, and hence $P_{\det}(\mu)$ has integer coefficients as well. Together with Claim 3.3 this gives the claim. □

Claim 3.5. *For all integer $\mu > 0$, we have that the largest singular value of $A^+_{\mathbb{R}}(\overline{m}(\mu))$, denoted $\sigma_{\max}(A^+_{\mathbb{R}}(\overline{m}(\mu)))$, satisfies $\sigma_{\max}(A^+_{\mathbb{R}}(\overline{m}(\mu))) \le (\ell+1)^2 \cdot (2\ell+1)! \cdot \mu^{2\ell+1}$.*

Proof. We have $\sigma_{\max}(A^+_{\mathbb{R}}(\overline{m}(\mu))) \le \|A^+_{\mathbb{R}}(\overline{m}(\mu))\|_F$. We use the following simple upper bound on the kth moment of the mean-μ, variance-1 Gaussian $\mathcal{N}(\mu,1)$:

$$\mathbf{E}[\mathcal{N}(\mu,1)^k] = \mathbf{E}[(\mathcal{N}(0,1)+\mu)^k] = \sum_{j=0}^{\lfloor k/2 \rfloor} \binom{k}{2j} (2j-1)!! \cdot \mu^{k-2j} < (\lfloor k/2 \rfloor + 1) \cdot k! \cdot \mu^k.$$

The claim follows by combining the two inequalities. □

Claim 3.6. *For all integer $\mu > \mu_0$, we have that the smallest singular value of $A^+_{\mathbb{R}}(\overline{m}(\mu))$, denoted $\sigma_{\min}(A^+_{\mathbb{R}}(\overline{m}(\mu)))$, satisfies*

$$\sigma_{\min}(A^+_{\mathbb{R}}(\overline{m}(\mu))) \ge \frac{1}{(2\ell+1)^{\ell(2\ell+2)} \cdot \mu^{\ell(2\ell+1)}}.$$

Proof. We just use the simple inequality that for any symmetric matrix $A \in \mathbb{R}^{(\ell+1)\times(\ell+1)}$:

$$\sigma_{\min}(A) \ge \frac{|\det(A)|}{\sigma_{\max}(A)^\ell} \ge \frac{1}{(2\ell+1)^{\ell(2\ell+2)} \cdot \mu^{\ell(2\ell+1)}}.$$

The last inequality uses Claim 3.4 and Claim 3.5. □

With these technical claims in hand, we proceed to establish (6). In case $A^+_{\mathbb{R}}(\overline{m}(\mu)) \succeq 0$ for some integer $\mu > \mu_0$, we are done. Otherwise, towards a contradiction, let us assume that $A^+_{\mathbb{R}}(\overline{m}(\mu))$ has a negative eigenvalue for every integer $\mu > \mu_0$. This means that for every integer $\mu > \mu_0$, we have

$$\lambda_{\min}(A^+_{\mathbb{R}}(\overline{m}(\mu))) \le -\frac{1}{(2\ell+1)^{\ell(2\ell+2)} \cdot \mu^{\ell(2\ell+1)}}. \tag{7}$$

Let us define the random variable $\boldsymbol{z}'_\mu$ to be distributed as $\boldsymbol{z}'_\mu = \max\{\boldsymbol{z}_\mu, 0\}$. It is straightforward to upper bound the difference in moments between the random variables $\boldsymbol{z}_\mu$ and $\boldsymbol{z}'_\mu$:

Claim 3.7. *For $k \in \mathbb{N}$, we have* $\left|\mathbf{E}[\boldsymbol{z}'^k_\mu] - \mathbf{E}[\boldsymbol{z}^k_\mu]\right| \le e^{-\frac{\mu^2}{2}} \cdot (k-1)!!$.

Proof. We have

$$\begin{aligned}\left|\mathbf{E}[\boldsymbol{z}'^k_\mu] - \mathbf{E}[\boldsymbol{z}^k_\mu]\right| &= \int_{y=-\infty}^{0} \frac{1}{\sqrt{2\pi}} \cdot |y|^k \cdot e^{-\frac{(y-\mu)^2}{2}} dy \\ &\le e^{-\frac{\mu^2}{2}} \int_{y=-\infty}^{0} \frac{1}{\sqrt{2\pi}} \cdot |y|^k \cdot e^{-\frac{y^2}{2}} dy \\ &= e^{-\frac{\mu^2}{2}} \cdot (k-1)!!\end{aligned}$$

where the last line used the fact that the k-th absolute moment of $\mathcal{N}(0,1)$ is at most $(k-1)!!$. □

To conclude the proof, let $m'_i(\mu) = \mathbf{E}[\boldsymbol{z}'^i_\mu]$. Then by Claim 3.7, we have (where $\|M\|_\infty$ denotes the entrywise maximum absolute value of any element of the matrix M)

$$\|A^+_{\mathbb{R}}(\overline{m}(\mu)) - A^+_{\mathbb{R}}(\overline{m}'(\mu))\|_\infty \le e^{-\frac{\mu^2}{2}} \cdot (2\ell)!!. \tag{8}$$

Let $u \in \mathbb{R}^{\ell+1}$ be the unit vector minimizing $u^T A^+_{\mathbb{R}}(\overline{m}(\mu))u$, so $\lambda_{\min}(A^+_{\mathbb{R}}(\overline{m}(\mu))) = u^T A^+_{\mathbb{R}}(\overline{m}(\mu))u$. We have

$$\begin{aligned}\lambda_{\min}(A^+_{\mathbb{R}}(\overline{m}'(\mu))) - \lambda_{\min}(A^+_{\mathbb{R}}(\overline{m}(\mu))) &\le u^T(A^+_{\mathbb{R}}(\overline{m}'(\mu)) - A^+_{\mathbb{R}}(\overline{m}(\mu)))u \\ &\le (\ell+1)\|A^+_{\mathbb{R}}(\overline{m}'(\mu)) - A^+_{\mathbb{R}}(\overline{m}(\mu))\|_\infty,\end{aligned}$$

so by (7) and (8) we get that

$$\lambda_{\min}(A^+_{\mathbb{R}}(\overline{m}'(\mu))) \le -\frac{1}{(2\ell+1)^{\ell(2\ell+2)} \cdot \mu^{\ell(2\ell+1)}} + (\ell+1) \cdot e^{-\frac{\mu^2}{2}} \cdot (2\ell)!!.$$

By choosing $\mu = \mu(\ell)$ to be a sufficiently large integer relative to ℓ we can make $\lambda_{\min}(A^+_{\mathbb{R}}(\overline{m'}(\mu))) < 0$, which is a contradiction with Theorem 5 and the fact that $\boldsymbol{z}'_\mu$ is supported on $[0,\infty)$.

4 Warmup: an $\Omega(n^{1/4-c})$ lower bound via higher moments

In this section we give the basic Lindeberg argument using matching higher moments. This immediately improves the $\tilde{\Omega}(n^{1/5})$ lower bound in [CST14] to $\Omega(n^{1/4-c})$ for any constant $c > 0$ (see the end of Section 4.2) and is the first step in our proof of the $\Omega(n^{1/2-c})$ lower bound. The main technical ingredient is a higher-moments extension of the [GOWZ10] multidimensional CLT, which we use in place of the [VV11] multidimensional CLT used in [CST14].

4.1 A useful mollifier

We begin with a couple of basic propositions:

Proposition 4.1. *Let $\mathcal{A}, \mathcal{A}_{in} \subseteq \mathbb{R}^d$ where $\mathcal{A}_{in} \subseteq \mathcal{A}$. Let $\Psi_{in} : \mathbb{R}^d \to [0,1]$ be a function satisfying $\Psi_{in}(X) = 1$ for all $X \in \mathcal{A}_{in}$ and $\Psi_{in}(X) = 0$ for all $X \notin \mathcal{A}$. Then for all random variables $\mathbf{S}, \mathbf{T}$:*

$$\left|\mathbf{Pr}[\mathbf{S} \in \mathcal{A}] - \mathbf{Pr}[\mathbf{T} \in \mathcal{A}]\right| \le \left|\mathbf{E}[\Psi_{in}(\mathbf{S})] - \mathbf{E}[\Psi_{in}(\mathbf{T})]\right| + \max\left\{\mathbf{Pr}[\mathbf{S} \in \mathcal{A} \setminus \mathcal{A}_{in}], \mathbf{Pr}[\mathbf{T} \in \mathcal{A} \setminus \mathcal{A}_{in}]\right\}.$$

Proof. Observe that $\mathbf{Pr}[\mathbf{S} \in \mathcal{A}] \geq \mathbf{E}[\Psi_{in}(\mathbf{S})]$ and $\mathbf{Pr}[\mathbf{S} \in \mathcal{A}] \leq \mathbf{E}[\Psi_{in}(\mathbf{S})] + \mathbf{Pr}[\mathbf{S} \in \mathcal{A} \setminus \mathcal{A}_{in}]$, and likewise for $\mathbf{T}$. As a result, we have

$$\begin{aligned}\mathbf{Pr}[\mathbf{S} \in \mathcal{A}] - \mathbf{Pr}[\mathbf{T} \in \mathcal{A}] &\leq \mathbf{E}[\Psi_{in}(\mathbf{S})] + \mathbf{Pr}[\mathbf{S} \in \mathcal{A} \setminus \mathcal{A}_{in}] - \mathbf{E}[\Psi_{in}(\mathbf{T})], \quad \text{and} \\ \mathbf{Pr}[\mathbf{S} \in \mathcal{A}] - \mathbf{Pr}[\mathbf{T} \in \mathcal{A}] &\geq \mathbf{E}[\Psi_{in}(\mathbf{S})] - \mathbf{Pr}[\mathbf{T} \in \mathcal{A} \setminus \mathcal{A}_{in}] - \mathbf{E}[\Psi_{in}(\mathbf{T})].\end{aligned}$$

Combining these, we have the proposition. □

We will use the following lemma of Bentkus [Ben03]. For completeness we include its proof.

Proposition 4.2 (Lemma 2.1 of [Ben03])**.** *Let $\mathcal{A}, \mathcal{A}_{in}, \mathcal{A}_{out} \subseteq \mathbb{R}^d$ with $\mathcal{A}_{in} \subseteq \mathcal{A} \subseteq \mathcal{A}_{out}$. Let $\Psi_{in} : \mathbb{R}^d \to [0,1]$ be a function where $\Psi_{in}(X) = 1$ for all $X \in \mathcal{A}_{in}$ and $\Psi_{in}(X) = 0$ for all $X \notin \mathcal{A}$, and let $\Psi_{out} : \mathbb{R}^d \to [0,1]$ be a function where $\Psi_{out}(X) = 1$ for all $X \in \mathcal{A}$ and $\Psi_{out}(X) = 0$ for all $X \notin \mathcal{A}_{out}$. Then for all random variables $\mathbf{S}, \mathbf{T}$ we have that*

$$\begin{aligned}\big|\mathbf{Pr}[\mathbf{S} \in \mathcal{A}] - \mathbf{Pr}[\mathbf{T} \in \mathcal{A}]\big| \leq \max\big\{&\big|\mathbf{E}[\Psi_{in}(\mathbf{S})] - \mathbf{E}[\Psi_{in}(\mathbf{T})]\big|, \big|\mathbf{E}[\Psi_{out}(\mathbf{S})] - \mathbf{E}[\Psi_{out}(\mathbf{T})]\big|\big\} \\ &+ \max\big\{\mathbf{Pr}[\mathbf{T} \in \mathcal{A}_{out} \setminus \mathcal{A}], \mathbf{Pr}[\mathbf{T} \in \mathcal{A} \setminus \mathcal{A}_{in}]\big\}.\end{aligned}$$

Proof. For the case when $\mathbf{Pr}[\mathbf{S} \in \mathcal{A}] \geq \mathbf{Pr}[\mathbf{T} \in \mathcal{A}]$, we have

$$\mathbf{Pr}[\mathbf{S} \in \mathcal{A}] - \mathbf{Pr}[\mathbf{T} \in \mathcal{A}] \leq \mathbf{E}[\Psi_{out}(\mathbf{S})] - \mathbf{E}[\Psi_{out}(\mathbf{T})] + \mathbf{E}[\Psi_{out}(\mathbf{T})] - \mathbf{Pr}[\mathbf{T} \in \mathcal{A}].$$

The proposition follows from $\mathbf{E}[\Psi_{out}(\mathbf{T})] \leq \mathbf{Pr}[\mathbf{T} \in \mathcal{A}_{out}] = \mathbf{Pr}[\mathbf{T} \in \mathcal{A}] + \mathbf{Pr}[\mathbf{T} \in \mathcal{A}_{out} \setminus \mathcal{A}]$.

Now for the case when $\mathbf{Pr}[\mathbf{T} \in \mathcal{A}] > \mathbf{Pr}[\mathbf{S} \in \mathcal{A}]$, we have

$$\mathbf{Pr}[\mathbf{T} \in \mathcal{A}] - \mathbf{Pr}[\mathbf{S} \in \mathcal{A}] \leq \mathbf{Pr}[\mathbf{T} \in \mathcal{A}] - \mathbf{E}[\Psi_{in}(\mathbf{T})] + \mathbf{E}[\Psi_{in}(\mathbf{T})] - \mathbf{E}[\Psi_{in}(\mathbf{S})].$$

The proposition follows from $\mathbf{Pr}[\mathbf{T} \in \mathcal{A}] \leq \mathbf{E}[\Psi_{in}(\mathbf{T})] + \mathbf{Pr}[\mathbf{T} \in \mathcal{A} \setminus \mathcal{A}_{in}]$. □

For the rest of this section, we need to define a sufficiently fast growing function $\alpha : \mathbb{N} \to \mathbb{R}^+$:

$$\alpha(k) = 2e \cdot (64)^k \cdot k! \cdot k^{2k+2}.$$

As is standard in Lindeberg-type arguments, our proof will employ a "mollifier", i.e. a smooth function which approximates the indicator function of a set. In this work we require a specific mollifier whose properties are tailored to our sets of interest (unions of orthants) and are given in the following proposition.

Proposition 4.3 (Product mollifier)**.** *Let $\mathcal{O}$ be a union of orthants in $\mathbb{R}^d$. For all $\varepsilon > 0$, there exists a smooth function $\Psi_{\mathcal{O}} : \mathbb{R}^d \to [0,1]$ with the following properties:*

1. *$\Psi_{\mathcal{O}}(X) = 0$ for all $X \notin \mathcal{O}$.*
2. *$\Psi_{\mathcal{O}}(X) = 1$ for all $X \in \mathcal{O}$ with $\min_i\{|X_i|\} \geq \varepsilon$.*
3. *For any multi-index $J \in \mathbb{N}^d$ such that $|J| = k$, $\|\Psi_{\mathcal{O}}^{(J)}\|_\infty \leq \alpha(k) \cdot (1/\varepsilon)^k$.*
4. *For any $J \in \mathbb{N}^d$, $\Psi_{\mathcal{O}}^{(J)}(X) \neq 0$ only if $X \in \mathcal{O}$ and $|X_i| \leq \varepsilon$ for all i such that $J_i \neq 0$. Equivalently, $\Psi_{\mathcal{O}}^{(J)}(X) \neq 0$ only if $X \in \mathcal{O}$ and $\|X|_J\|_\infty \leq \varepsilon$.*

We note that while properties (1)–(3) above are entirely standard, we are not aware of previous work which uses property (4). As we shall see this property is particularly useful in our setting where the goal is to bound the union-of-orthants distance d_{UO}. To prove Proposition 4.3, we first prove the following easier version of it.

Proposition 4.4. *Let $\mathcal{O}_1$ be an orthant in $\mathbb{R}^d$. For all $\varepsilon > 0$, there exists a smooth function $\Psi_{\mathcal{O}_1} : \mathbb{R}^d \to [0,1]$ with the following properties:*

1. *$\Psi_{\mathcal{O}_1}(X) = 0$ for all $X \notin \mathcal{O}_1$.*
2. *$\Psi_{\mathcal{O}_1}(X) = 1$ for all $X \in \mathcal{O}_1$ with $\min_i\{|X_i|\} \geq \varepsilon$.*
3. *For any multi-index $J \in \mathbb{N}^d$ such that $|J| = k$, $\|\Psi^{(J)}_{\mathcal{O}_1}\|_\infty \leq \alpha(k) \cdot (1/\varepsilon)^k$.*
4. *For any $J \in \mathbb{N}^d$, $\Psi^{(J)}_{\mathcal{O}_1}(X) \neq 0$ only if $X \in \mathcal{O}_1$ and $|X_i| \leq \varepsilon$ for all i such that $J_i \neq 0$. Equivalently, $\Psi^{(J)}_{\mathcal{O}_1}(X) \neq 0$ only if $X \in \mathcal{O}_1$ and $\|X|_J\|_\infty \leq \varepsilon$*

We first see how Proposition 4.4 can be used to prove Proposition 4.3.

Proof of Proposition 4.3. Let $\mathcal{O} = \cup_{i\in[m]}\mathcal{O}_i$, where the $\mathcal{O}_i$'s are (disjoint) orthants in $\mathbb{R}^d$. Let $\Psi_{\mathcal{O}_i}$ be the function obtained by applying Proposition 4.4 to the orthant $\mathcal{O}_i$, and let $\Psi_{\mathcal{O}} = \sum_{i\in[m]} \Psi_{\mathcal{O}_i}$. We claim that $\Psi_{\mathcal{O}}$ satisfies the required conditions. Properties (1) and (2) follow immediately from the corresponding properties of $\Psi_{\mathcal{O}_i}$.

For properties (3) and (4), observe that from Proposition 4.4, for each $i \in [m]$, $\Psi^{(J)}_{\mathcal{O}_i}(X) = 0$ if $X \notin \mathcal{O}_i$. Also by the definition of $\Psi_{\mathcal{O}}$ we have

$$\Psi^{(J)}_{\mathcal{O}}(X) = \sum_{i\in[m]} \Psi^{(J)}_{\mathcal{O}_i}(X).$$

Since the $\mathcal{O}_i$'s are pairwise disjoint, we have that for any $X \in \mathbb{R}^d$, at most one of the summands is non-zero. Thus, using property (3) from Proposition 4.4, we get property (3) for $\Psi_{\mathcal{O}}$. Using the same reasoning and property (4) from Proposition 4.4, we get property (4) for $\Psi_{\mathcal{O}}$. □

To prove Proposition 4.4 we will need the following *one-dimensional* version of $\Psi_{\mathcal{O}_1}$. This is the standard mollifier construction in one-dimension. For completeness we prove it in Appendix A.

Claim 4.5. *For all $\varepsilon > 0$, there exists a smooth function $\Phi_\varepsilon : \mathbb{R} \to [0,1]$ which satisfies:*

1. *If $x < 0$, then $\Phi_\varepsilon(x) = 0$.*
2. *If $x > \varepsilon$, then $\Phi_\varepsilon(x) = 1$.*
3. *$\|\Phi^{(k)}_\varepsilon\|_\infty \leq \alpha(k) \cdot (1/\varepsilon)^k$.*

Proof of Proposition 4.4. Without loss of generality we may assume our orthant $\mathcal{O}_1$ is $(\mathbb{R}^+)^d$. Let

$$\Psi_{\mathcal{O}_1}(X) = \prod_{i\in[d]} \Phi_\varepsilon(X_i).$$

Then Properties (1) and (2) of Proposition 4.4 follow directly from Properties (1) and (2) of Claim 4.5. By the product rule and the definition of $\Psi_{\mathcal{O}_1}$, we have for any multi-index J:

$$\Psi_{\mathcal{O}_1}^{(J)}(X) = \prod_{i\in[d]} \Phi_{\varepsilon}^{(J_i)}(X_i). \tag{9}$$

Using property (3) of Claim 4.5, we get

$$\|\Psi_{\mathcal{O}_1}^{(J)}\|_\infty = \prod_{i\in[d]} \|\Phi_{\varepsilon}^{(J_i)}\|_\infty \le \prod_{i\in\mathrm{supp}(J)} \alpha(J_i)\cdot(1/\varepsilon)^{J_i} \le \alpha(k)\cdot(1/\varepsilon)^k.$$

The last inequality uses that $\sum_{i=1}^d J_i = k$ and that $\log\alpha(\cdot)$ is sub-additive. This gives property (3). To prove property (4), we again use (9) and observe

$$\Psi_{\mathcal{O}_1}^{(J)}(X) \neq 0 \;\Rightarrow\; \Phi_{\varepsilon}^{(J_i)}(X_i) \neq 0 \;\text{ for all } i \in [d].$$

For any $i \in [d]$ such that $J_i \neq 0$, the latter implies that $0 \le X_i \le \varepsilon$ since $\Phi_\varepsilon(X_i)$ is constant outside $0 \le X_i \le \varepsilon$. This finishes the proof of Proposition 4.4. □

4.2 Lindeberg's replacement method and an $\Omega(n^{1/4-c})$-query lower bound

Let $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, $i \in [n]$, denote independent random variables distributed according to $\boldsymbol{u}$ and $\boldsymbol{v}$ from Proposition 3.1 and 3.2 with $\ell = h$ and $\mu = \mu(h)$, for some odd constant $h = h(c) \in \mathbb{N}$ to be specified at the end of this subsection. We note that only in this subsection, Section 4.2, do we take $\ell = h$ rather than $\ell = h^3$ (for the $\Omega(n^{1/4-c})$ lower bound that we establish in this subsection, we only require $\ell = h$).

Let $\mathcal{X} \in \{\pm 1/\sqrt{n}\}^{d\times n}$ denote a query matrix, and let $\mathcal{X}^{(i)}$ denote its ith column. Recall that

$$\mathbf{S} = \sum_{i=1}^n \boldsymbol{u}_i \mathcal{X}^{(i)} \quad \text{and} \quad \mathbf{T} = \sum_{i=1}^n \boldsymbol{v}_i \mathcal{X}^{(i)}. \tag{10}$$

Our goal is to show that $d_{\mathrm{UO}}(\mathbf{S},\mathbf{T}) \le 0.1$ when $d = O(n^{1/4-c})$.

To this end, let $\mathcal{O}$ denote a union of orthants such that

$$d_{\mathrm{UO}}(\mathbf{S},\mathbf{T}) = \big|\mathbf{Pr}[\mathbf{S}\in\mathcal{O}] - \mathbf{Pr}[\mathbf{T}\in\mathcal{O}]\big|. \tag{11}$$

Following [Mos08, GOWZ10], we first use the Lindeberg replacement method to bound

$$\big|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{S})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{T})]\big|,$$

and then apply Proposition 4.1 to bound (11).

For all $i \in \{0, 1\ldots, n\}$ we introduce the $\mathbb{R}^d$-valued hybrid random variable:

$$\mathbf{Q}^{(i)} = \sum_{j=1}^{i} \boldsymbol{v}_j \mathcal{X}^{(j)} + \sum_{j=i+1}^{n} \boldsymbol{u}_j \mathcal{X}^{(j)},$$

and note that $\mathbf{Q}^{(0)} = \mathbf{S}$ and $\mathbf{Q}^{(n)} = \mathbf{T}$. Informally we think of getting $\mathbf{T}$ from $\mathbf{S}$ via $\mathbf{Q}^{(1)},\ldots,\mathbf{Q}^{(n-1)}$ by swapping out each of the summands $\boldsymbol{u}_j\mathcal{X}^{(j)}$ for $\boldsymbol{v}_j\mathcal{X}^{(j)}$ one by one. The main idea is to bound the difference in expectations

$$\big|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i-1)})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i)})]\big|, \tag{12}$$

since summing over all $i \in [n]$ gives an upper bound on

$$\big|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{S})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{T})]\big| = \big|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(0)})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(n)})]\big| \le \sum_{i=1}^{n} \big|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i-1)})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i)})]\big|$$

via the triangle inequality.

To bound (12), we define the random variable

$$\mathbf{R}_{-i} = \sum_{j=1}^{i-1} \boldsymbol{v}_j \mathcal{X}^{(j)} + \sum_{j=i+1}^{n} \boldsymbol{u}_j \mathcal{X}^{(j)} \tag{13}$$

and note that

$$\big|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i-1)})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i)})]\big| = \big|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{R}_{-i} + \boldsymbol{v}_i \mathcal{X}^{(i)})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{R}_{-i} + \boldsymbol{u}_i \mathcal{X}^{(i)})]\big|.$$

Truncating the Taylor expansion of $\Psi_{\mathcal{O}}$ at the h-th term (Fact 2.2), we get

$$\begin{aligned}\mathbf{E}\left[\Psi_{\mathcal{O}}(\mathbf{R}_{-i} + \boldsymbol{v}_i \mathcal{X}^{(i)})\right] &= \sum_{|J|\le h} \frac{1}{J!} \cdot \mathbf{E}\left[\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i}) \cdot (\boldsymbol{v}_i \mathcal{X}^{(i)})^J\right] \\ &\quad + \sum_{|J|=h+1} \frac{h+1}{J!} \cdot \mathbf{E}\left[(1-\boldsymbol{\tau})^h \cdot \Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i} + \boldsymbol{\tau} \cdot \boldsymbol{v}_i \mathcal{X}^{(i)}) \cdot (\boldsymbol{v}_i \mathcal{X}^{(i)})^J\right]\end{aligned} \tag{14}$$

where $\boldsymbol{\tau}$ is a random variable uniformly distributed on the interval $[0,1]$ (so the very last expectation is with respect to $\boldsymbol{\tau}$, $\boldsymbol{v}_i$ and $\mathbf{R}_{-i}$). Writing the analogous expression for $\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{R}_{-i} + \boldsymbol{u}_i \mathcal{X}^{(i)})]$, we observe that by Propositions 3.1 and 3.2 the first sums are equal term by term, i.e. we have

$$\sum_{|J|\le h} \frac{1}{J!} \cdot \mathbf{E}\left[\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i}) \cdot (\boldsymbol{v}_i \mathcal{X}^{(i)})^J\right] = \sum_{|J|\le h} \frac{1}{J!} \cdot \mathbf{E}\left[\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i}) \cdot (\boldsymbol{u}_i \mathcal{X}^{(i)})^J\right]$$

for each $|J| \le h$. Thus we may cancel all but the last terms to obtain

$$\big|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i-1)})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i)})]\big| \le \sum_{|J|=h+1} \frac{h+1}{J!} \cdot \|\Psi_{\mathcal{O}}^{(J)}\|_\infty \cdot \Big(\mathbf{E}\left[|(\boldsymbol{v}_i \mathcal{X}^{(i)})^J|\right] + \mathbf{E}\left[|(\boldsymbol{u}_i \mathcal{X}^{(i)})^J|\right]\Big).$$

Observe that there are $|\{J \in \mathbb{N}^d : |J| = h+1\}| = \Theta(d^{h+1})$ many terms in this sum. Recalling that each coordinate of $\mathcal{X}^{(i)}$ has magnitude $1/\sqrt{n}$, that both $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are supported on at most $h+1$ real values that depend only on h (by Propositions 3.1 and 3.2), and Proposition 4.3, we have

$$\big|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i-1)})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i)})]\big| = O_h(1) \cdot \left(\frac{d}{\varepsilon}\right)^{h+1} \cdot \frac{1}{n^{(h+1)/2}}. \tag{15}$$

Summing over all $i \in [n]$ costs us a factor of n and so we get

$$\big|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{S})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{T})]\big| = O_h(1) \cdot \left(\frac{d}{\varepsilon}\right)^{h+1} \cdot \frac{1}{n^{(h-1)/2}}.$$

With this in hand we are in place to apply Proposition 4.1. Let

$$\mathcal{B}_\varepsilon = \big\{X \in \mathcal{O} : |X_i| \le \varepsilon \text{ for some } i \in [d]\big\}.$$

Since both $\boldsymbol{v}$ and $\boldsymbol{u}$ are supported on values of magnitude $O_h(1)$, we have that both $\mathbf{Pr}[\mathbf{S} \in \mathcal{B}_\varepsilon]$ and $\mathbf{Pr}[\mathbf{T} \in \mathcal{B}_\varepsilon]$ are bounded by $O_h(d\varepsilon) + O_h(d/\sqrt{n})$ by using the standard 1-dimensional Berry-Esseen inequality (Theorem 3) together with a union bound across the d dimensions. So all in all we have

$$d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) \le O_h(d\varepsilon) + O_h(d/\sqrt{n}) + O_h(1) \cdot \left(\frac{d}{\varepsilon}\right)^{h+1} \cdot \frac{1}{n^{(h-1)/2}}.$$

We note as an aside at this point that given any $0 < c < 1/4$, we may take $\varepsilon = n^{-1/4}$ and take h to be the smallest odd integer at least $1/c$. Then the RHS above is $O_h(n^{-c})$ when $d = O(n^{1/4-c})$ as desired. This gives the $\Omega(n^{1/4-c})$ query lower bound claimed earlier:

Proposition 4.6. *Given any $0 < c < 1/4$, there is a $\kappa = \kappa(c) > 0$ such that any non-adaptive algorithm for testing whether $f : \{-1,1\}^n \to \{-1,1\}$ is monotone versus κ-far from monotone must use $\Omega(n^{1/4-c})$ queries.*

4.3 Going beyond $\Omega(n^{1/4})$

The setup for the $\Omega(n^{1/2-c})$ bound is exactly the same as that of the $\Omega(n^{1/4-c})$ bound except that $\boldsymbol{u}_i, \boldsymbol{v}_i$ are distributed according to $\boldsymbol{u}$ and $\boldsymbol{v}$ from Proposition 3.1 and 3.2, respectively, with $\ell = h^3$ and $\mu = \mu(\ell)$, for some odd constant $h = h(c) \in \mathbb{N}$ to be specified later (see Equation (22)). We then repeat Lindeberg's replacement method on two random variables $\mathbf{S}$ and $\mathbf{T}$ as defined in (10), but only using the first h matching moments of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ (with the higher $h^3 - h$ matching moments being reserved for another application of Lindeberg's method later, as mentioned in "(4): Handing pruned query sets" in Section 1.3 above).

The improvement to the $\Omega(n^{1/2-c})$ bound comes from a more careful analysis of the sum in (14) which in turn translates into a stronger bound on the difference (12) than that was given in (15). Specifically, rather than using the naive bound

$$\left|\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i} + \boldsymbol{\tau} \cdot \boldsymbol{v}_i \mathcal{X}^{(i)})\right| \le \|\Psi_{\mathcal{O}}^{(J)}\|_\infty = O_h(1) \cdot (1/\varepsilon)^{h+1}$$

for each of the $\Theta(d^{h+1})$ possible outcomes of $J \in \mathbb{N}^d$ (which shows up as the $O_h(1) \cdot (d/\varepsilon)^{h+1}$ term in (15)), we shall instead argue that almost all of these outcomes actually make a much smaller contribution than $O_h(1) \cdot (1/\varepsilon)^{h+1}$. For this purpose, we will leverage the fourth property of $\Psi_{\mathcal{O}}$ from Proposition 4.3; note that the proof of the $\Omega(n^{1/4-c})$ lower bound in Section 4.2 uses the first three properties of $\Psi_{\mathcal{O}}$ from Proposition 4.3, but not the fourth.

Recall ε is the parameter of our mollifier $\Psi_{\mathcal{O}}(\cdot)$. Throughout the rest of the paper we shall take

$$\varepsilon = n^{4/h-1/2} \quad \text{and} \quad \delta = n^{-1/2} \tag{16}$$

but we continue to write "ε" and "δ" as separate parameters for conceptual clarity. See Table 1 as a reference for parameter settings used from Section 4.3 through the rest of the paper.

Revisiting equation (14) of the proof above, we have that

$$\begin{aligned}&\left|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i-1)})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i)})]\right| \\ &\le O_h(1) \sum_{|J|=h+1} \left(\mathbf{E}\left[\left|\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i} + \boldsymbol{\tau} \cdot \boldsymbol{v}_i \mathcal{X}^{(i)}) \cdot (\boldsymbol{v}_i \mathcal{X}^{(i)})^J\right|\right] + \mathbf{E}\left[\left|\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i} + \boldsymbol{\tau} \cdot \boldsymbol{u}_i \mathcal{X}^{(i)}) \cdot (\boldsymbol{u}_i \mathcal{X}^{(i)})^J\right|\right]\right)\end{aligned}$$

Parameter settings	Where the parameters are set
$h = h(c) =$ smallest odd integer $> 5/c$	Equation (22)
$\ell = h^3$	Section 3
$\mu = \mu(\ell)$	Proposition 3.1
$\varepsilon = n^{4/h-1/2}$	Equation (16)
$\delta = n^{-1/2}$	Equation (16)
$\beta = O_h(1)$	Equation (17)

Table 1: Parameter settings used from Section 4.3 onward. The value "c" may be any positive absolute constant.

For each multi-index J with $|J| = h+1$ we relax

$$\mathbf{E}\left[\left|\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i} + \boldsymbol{\tau} \cdot \boldsymbol{v}_i \mathcal{X}^{(i)}) \cdot (\boldsymbol{v}_i \mathcal{X}^{(i)})^J\right|\right] \leq \mathbf{E}\left[\left|(\boldsymbol{v}_i \mathcal{X}^{(i)})^J\right| \cdot \sup_{T \in [-\beta\delta, \beta\delta]^d} \mathbf{E}\left[\left|\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i} + T)\right|\right]\right], \quad (17)$$

where $\beta = O_h(1)$ is an absolute constant that depends only on the largest value in the support of $\boldsymbol{v}$ (which depends only on h). Observe that since each coordinate of $\mathcal{X}^{(i)}$ has magnitude $1/\sqrt{n}$, each coordinate of the vector-valued random variable $\boldsymbol{\tau} \cdot \boldsymbol{v}_i \mathcal{X}^{(i)}$ is supported on values in $[-\beta\delta, \beta\delta]$, for the β as described above. Combining the above with an analogous bound for the $\boldsymbol{u}_i$ term, we have

$$\left|\mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i-1)})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{(i)})]\right| \leq \frac{O_h(1)}{n^{(h+1)/2}} \sum_{|J|=h+1} \left(\sup_{T \in [-\beta\delta, \beta\delta]^d} \mathbf{E}\left[\left|\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i} + T)\right|\right] \right). \quad (18)$$

We obtain an improved upper bound on this sum by exploiting the distributional properties of the d-dimensional random variable $\mathbf{R}_{-i} + T$. In particular we would like to show that for most ways of choosing $h+1$ out of the d coordinates, it is quite unlikely that all $h+1$ chosen coordinates can simultaneously take a value in the small interval $[-\beta\delta, \beta\delta]$. (Note that almost all J with $|J| = h+1$ satisfy $\#J = h+1$.) The fourth property of $\Psi_{\mathcal{O}}$ from Proposition 4.3 implies that having all these coordinates be small is the only way an outcome of $\mathbf{R}_{-i} + T$ can have

$$\mathbf{E}\left[\left|\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i} + T)\right|\right]$$

make a nonzero contribution to the sum in (18). In other words, we would like to use the fact that for all $J \in \mathbb{N}^d$ with $|J| = h+1$ we have

$$\sup_{T \in [-\beta\delta, \beta\delta]^d} \mathbf{E}\left[\left|\Psi_{\mathcal{O}}^{(J)}(\mathbf{R}_{-i} + T)\right|\right] \leq O_h(1) \cdot \left(\frac{1}{\varepsilon}\right)^{h+1} \cdot \mathbf{Pr}\left[(\mathbf{R}_{-i})|_J \in \mathcal{B}_J\right], \quad (19)$$

where we use $\mathcal{B}_J$ to denote the origin-centered $(\#J)$-dimensional box $[-\varepsilon - \beta\delta, \varepsilon + \beta\delta]^{\#J}$. Recall that the analysis of the previous subsection simply used the weaker bound obtained from (19) by upper bounding $\mathbf{Pr}[(\mathbf{R}_{-i})|_J \in \mathcal{B}_J]$ by 1.

Unfortunately, given an arbitrary query set, we cannot argue that the RHS of (19) is typically small. Indeed, consider a d-query set $\mathcal{X}$ in which a single fixed string $Q \in \{\pm 1/\sqrt{n}\}^n$ is repeated d times. In such a situation, *every* outcome of J will have

$$\mathbf{Pr}\left[(\mathbf{R}_{-i})|_J \in \mathcal{B}_J\right] = \mathbf{Pr}\left[(\mathbf{R}_{-i})_1 \in [-\varepsilon - \beta\delta, \varepsilon + \beta\delta]\right],$$

because every coordinate of every outcome of $(\mathbf{R}_{-i})$ is the same, and this probability over the 1-dimensional random variable $(\mathbf{R}_{-i})_1$ may be as large as $\Omega(\varepsilon)$; thus no significant savings is achieved over the earlier analysis. However, it is clear that such a query set $\mathcal{X}$ is highly "degenerate," in the sense that it can be replaced by a 1-query set (which we denote by $\mathcal{X}^*$) consisting of just one copy of Q, which will serve just as well as $\mathcal{X}$ for the purpose of monotonicity testing. (More precisely, the "union-of-orthants" distance $d_{\mathrm{UO}}(\mathbf{S},\mathbf{T})$ corresponding to the original query set will be precisely the same as the union-of-orthants distance $d_{\mathrm{UO}}(\mathbf{S}^*,\mathbf{T}^*)$ corresponding to the reduced query set $\mathcal{X}^*$.)

Is it possible that *every* "degenerate" query set (for which (18) is large) can be "pruned" down to an essentially equivalent query set (in terms of our d_{UO} measure) for which we can give a strong upper bound? Perhaps surprisingly, the answer is yes; however, doing this requires significant work and careful analysis. In the next section we describe and analyze our pruning procedure, and in Section 6 we show how an analysis based on (19) can handle pruned query sets.

5 Pruning a query set

In this section we explain how an arbitrary query set can be "pruned" so as to make it "scattered." (The definition of a "scattered" query set is somewhat complicated, involving the density of points that lie close to the linear span of other sets of points, so we defer it to Section 5.3.) We show that the pruning procedure has only a negligible effect on the variation distance $d_{\mathrm{UO}}(\mathbf{S},\mathbf{T})$ that we are aiming to bound. In later sections we give a lower bound against scattered query sets and thereby prove our main result.

We give some preliminary geometric results in Section 5.1, and after some setup in Section 5.2, describe and analyze the pruning procedure in Section 5.3.

5.1 Useful results about hypercubes and subspaces

The first geometric result we require is a variant of a well known fact due to Odlyzko [Odl88]. We begin by recalling the original fact:

Fact 5.1. *Let $\mathcal{V} \subseteq \mathbb{R}^n$ be a subspace of dimension k. Then $|\mathcal{V} \cap \{\pm 1/\sqrt{n}\}^n| \le 2^k$.*

Our variant is more restrictive than the original statement in that it only deals with subspaces $\mathcal{V}$ of the form $\mathcal{V} = \mathrm{span}\{V^{(1)},\dots,V^{(k)}\}$, for some $V^{(1)},\dots,V^{(k)} \in \{\pm 1/\sqrt{n}\}^n$ (though see Remark 6). However, the variant is significantly more general in that it gives us a bound on the number of Hamming balls that are required to cover all points of $\{\pm 1/\sqrt{n}\}^n$ that lie *close* to (and need not lie exactly on) the subspace $\mathcal{V}$. (Odlyzko's fact may be viewed as giving a bound on the number of radius-0 Hamming balls that are required to cover all points of $\{\pm 1/\sqrt{n}\}^n$ that lie exactly on $\mathcal{V}$.) A detailed statement and proof of our variant follow.

Given $r \ge 0$ and a subspace $\mathcal{V} \subseteq \mathbb{R}^n$, we define the *$r$-dilation of $\mathcal{V}$* to be the set

$$B_{\ell_2}(\mathcal{V}, r) := \bigcup_{V \in \mathcal{V}} B_{\ell_2}(V, r).$$

Our lemma is the following:

Lemma 5.2. *Given any set $\mathcal{A} = \{V^{(1)}, \dots, V^{(k)}\} \subseteq \{\pm 1/\sqrt{n}\}^n$ and any $r \geq 0$, there exists a set of at most 2^{k^2} points $\mathrm{cover}(\mathcal{A}) \subseteq \{\pm 1/\sqrt{n}\}^n$ such that*

$$B_{\ell_2}(\mathrm{span}(\mathcal{A}), r) \cap \{\pm 1/\sqrt{n}\}^n \subseteq \bigcup_{Y \in \mathrm{cover}(\mathcal{A})} B_{\mathrm{Ham}}(Y, r^2 n).$$

Observe that by taking $r = 0$, Lemma 5.2 recovers Fact 5.1 for $\mathcal{V} = \mathrm{span}\{V^{(1)}, \dots, V^{(k)}\}$ where $V^{(1)}, \dots, V^{(k)} \in \{\pm 1/\sqrt{n}\}^n$, with the somewhat weaker bound 2^{k^2} compared to 2^k.

Proof. Fix any $V \in \{\pm 1/\sqrt{n}\}^n$ such that $V \in B_{\ell_2}(\mathrm{span}(\mathcal{A}), r)$, so there exists a $U = \sum_{j=1}^k \alpha_j V^{(j)}$ such that $\|U - V\|_2 \leq r$. Let the vector $U_{\mathrm{round}} \in \{\pm 1/\sqrt{n}\}^n$ be defined by taking

$$(U_{\mathrm{round}})_i = \mathrm{sign}(U_i)/\sqrt{n} \in \{\pm 1/\sqrt{n}\}$$

for each $i \in [n]$. It is clear that we have

$$\|U_{\mathrm{round}} - V\|_2 = \frac{2}{\sqrt{n}} \sqrt{\sum_{i=1}^n \mathbf{1}\big[V_i \neq (U_{\mathrm{round}})_i\big]} \leq 2 \cdot \sqrt{\sum_{i=1}^n (U_i - V_i)^2} = 2 \cdot \|U - V\|_2 \leq 2r,$$

and also that

$$\|U_{\mathrm{round}} - V\|_2 = \sqrt{\frac{4 \cdot d_{\mathrm{Ham}}(U_{\mathrm{round}}, V)}{n}}.$$

As a result, we have $d_{\mathrm{Ham}}(U_{\mathrm{round}}, V) \leq r^2 n$.

Let $\mathrm{cover}(\mathcal{A}) \subseteq \{\pm 1/\sqrt{n}\}^n$ denote the following set of points:

$$\mathrm{cover}(\mathcal{A}) = \big\{U_{\mathrm{round}} : U \in \mathrm{span}(\mathcal{A})\big\}.$$

We will show that $|\mathrm{cover}(\mathcal{A})| \leq 2^{k^2}$; this establishes the lemma. To see this, note that

$$(U_{\mathrm{round}})_i = \mathrm{sign}\left(\sum_{j=1}^k \alpha_j \cdot V_i^{(j)}\right), \qquad \text{given } U = \sum_{j=1}^k \alpha_j \cdot V^{(j)}.$$

In other words, the i-th entry of U_{round} is given by the value of the k-variable LTF

$$f(Y) = \mathrm{sign}\Big(\textstyle\sum_{j=1}^k \alpha_j Y_j\Big)$$

evaluated on the fixed input $X^{(i)} = (V_i^{(1)}, \dots, V_i^{(k)}) \in \{\pm 1/\sqrt{n}\}^k$ (note that different U_{round}'s correspond to LTFs with different coefficients, but the n inputs $X^{(1)}, \dots, X^{(n)}$ on which the LTFs are evaluated are the same over all U_{round}'s). Thus we can upper bound the number of distinct vectors U_{round} by the number of distinct k-variable LTFs (viewed as Boolean functions) over $\{\pm 1/\sqrt{n}\}^k$, which is at most 2^{k^2} by Fact 2.1. □

Remark 6. Though we do not need it, we note that Lemma 5.2 may easily be generalized to allow each of $V^{(1)}, \dots, V^{(k)}$ to be an arbitrary point in $\mathbb{R}^n$, at the cost of having the RHS become n^{k+1} instead of 2^{k^2}. As the VC dimension of the class of all LTFs over $\mathbb{R}^k$ is $k+1$, Sauer's lemma tells us that the number of different ways that LTFs can label a fixed set of n points in $\mathbb{R}^k$ (like the points $X^{(1)}, \dots, X^{(n)}$) is at most $(en/(k+1))^{k+1} \leq n^{k+1}$.

The next geometric lemma that we require is the following:

Lemma 5.3. *Fix any positive integer h. There exist two constants $\gamma_1 = \gamma_1(h)$ and $\gamma_2 = \gamma_2(h)$ with the following property. For any $\mathcal{A} = \{V^{(1)}, \ldots, V^{(k)}\} \subset \{\pm 1/\sqrt{n}\}^n$ with $k \le h$ and any $V \in \{\pm 1/\sqrt{n}\}^n$, there is a vector $U = \beta_1 V^{(1)} + \cdots + \beta_k V^{(k)} \in \operatorname{span}(\mathcal{A})$ such that $|\beta_i| \le \gamma_1$ for all i and*

$$\|V - U\|_2 \le \gamma_2 \cdot d_{\ell_2}\big(V, \operatorname{span}(\mathcal{A})\big).$$

Roughly speaking, Lemma 5.3 shows that given any set of $k \le h$ vectors $\mathcal{A} = \{V^{(1)}, \ldots, V^{(k)}\}$ from $\{\pm 1/\sqrt{n}\}^n$ and a "target vector" $V \in \{\pm 1/\sqrt{n}\}^n$, there exists $U \in \operatorname{span}(\mathcal{A})$ such that U is almost as close to V in Euclidean distance as the closest point in $\operatorname{span}(\mathcal{A})$, and U can be written as a "low-weight" linear combination of the elements in $\mathcal{A}$. Note that there are competing demands imposed by keeping both parameters γ_1 and γ_2 small; for example, it is easy to see that either one may individually be made to be 1, but doing this may potentially cause the other one to become large. The crux of Lemma 5.3 is that it is possible to simultaneously have both γ_1 and γ_2 bounded by $O_h(1)$ independent of n.

Proof. Given $\mathcal{A}$ and V, we let $U = \beta_1 V^{(1)} + \cdots + \beta_k V^{(k)}$ denote the closest point to V in $\operatorname{span}(\mathcal{A})$. Below we view $\mathcal{A}$ as a $k \times n$ matrix, with $V^{(i)}$ being its i-th row vector. Note that $\mathcal{A}$ has $m \le 2^k$ many distinct columns, and we let $P^{(1)}, \ldots, P^{(m)}$ denote these column vectors in $\{\pm 1/\sqrt{n}\}^k$. Let $I \subseteq [m]$ denote the set of indices $i \in [m]$ such that coordinates of U that correspond to columns of type $P^{(i)}$ have absolute value at most $2/\sqrt{n}$. (Note that if two coordinates U_a, U_b of U correspond to the same column type $P^{(i)}$ then $U_a = U_b$.)

We consider two cases. For Case 1, we show that $\beta_1, \ldots, \beta_k$ already satisfy $|\beta_i| = O_h(1)$ for all $i \in [k]$, and we are done. For Case 2, we use $\beta_1, \ldots, \beta_k$ to obtain $\alpha_1, \ldots, \alpha_k$ such that $|\alpha_i| = O_h(1)$ for all $i \in [k]$ and $W = \alpha_1 V^{(1)} + \cdots + \alpha_k V^{(k)}$ has small Euclidean distance from V as claimed.

Case 1: The set of columns in $\{P^{(i)} : i \in I\}$ spans full dimension k. For this case we pick any k such columns, say $P^{(1)}, \ldots, P^{(k)}$ without loss of generality, in I. Then $(\beta_1, \ldots, \beta_k)$ is the unique solution to the following linear system of k equations in variables $x_1, \ldots, x_k$:

$$P^{(i)} \cdot (x_1, \ldots, x_k) = P^{(i)} \cdot (\beta_1, \ldots, \beta_k), \quad \text{for } i \in [k].$$

Each entry of the $k \times k$ coefficient matrix given by the $P^{(i)}$'s is $\pm 1/\sqrt{n}$, and the right side of each of the k equations has absolute value at most $2/\sqrt{n}$. By Cramer's rule it follows that $|\beta_i| = O_k(1) = O_h(1)$ for all i, and the lemma is proved in this case.

Case 2: The set of columns in $\{P^{(i)} : i \in I\}$ spans a space of dimension $j < k$. For this case we pick j independent columns from $\{P^{(i)} : i \in I\}$, say $P^{(1)}, \ldots, P^{(j)}$. Then we pick arbitrarily $k - j$ vectors $T^{(j+1)}, \ldots, T^{(k)}$ from $\{\pm 1/\sqrt{n}\}^n$ so that they together with $P^{(1)}, \ldots, P^{(j)}$ span full dimension k (note that $T^{(i)}$'s are not necessarily column vectors of $\mathcal{A}$). Solving the following linear system we get an alternative set of coefficients $\alpha_1, \ldots, \alpha_k$:

1. For each $i \in [j]$, we require $P^{(i)} \cdot (x_1, \ldots, x_k) = P^{(i)} \cdot (\beta_1, \ldots, \beta_k) \in [-2/\sqrt{n}, 2/\sqrt{n}]$.
2. For each $i \in [j+1 : k]$, we require $T^{(i)} \cdot (x_1, \ldots, x_k) = 0$.

Let $(\alpha_1, \ldots, \alpha_k)$ denote the unique solution to this linear system. Similar to Case 1, Cramer's rule implies that $|\alpha_i| = O_k(1) = O_h(1)$ for all $i \in [k]$.

Finally we complete the proof by showing that the vector $W = \alpha_1 V^{(1)} + \cdots + \alpha_k V^{(k)}$ is close to V; more precisely, we show that

$$\|V - W\|_2 = O_h(1) \cdot \|V - U\|_2 = O_h(1) \cdot d_{\ell_2}\big(V, \mathrm{span}(\mathcal{A})\big).$$

For this we just compare $W = \alpha_1 V^{(1)} + \cdots + \alpha_k V^{(k)}$ with $U = \beta_1 V^{(1)} + \cdots + \beta_k V^{(k)}$ entry by entry. Fix any $a \in [n]$ and suppose that the a-th column of $\mathcal{A}$ is of type $P^{(b)}$, for some $b \in [m]$. If $b \in I$, then it is clear that $U_a = W_a$. If $b \notin I$, then we have $|U_a - V_a| > 1/\sqrt{n}$ since $|U_a| > 2/\sqrt{n}$. On the other hand, from $|\alpha_i| = O_h(1)$ for all i we also have $|W_a| \le k \cdot O_h(1)/\sqrt{n}$ and thus,

$$|W_a - V_a| = O_h(1)/\sqrt{n} < O_h(1) \cdot |U_a - V_a|.$$

The claim now follows. □

5.2 Setup for the pruning procedure: compatibility between points and sets

We will use the following simple lemma, which follows directly from the Hoeffding inequality and the fact that $\boldsymbol{u}, \boldsymbol{v}$ are bounded and $\mathbf{E}[\boldsymbol{u}] = \mathbf{E}[\boldsymbol{v}]$. Recall that $h = h(c)$ is an odd integer constant.

Lemma 5.4. *Let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ denote n independent random variables, where each $\boldsymbol{w}_i$ is distributed according to either $\mathbf{u}$ or $\boldsymbol{v}$ given in Proposition 3.2 or 3.1 with $\ell = h^3$ and $\mu = \mu(\ell)$. Let $W \in \mathbb{R}^n$ and $\boldsymbol{x} = \sum_{i \in [n]} \boldsymbol{w}_i W_i$, then $\mathbf{E}[\boldsymbol{x}] = \mu \cdot \sum_{i \in [n]} W_i$. Moreover, we have*

$$\mathbf{Pr}\Big[|\boldsymbol{x} - \mathbf{E}[\boldsymbol{x}]| \ge \|W\|_2 \cdot (\log n)^{3/4}\Big] \le \frac{1}{n^{\omega(1)}}.$$

Now we define compatibility between a point $V \in \{\pm 1/\sqrt{n}\}^n$ and a set $\mathcal{A} \subset \{\pm 1/\sqrt{n}\}^n$. Let $\gamma_1 = \gamma_1(h)$ and $\gamma_2 = \gamma_2(h)$ denote the constants from Lemma 5.3. Recall that $\varepsilon = n^{4/h - 1/2}$.

Definition 7 (Compatibility)**.** Given $\mathcal{A} = \{V^{(1)}, \ldots, V^{(k)}\} \subset \{\pm 1/\sqrt{n}\}^n$ for some $k \le h$ and $V \in \{\pm 1/\sqrt{n}\}^n$, we say that V is *incompatible* with $\mathcal{A}$ if there *exist* real numbers $\beta_1, \ldots, \beta_k$ such that both (i) $|\beta_i| \le \gamma_1(h)$ for all $i \in [k]$ and (ii) the vector $U = \beta_1 V^{(1)} + \cdots + \beta_k V^{(k)} \in \mathrm{span}(\mathcal{A})$ satisfies

$$\Big| \textstyle\sum_{i \in [n]} (V_i - U_i) \Big| > \Big(\|V - U\|_2 + \varepsilon \Big) \cdot \log n.$$

Otherwise we say V is *compatible* with $\mathcal{A}$.

We may equivalently define compatibility as follows: V is compatible with $\mathcal{A}$ if for *every* $\beta_1, \ldots \beta_k$ of magnitude at most $\gamma_1(h)$, the vector $U = \beta_1 V^{(1)} + \cdots + \beta_k V^{(k)}$ satisfies

$$\Big| \textstyle\sum_{i \in [n]} (V_i - U_i) \Big| \le \Big(\|V - U\|_2 + \varepsilon \Big) \cdot \log n.$$

Recall from (19) that we would like to give a strong upper bound on $\mathbf{Pr}[(\mathbf{R}_{-i})|_J \in \mathcal{B}_J]$ for as many multi-indices J with $|J| = h+1$ as possible. Given a fixed set $\mathcal{X}$ of d query strings, a subset $\mathcal{A} \subset \mathcal{X} \subset \{\pm 1/\sqrt{n}\}^n$ of size $k \le h$ corresponds naturally to a multi-index J with $|J| = |\mathcal{A}|$. It is intuitively helpful to think of a multi-index J as being "built up" by successively adding elements from $\mathcal{X}$ to $\mathcal{A}$ one by one, starting with $\emptyset$. This motivates the above definition of incompatibility; as the following lemma shows, if a query string V is incompatible with $\mathcal{A}$, then we get a very strong bound on the probability $\mathbf{Pr}[(\mathbf{R}_{-i})|_J \in \mathcal{B}_J]$ for the multi-index J corresponding to $\{V\} \cup \mathcal{A}$ (which is desirable for our analysis). We will use this lemma later in Section 6.1 to deal with multi-indices corresponding to subsets of queries that contain a query that is incompatible with the other queries.

Lemma 5.5. *Suppose $V \in \{\pm 1/\sqrt{n}\}^n$ is incompatible with set $\mathcal{A} \subset \{\pm 1/\sqrt{n}\}^n$ where $k = |\mathcal{A}| \le h$. Let $(\mathcal{A}, V)$ be the $(k+1) \times n$ matrix whose rows are given by the vectors of $\mathcal{A}$ followed by V. Then*

$$\mathbf{Pr}\left[(\mathcal{A}, V) \cdot (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) \in [-2\varepsilon, 2\varepsilon]^{k+1}\right] = \frac{1}{n^{\omega(1)}},$$

where $\boldsymbol{w}_i$'s are independent random variables each of which is distributed according to $\boldsymbol{u}$ or $\boldsymbol{v}$.

Proof. Let $U \in \mathbb{R}^n$ be a linear combination of the elements of $\mathcal{A}$ that satisfies conditions (i) and (ii) of Definition 7 (the existence of U is guaranteed by the incompatibility of V with $\mathcal{A}$). Because U is a "low-weight" linear combination of vectors in $\mathcal{A}$, having $(\mathcal{A}, V) \cdot (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) \in [-2\varepsilon, 2\varepsilon]^{k+1}$ implies that $W := V - U$ also satisfies

$$\big|W \cdot (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)\big| = O_h(\varepsilon). \tag{20}$$

Next observe that by condition (ii) of Definition 7, we have that

$$\left|\textstyle\sum_{i\in[n]} W_i\right| > \big(\|W\|_2 + \varepsilon\big) \log n.$$

On the other hand, Lemma 5.4 gives us that

$$\mathbf{Pr}\left[\left|W \cdot (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) - \mu \textstyle\sum_{i\in[n]} W_i\right| \ge \|W\|_2 \cdot (\log n)^{3/4}\right] \le \frac{1}{n^{\omega(1)}};$$

together with the previous inequality, recalling that $0 < \mu = O_h(1)$, this gives

$$\mathbf{Pr}\left[\big|W \cdot (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)\big| = \Omega\big((\|W\|_2 + \varepsilon) \log n\big)\right] \ge 1 - \frac{1}{n^{\omega(1)}}.$$

This then implies that $|W \cdot (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)| = O_h(\varepsilon)$ with probability at most $1/n^{\omega(1)}$, which together with (20) establishes the lemma. □

Finally, the following lemma plays a key role in arguing about our pruning procedure:

Lemma 5.6. *Let $\mathcal{A} = \{V^{(1)}, \ldots, V^{(k)}\} \subset \{\pm 1/\sqrt{n}\}^n$ where $k \le h$, and $r \ge 0$. Let $\mathcal{R} \subset \{\pm 1/\sqrt{n}\}^n$ denote a set of points such that $\mathcal{R} \cap \mathcal{A} = \emptyset$ and $\mathcal{R} \subset B_{\ell_2}(\mathrm{span}(\mathcal{A}), r)$. Then one can partition the set $\mathcal{R}$ into three disjoint sets $\mathcal{R} = \mathcal{R}_{cover} \cup \mathcal{R}_{remove} \cup \mathcal{R}_{incomp}$ with the following properties:*

1. *$\mathcal{R}_{incomp}$ consists of all the points in $\mathcal{R}$ that are incompatible with $\mathcal{A}$;*
2. *$|\mathcal{R}_{cover}| \le 2^{h^2}$; and*
3. *For each point $W \in \mathcal{R}_{remove}$, there exists at least one point $V \in \mathcal{R}_{cover}$ such that $\|V - W\|_2 \le 4r$. Moreover, every such $V \in \mathcal{R}_{\mathrm{cover}}$ satisfies*

$$\left|\textstyle\sum_i (V_i - W_i)\right| \le (r + \varepsilon) \log^2 n. \tag{21}$$

As their names suggest, the points in $\mathcal{R}_{cover}$ will be used as a "cover" of the points in $\mathcal{R}_{remove}$, which will be removed from the query set in the pruning procedure described later. Also note that by condition (3), we must have $\mathcal{R}_{remove} = \emptyset$ when $r = 0$.

Proof. Let $\mathcal{R}_{incomp}$ be as described in (1) above, and let $\mathcal{R}' = \mathcal{R} \setminus \mathcal{R}_{incomp}$.

From Lemma 5.2, we know that there is a set $\mathrm{cover}(\mathcal{A}) \subset \{\pm 1/\sqrt{n}\}^n$ such that $|\mathrm{cover}(\mathcal{A})| \le 2^{h^2}$ and for any $W \in \mathcal{R}'$, there is a $V \in \mathrm{cover}(\mathcal{A})$ such that $\|V - W\|_2 \le 2r$. It follows that there exists a set $\mathcal{R}_{cover} \subseteq \mathcal{R}'$ with $|\mathcal{R}_{cover}| \le |\mathrm{cover}(\mathcal{A})| \le 2^{h^2}$ such that for any $W \in \mathcal{R}'$, there is a $V \in \mathcal{R}_{cover}$ such that $\|V - W\|_2 \le 4r$. Let $\mathcal{R}_{remove} = \mathcal{R}' \setminus \mathcal{R}_{cover}$. Then the only requirement that remains to be proven is the second inequality in (21).

To prove this, we let $\mathcal{A} = \{V^{(1)}, \ldots, V^{(k)}\}$ and let $U = \beta_1 V^{(1)} + \cdots + \beta_k V^{(k)}$ denote the vector guaranteed by Lemma 5.3 for $\mathcal{A}$ and V, with $\|\beta_i\| \le \gamma_1(h)$ for all i and

$$\|U - V\|_2 \le \gamma_2(h) \cdot d_{\ell_2}(\mathrm{span}(\mathcal{A}), V) \le \gamma_2(h) \cdot r.$$

Note that $\beta_1, \ldots, \beta_k$ satisfy condition (i) of Definition 7. As V is compatible with $\mathcal{A}$, we have

$$\Big|\textstyle\sum_i (V_i - U_i)\Big| \le \Big(\|V - U\|_2 + \varepsilon\Big) \cdot \log n.$$

Similarly, as W is compatible with $\mathcal{A}$ as well, we have

$$\Big|\textstyle\sum_i (W_i - U_i)\Big| \le \Big(\|W - U\|_2 + \varepsilon\Big) \cdot \log n.$$

Combining these two inequalities, we have

$$\Big|\sum (V_i - W_i)\Big| \le \Big|\sum (V_i - U_i)\Big| + \Big|\sum (W_i - U_i)\Big| \le \Big(\|V - U\|_2 + \|W - U\|_2 + 2\varepsilon\Big) \cdot \log n.$$

Combining this with $\|V - U\|_2 \le \gamma_2(h) \cdot r$ and

$$\|W - U\|_2 \le \|W - V\|_2 + \|V - U\|_2 \le \big(4 + \gamma_2(h)\big) \cdot r,$$

the second part of (21) is proven. This finishes the proof of the lemma. □

5.3 The pruning procedure and its analysis

Let $\mathcal{X} = \{X^{(1)}, \ldots, X^{(d)}\} \subseteq \{\pm 1/\sqrt{n}\}^n$ denote a query set of size d. We view $\mathcal{X}$ as a $d \times n$ matrix with $X^{(i)} \in \{\pm 1/\sqrt{n}\}^n$ being its i-th row vector and $\mathcal{X}^{(j)} \in \{\pm 1/\sqrt{n}\}^d$ its j-th column vector.

Fix a $c > 0$, we now specify the function h:

$$h(c) = \text{the smallest odd integer} \ge 5/c, \tag{22}$$

and recall that $\ell = h^3$. Recall our goal is to show that any query set $\mathcal{X}$ of size $d \le n^{1/2-c}$ satisfies

$$d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) = \max\Big\{\big|\mathbf{Pr}[\mathbf{S} \in \mathcal{O}] - \mathbf{Pr}[\mathbf{T} \in \mathcal{O}]\big| : \mathcal{O} \text{ is a union of orthants in } \mathbb{R}^d\Big\} \le 0.1.$$

Here $\mathbf{S} = \sum_j \boldsymbol{u}_j \mathcal{X}^{(j)}$ and $\mathbf{T} = \sum_j \boldsymbol{v}_j \mathcal{X}^{(j)}$, where $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are independent random variables with the same distribution as $\boldsymbol{u}$ and $\boldsymbol{v}$ from Proposition 3.2 and 3.1, given constants ℓ and $\mu(\ell)$.

Next we describe a procedure that "prunes" $\mathcal{X}$ and outputs a new query set $\mathcal{X}^* \subseteq \mathcal{X}$, which is almost as good as $\mathcal{X}$ for monotonicity testing, and is what we call a *scattered* query set.

Definition 8 (Scattered query sets)**.** Fix $\mathcal{A} \subseteq \mathcal{X}$ with $0 <|\mathcal{A}| \leq h$ and a value $r > 0$. Let

$$\mathcal{R} = \big(\mathcal{X} \cap B_{\ell_2}(\mathrm{span}(\mathcal{A}), r)\big) \setminus \mathcal{A},$$

and let $\mathcal{R} = \mathcal{R}_{cover} \cup \mathcal{R}_{remove} \cup \mathcal{R}_{incomp}$ denote the partition of $\mathcal{R}$ promised by Lemma 5.6. We say that $\mathcal{A}$ is *r-scattered* if $\mathcal{R}_{remove}$ satisfies

$$|\mathcal{R}_{remove}| \leq r|\mathcal{X}| \log^5 n. \tag{23}$$

We say that $\mathcal{X}$ is *scattered* if $\mathcal{A}$ is r-scattered for every $\mathcal{A} \subseteq \mathcal{X}$ with $0 <|\mathcal{A}| \leq h$ and every $r > 0$.

The parameter r above should be thought of as close to zero. Thus the rough idea is that in a scattered query set $\mathcal{X}$, for every small subset $\mathcal{A} \subset \mathcal{X}$, only a small number of points in $\mathcal{X}$ that lie close to the span of $\mathcal{A}$ are compatible with $\mathcal{A}$. Recall that as discussed earlier, small subsets $\mathcal{A}$ (of size at most h) correspond to different choices of the multi-index $J \in \mathbb{N}^d$ in (18). Intuitively, our analysis can handle points that do not lie close to the span of $\mathcal{A}$ (we make this intuition precise in Proposition 6.3), and as discussed above in Lemma 5.5, points that are incompatible with $\mathcal{A}$ are also good for our analysis. Having a query set be scattered will aid us in bounding the sum in (18); in particular, we will show that for a scattered query set, most multi-indices J are such that $\mathbf{Pr}\left[(\mathbf{R}_{-i})|_J \in \mathcal{B}_J\right] \lessapprox \varepsilon^{\#J}$. This will result in a substantially better bound in (18).

We now state the main lemma, which describes the effect of our pruning procedure:

Lemma 5.7. *Fix $c > 0$, and let $h = h(c)$ be as defined in* (22)*. Given a query set $\mathcal{X} \subseteq \{\pm 1/\sqrt{n}\}^n$ with $|\mathcal{X}| \leq n^{1/2-c}$, there exists a scattered query set $\mathcal{X}^* \subseteq \mathcal{X}$ (so $|\mathcal{X}^*| \leq |\mathcal{X}|$) such that*

$$d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) \leq d_{\mathrm{UO}}(\mathbf{S}^*, \mathbf{T}^*) + 0.01,$$

where $\mathbf{S}^* = \sum_j \boldsymbol{u}_j \mathcal{X}^{*(j)}$ and $\mathbf{T}^* = \sum_j \boldsymbol{v}_j \mathcal{X}^{*(j)}$.

Assuming Lemma 5.7, it now suffices to show that $d_{\mathrm{UO}}(\mathbf{S}^*, \mathbf{T}^*) \leq 0.09$ for any scattered query set $\mathcal{X}^* \subseteq \{\pm 1/\sqrt{n}\}^n$ of size $|\mathcal{X}^*| \leq n^{1/2-c}$, which we will do in the following sections.

The basic step of our pruning procedure is quite straightforward:

Pruning($\mathcal{X}$):

1. If $\mathcal{X}$ is not scattered, find any pair $(\mathcal{A}, r)$ with $\mathcal{A} \subseteq \mathcal{X}$, $0 <|\mathcal{A}| \leq h$, $r > 0$ such that $\mathcal{A}$ is not r-scattered (i.e. (23) is violated). For any such $\mathcal{A}$ choose the largest possible r which violates (23).
2. Let $\mathcal{R} = (\mathcal{X} \cap B_{\ell_2}(\mathrm{span}(\mathcal{A}), r)) \setminus \mathcal{A}$ and let $\mathcal{R} = \mathcal{R}_{cover} \cup \mathcal{R}_{remove} \cup \mathcal{R}_{incomp}$ denote the partition as promised by Lemma 5.6.
3. Remove all points of $\mathcal{R}_{remove}$ from $\mathcal{X}$.

Given a query set $\mathcal{X} \subset \{\pm 1/\sqrt{n}\}^n$, we can iteratively prune $\mathcal{X}$ via the Pruning procedure above until we obtain a scattered query set as defined in Definition 8. Starting with a query set $\mathcal{X}$ with $|\mathcal{X}| \leq n^{1/2-c}$, we write $\mathcal{X} = \mathcal{X}_0 \supset \mathcal{X}_1 \supset \cdots \supset \mathcal{X}_t = \mathcal{X}^*$ to denote the sequence of query sets we get from calling Pruning repeatedly until $\mathcal{X}^*$ is scattered. Note that the final set $\mathcal{X}^*$ will be nonempty, because $\mathcal{A} \cap \mathcal{R} = \emptyset$ for the sets $\mathcal{A}, \mathcal{R}$ used in the final application of Pruning and thus $\mathcal{A}$ remains in $\mathcal{X}$ at the end of Pruning.

To prove Lemma 5.7, we show that $\mathcal{X}^*$ is almost as effective as $\mathcal{X}$ in the following sense:

Claim 5.8. $d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) \le d_{\mathrm{UO}}(\mathbf{S}^*, \mathbf{T}^*) + 0.01.$

Proof. For each $i = 0, 1, \ldots, t-1$, let $(\mathcal{A}_i, r_i)$ denote the pair identified in Step 1 of Pruning, when it is run on $\mathcal{X}_i$. From (23) we have $|\mathcal{X}_i| - |\mathcal{X}_{i+1}| > r_i |\mathcal{X}_i| \log^5 n$. On the other hand, we have

$$\sum_{i=0}^{t-1} \frac{|\mathcal{X}_i| - |\mathcal{X}_{i+1}|}{|\mathcal{X}_i|} \le \sum_{i=0}^{t-1} \left(\frac{1}{|\mathcal{X}_{i+1}| + 1} + \ldots + \frac{1}{|\mathcal{X}_i|} \right) = O(\log |\mathcal{X}|) = O(\log n).$$

We conclude that $\sum_i r_i = O(1/\log^4 n)$. Next, let

$$\mathbf{S}_i = \textstyle\sum_j \boldsymbol{u}_j \mathcal{X}_i^{(j)}, \quad \mathbf{T}_i = \sum_j \boldsymbol{v}_j \mathcal{X}_i^{(j)}, \quad \mathbf{S}_{i+1} = \sum_j \boldsymbol{u}_j \mathcal{X}_{i+1}^{(j)} \quad \text{and} \quad \mathbf{T}_{i+1} = \sum_j \boldsymbol{v}_j \mathcal{X}_{i+1}^{(j)}.$$

We compare $d_{\mathrm{UO}}(\mathbf{S}_i, \mathbf{T}_i)$ and $d_{\mathrm{UO}}(\mathbf{S}_{i+1}, \mathbf{T}_{i+1})$, and our goal is to show that

$$d_{\mathrm{UO}}(\mathbf{S}_i, \mathbf{T}_i) \le d_{\mathrm{UO}}(\mathbf{S}_{i+1}, \mathbf{T}_{i+1}) + O\big((r_i + \varepsilon) \log^3 n + (1/\sqrt{n})\big). \tag{24}$$

It then follows that

$$d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) \le d_{\mathrm{UO}}(\mathbf{S}^*, \mathbf{T}^*) + O(t/\sqrt{n} + t\varepsilon \log^3 n) + \textstyle\sum_i O(r_i \log^3 n) < d_{\mathrm{UO}}(\mathbf{S}^*, \mathbf{T}^*) + 0.01,$$

since we have $\sum r_i = O(1/\log^4 n)$, $t \le n^{1/2-c}$ and $\varepsilon = n^{4/h-1/2}$ with $h \ge 5/c$ (as defined in (22)).

Fix an i and let $\mathcal{R} = (\mathcal{X}_i \cap B_{\ell_2}(\mathrm{span}(\mathcal{A}_i), r_i)) \setminus \mathcal{A}_i$ and write $\mathcal{R} = \mathcal{R}_{cover} \cup \mathcal{R}_{remove} \cup \mathcal{R}_{incomp}$ as in Lemma 5.6. Let $\mathcal{O}_i$ be a union of orthants in $|\mathcal{X}_i|$-dimensional space with

$$d_{\mathrm{UO}}(\mathbf{S}_i, \mathbf{T}_i) = \big| \mathbf{Pr}[\mathbf{S}_i \in \mathcal{O}_i] - \mathbf{Pr}[\mathbf{T}_i \in \mathcal{O}_i] \big|.$$

Given $\mathcal{O}_i$, below we define a union of orthants $\mathcal{O}_{i+1}$ in $|\mathcal{X}_{i+1}|$-dimensional space. We will then show that $\mathcal{O}_{i+1}$ satisfies (27) below and thereby obtain (24).

We start with some terminology. Recall that an orthant in $|\mathcal{X}_i|$-dimensional space can be viewed as an assignment of a $\{\pm 1\}$ value to each element of $\mathcal{X}_i$. Given $V \in \mathcal{X}_i$ and an orthant $\mathcal{T}$ in $|\mathcal{X}_i|$-dimensional space, we let $\mathcal{T}(V) \in \{\pm 1\}$ denote the value assigned to V by $\mathcal{T}$. We say an orthant $\mathcal{T}$ in $|\mathcal{X}_i|$-dimensional space (but not necessarily in $\mathcal{O}_i$) is *bad* if there exist $W \in \mathcal{R}_{remove}, V \in \mathcal{R}_{cover}$ such that $\|V - W\|_2 \le 4r_i$ but $\mathcal{T}(V) \ne \mathcal{T}(W)$; otherwise we say $\mathcal{T}$ is a *good* orthant. Observe that by Lemma 5.6, a good orthant $\mathcal{T}$ is uniquely determined by its values $\mathcal{T}(V)$, $V \in \mathcal{X}_{i+1}$. We let $\mathcal{O}_{i,b}$ denote the union of bad orthants in $\mathcal{O}_i$, and let $\mathcal{O}_{i,g}$ denote the union of good orthants in $\mathcal{O}_i$.

As we will see below in Claim 5.9, the probability of $\mathbf{S}_i$ or $\mathbf{T}_i$ lying in a bad orthant is negligible. Thus, most of $|\mathbf{Pr}[\mathbf{S}_i \in \mathcal{O}_i] - \mathbf{Pr}[\mathbf{T}_i \in \mathcal{O}_i]|$ comes from good orthants of $\mathcal{O}_i$. Inspired by this, we will take $\mathcal{O}_{i+1}$ to be the projection of good orthants of $\mathcal{O}_i$ onto the $|\mathcal{X}_{i+1}|$-dimensional space.

We define formally $\mathcal{O}_{i+1}$ as follows. We say orthants $\mathcal{T}$ and $\mathcal{T}'$ in $|\mathcal{X}_i|$- and $|\mathcal{X}_{i+1}|$-dimensional space, respectively, are *consistent* if every $V \in \mathcal{X}_{i+1}$ satisfies $\mathcal{T}(V) = \mathcal{T}'(V)$. Given $\mathcal{O}_i$, we define $\mathcal{O}_{i+1}$ to be the union of orthants in $|\mathcal{X}_{i+1}|$-dimensional space each of which is consistent with a good orthant of $\mathcal{O}_i$. By definition, there is a bijection between orthants of $\mathcal{O}_{i+1}$ and good orthants of $\mathcal{O}_i$. For each orthant $\mathcal{T}'$ of $\mathcal{O}_{i+1}$, we let $g(\mathcal{T}')$ denote the corresponding good orthant $\mathcal{T}$ of $\mathcal{O}_i$; let $b(\mathcal{T}')$ denote the *union* of all bad $|\mathcal{X}_i|$-dimensional orthants $\mathcal{T}$ (not necessarily in $\mathcal{O}_i$) that are consistent with $\mathcal{T}'$.

We delay the proof of the following claim:

Claim 5.9. *Let $\mathcal{O}^*$ denote the union of all bad orthants in $|\mathcal{X}_i|$-dimensional space. Then*

$$\mathbf{Pr}[\mathbf{S}_i \in \mathcal{O}^*],\, \mathbf{Pr}[\mathbf{T}_i \in \mathcal{O}^*] = O\big((r_i + \varepsilon)\log^3 n\big) + O(1/\sqrt{n}). \tag{25}$$

Returning to the proof of Claim 5.8, for each orthant $\mathcal{T}'$ in $\mathcal{O}_{i+1}$ we have

$$\begin{aligned} \mathbf{Pr}[\mathbf{S}_{i+1} \in \mathcal{T}'] &= \mathbf{Pr}[\mathbf{S}_i \in g(\mathcal{T}')] + \mathbf{Pr}[\mathbf{S}_i \in b(\mathcal{T}')] \quad \text{and} \\ \mathbf{Pr}[\mathbf{T}_{i+1} \in \mathcal{T}'] &= \mathbf{Pr}[\mathbf{T}_i \in g(\mathcal{T}')] + \mathbf{Pr}[\mathbf{T}_i \in b(\mathcal{T}')]. \end{aligned} \tag{26}$$

Combining (25) and (26), we have

$$\begin{aligned} \big|\mathbf{Pr}[\mathbf{S}_i \in \mathcal{O}_i] - \mathbf{Pr}[\mathbf{T}_i \in \mathcal{O}_i]\big| &\le \big|\mathbf{Pr}[\mathbf{S}_i \in \mathcal{O}_{i,b}]\big| + \big|\mathbf{Pr}[\mathbf{T}_i \in \mathcal{O}_{i,b}]\big| + \big|\mathbf{Pr}[\mathbf{S}_i \in \mathcal{O}_{i,g}] - \mathbf{Pr}[\mathbf{T}_i \in \mathcal{O}_{i,g}]\big| \\ &\le O\big((r_i+\varepsilon)\log^3 n + 1/\sqrt{n}\big) + \big|\mathbf{Pr}[\mathbf{S}_{i+1} \in \mathcal{O}_{i+1}] - \mathbf{Pr}[\mathbf{T}_{i+1} \in \mathcal{O}_{i+1}]\big| \\ &\quad + \Big|\textstyle\sum_{\mathcal{T}' \text{ in } \mathcal{O}_{i+1}} \mathbf{Pr}[\mathbf{S}_i \in b(\mathcal{T}')]\Big| + \Big|\textstyle\sum_{\mathcal{T}' \text{ in } \mathcal{O}_{i+1}} \mathbf{Pr}[\mathbf{T}_i \in b(\mathcal{T}')]\Big| \\ &= O\big((r_i+\varepsilon)\log^3 n + 1/\sqrt{n}\big) + \big|\mathbf{Pr}[\mathbf{S}_{i+1} \in \mathcal{O}_{i+1}] - \mathbf{Pr}[\mathbf{T}_{i+1} \in \mathcal{O}_{i+1}]\big|, \end{aligned} \tag{27}$$

where the sums are over all orthants $\mathcal{T}'$ in $\mathcal{O}_{i+1}$. The last inequality used (25) as well as the fact that the $b(\mathcal{T}')$'s are unions of disjoint bad orthants in $|\mathcal{X}_i|$-dimensional space.

This finishes the proof of Claim 5.8. □

It remains to prove Claim 5.9.

Proof of Claim 5.9. We focus on $\mathbf{Pr}[\mathbf{S}_i \in \mathcal{O}^*]$ since the same argument works for $\mathbf{Pr}[\mathbf{T}_i \in \mathcal{O}^*]$.

By the definition of bad orthants in $|\mathcal{X}_i|$-dimensional space, we have

$$\mathbf{Pr}[\mathbf{S}_i \in \mathcal{O}^*] \le \sum_{V \in \mathcal{R}_{cover}} \mathbf{Pr}\Big[\exists W \in \mathcal{R}_{remove}\colon \|V-W\|_2 \le 4r_i \text{ but } \mathbf{S}_i \text{ has different signs on } V, W\Big].$$

Observe that the number of terms in the sum is $|\mathcal{R}_{cover}| = O_h(1)$.

Fix a $V \in \mathcal{R}_{cover}$. By Lemma 5.4, we have for every $W \in \mathcal{R}_{remove}$ such that $\|V-W\|_2 \le 4r_i$:

$$\mathbf{Pr}\left[\Big|\textstyle\sum_j \boldsymbol{u}_j(V_j - W_j) - \mathbf{E}[\boldsymbol{u}] \cdot \sum_j (V_j - W_j)\Big| \ge 4r_i \cdot (\log n)^{3/4}\right] \le \frac{1}{n^{\omega(1)}}. \tag{28}$$

Since the number of such W is at most $n^{1/2-c}$, we have

$$\mathbf{Pr}\Big[\exists W \in \mathcal{R}_{remove} \text{ s.t. } \|V-W\|_2 \le 4r_i \text{ and satisfies the condition in (28)}\Big] \le \frac{1}{n^{\omega(1)}}. \tag{29}$$

On the other hand, using the standard 1-dimensional Berry–Esséen Theorem (Theorem 3), and recalling that $\|V\|_2 = 1$ and each $\boldsymbol{u}_j$ has variance 1, we have

$$\mathbf{Pr}\left[\Big|\textstyle\sum_j \boldsymbol{u}_j V_j\Big| \ge (r_i+\varepsilon)\log^3 n\right] = 1 - O\big((r_i+\varepsilon)\log^3 n\big) - O(1/\sqrt{n}). \tag{30}$$

By Lemma 5.6 every $W \in \mathcal{R}_{remove}$ with $\|V-W\|_2 \le 4r_i$ satisfies $\big|\sum_j (V_j - W_j)\big| \le (r_i+\varepsilon)\log^2 n$. Combining this with (29) and (30), we have that the probability of

$$\operatorname{sign}\Big(\sum \boldsymbol{u}_j V_j\Big) = \operatorname{sign}\Big(\sum \boldsymbol{u}_j W_j\Big), \quad \text{for all } W \in \mathcal{R}_{remove} \text{ with } \|V-W\|_2 \le 4r_i,$$

is at least $1 - O((r_i+\varepsilon)\log^3 n) - O(1/\sqrt{n})$. Claim 5.9 then follows. □

6 A lower bound against scattered query sets

With the pruning procedure of the previous section in hand — showing how an arbitrary query set can be pruned so as to make it scattered — we now focus on proving a lower bound against scattered query sets via the approach outlined in Section 4.3. In Section 7 we will see how such a lower bound along with our analysis in the previous section can be easily combined to complete the proof of our main theorem, Theorem 1.

We briefly recall the setup of our approach. Fix a $c > 0$, and let h and ℓ be defined as in (22). Recall that $\varepsilon = n^{4/h-1/2}$ and $\delta = 1/\sqrt{n}$. Let $\mathcal{X} = \{X^{(1)}, \ldots, X^{(d)}\} \subseteq \{\pm 1/\sqrt{n}\}^n$ denote a query set with $d \leq n^{1/2-c}$. Let $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ denote independent random variables with the same distribution as $\boldsymbol{u}$ and $\boldsymbol{v}$ given in Proposition 3.1 and 3.2 with parameters ℓ and $\mu = \mu(\ell)$. Recall that

$$\mathbf{R}_{-i} = \sum_{j=1}^{i-1} \boldsymbol{v}_j \mathcal{X}^{(j)} + \sum_{j=i+1}^{n} \boldsymbol{u}_j \mathcal{X}^{(j)}$$

as defined in (13). Revisiting our discussion in Section 4.3, recall that our goal is to upper bound the quantity on the RHS of (18) using (19); to be precise we would like to show that the probability $\mathbf{Pr}[(\mathbf{R}_{-i})|_J \in \mathcal{B}_J]$ is typically small for most choices of $J \in \mathbb{N}^d$ with $|J| = h+1$, where $\mathcal{B}_J$ denotes the origin-centered $(\#J)$-dimensional box $[-\varepsilon - \beta\delta, \varepsilon + \beta\delta]^{\#J}$ and $\beta = O_h(1)$ is an absolute constant that depends only on the largest value in the support of $\boldsymbol{u}$ and $\boldsymbol{v}$. We do so in this section via the following lemma:

Lemma 6.1. *If $\mathcal{X}$ is a scattered query set of size $d \leq n^{1/2-c}$, then*

$$\sum_{|J|=h+1} \mathbf{Pr}\left[(\mathbf{R}_{-i})|_J \in \mathcal{B}_J\right] = O\Big(d^{h+1} \cdot \big(1/d + \varepsilon \log^6 n\big)^h\Big), \tag{31}$$

where $\mathcal{B}_J$ denotes the origin-centered $(\#J)$-dimensional box $\mathcal{B}_J = [-\varepsilon - \beta\delta, \varepsilon + \beta\delta]^{\#J}$.

Instead of focusing on the sum in (31) we let $\mathbf{I} = (\mathbf{V}^{(1)}, \ldots, \mathbf{V}^{(h+1)})$ denote a sequence of $h+1$ points sampled from $\mathcal{X}$ uniformly at random, with replacement. Let $\#\mathbf{I}$ denote the number of distinct points in $\mathbf{I}$, and $(\mathbf{R}_{-i})_{\mathbf{I}}$ denote the projection of $\mathbf{R}_{-i}$ onto the coordinates that correspond to points in $\mathbf{I}$. Lemma 6.1 then follows directly from the following lemma, as the distribution of $\mathbf{I}$ is close to the uniform distribution over J with $|J| = h+1$.

Lemma 6.2. *If $\mathcal{X}$ is a scattered query set of size $d \leq n^{1/2-c}$, then*

$$\mathop{\mathbf{E}}_{\mathbf{I}}\Big[\mathbf{Pr}\left[(\mathbf{R}_{-i})|_{\mathbf{I}} \in \mathcal{B}_{\mathbf{I}}\right]\Big] = O\Big(\big(1/d + \varepsilon \log^6 n\big)^h\Big), \tag{32}$$

where $\mathcal{B}_{\mathbf{I}}$ denotes the origin-centered $(\#\mathbf{I})$-dimensional box $\mathcal{B}_{\mathbf{I}} = [-\varepsilon - \beta\delta, \varepsilon + \beta\delta]^{\#\mathbf{I}}$.

Proof of Lemma 6.1. Let g denote the following natural map from $\mathcal{X}^{h+1}$ to $\{J \in \mathbb{N}^d : |J| = h+1\}$:

$$g(I) = (J_1, \ldots, J_d),$$

where J_i is the number of times $X^{(i)}$ appears in $I = (V^{(1)}, \ldots, V^{(h+1)}) \in \mathcal{X}^{h+1}$.

It is clear that g is surjective. As a result, we have

$$\sum_{|J|=h+1} \mathbf{Pr}\left[(\mathbf{R}_{-i})|_J \in \mathcal{B}_J\right] \leq \sum_{I \in \mathcal{X}^{h+1}} \mathbf{Pr}\left[(\mathbf{R}_{-i})|_I \in \mathcal{B}_I\right] = d^{h+1} \cdot \mathop{\mathbf{E}}_{\mathbf{I}}\Big[\mathbf{Pr}\left[(\mathbf{R}_{-i})|_{\mathbf{I}} \in \mathcal{B}_{\mathbf{I}}\right]\Big].$$

Lemma 6.1 then follows directly from Lemma 6.2. □

Let $I = (V^{(1)}, \ldots, V^{(h+1)}) \in \mathcal{X}^{h+1}$. For each $j \in [2 : h+1]$, we let

$$d_j = d_{\ell_2}\big(V^{(j)}, \mathrm{span}\{V^{(1)}, \ldots, V^{(j-1)}\}\big), \tag{33}$$

and define η_j for each $j \in [2 : h+1]$ as:

$$\eta_j = \begin{cases} 0 & \text{if } V^{(j)} \text{ is incompatible with } \{V^{(1)}, \ldots, V^{(j-1)}\} \\ 1 & \text{if } V^{(j)} \text{ is compatible with } \{V^{(1)}, \ldots, V^{(j-1)}\} \text{ but } d_j < \varepsilon \\ \varepsilon/d_j & \text{otherwise.} \end{cases} \tag{34}$$

We will use the following proposition to bound $\mathbf{Pr}[(\mathbf{R}_{-i})|_I \in \mathcal{B}_I]$:

Proposition 6.3. *Given an $I = (V^{(1)}, \ldots, V^{(h+1)}) \in \mathcal{X}^{h+1}$, we have*

$$\mathbf{Pr}\left[(\mathbf{R}_{-i})|_I \in \mathcal{B}_I\right] \le O\left(\prod_{j=2}^{h+1} \eta_j\right) + O\left(\frac{1}{n^h}\right),$$

where $\mathcal{B}_I$ denotes the origin-centered $(\#I)$-dimensional box $\mathcal{B}_I = [-\varepsilon - \beta\delta, \varepsilon + \beta\delta]^{\#I}$.

We delay the proof of Proposition 6.3 to the next section, but first use it to prove Lemma 6.2.

Proof of Lemma 6.2 assuming Proposition 6.3. Let $\mathbf{d}_j$ and $\boldsymbol{\eta}_j$ denote two random variables defined from $\mathbf{I}$ in the same fashion as (33) and (34). By Proposition 6.3, it suffices to show that

$$\mathop{\mathbf{E}}_{\mathbf{I}}\left[\prod_{j=2}^{h+1} \boldsymbol{\eta}_j\right] = O\Big(\big(1/d + \varepsilon \log^6 n\big)^h\Big).$$

Note that $\boldsymbol{\eta}_j$ is a nonnegative random variable with $\mathbf{Pr}[\boldsymbol{\eta}_j \le 1] = 1$.

Fix $j \in [2 : h+1]$. Let $(V^{(1)}, \ldots, V^{(j-1)}) \in \mathcal{X}^{j-1}$ denote a possible outcome of $(\mathbf{V}^{(1)}, \ldots, \mathbf{V}^{(j-1)})$ and let $\mathcal{A}$ denote the *set* that consists of $V^{(1)}, \ldots, V^{(j-1)}$, so $|\mathcal{A}| \le j - 1 \le h$. For any $r > 0$, let

$$\mathcal{R} = \big(\mathcal{X} \cap B_{\ell_2}(\mathrm{span}(\mathcal{A}), r)\big) \setminus \mathcal{A} \quad \text{and} \quad \mathcal{R} = \mathcal{R}_{cover} \cup \mathcal{R}_{remove} \cup \mathcal{R}_{incomp}$$

denotes the three-way partition of $\mathcal{R}$ promised by Lemma 5.6. By definition, we have

$$\big|\mathcal{R}_{cover} \cup \mathcal{R}_{remove}\big| \le rd\log^5 n + 2^{h^2} = rd\log^5 n + O_h(1).$$

This implies that, conditioning on $\mathcal{A}$ being the set of the first $j - 1$ points sampled in $\mathbf{I}$:

$$\mathbf{Pr}\Big[\boldsymbol{d}_j \le r \text{ and } \mathbf{V}^{(i)} \text{ is compatible with } \mathcal{A} \,|\, \mathcal{A}\Big] \le r\log^5 n + O_h(1/d), \quad \text{for all } r > 0. \tag{35}$$

By the definition of $\boldsymbol{\eta}_j$, we have (note that the smallest nonzero value for $\boldsymbol{\eta}_j$ is $\varepsilon/2$)

$$\mathbf{E}\big[\boldsymbol{\eta}_j \,|\, \mathcal{A}\big] = \int_0^1 \mathbf{Pr}\left[\boldsymbol{\eta}_j \ge x \,|\, \mathcal{A}\right] dx \le (\varepsilon/2) + \int_{\varepsilon/2}^1 \mathbf{Pr}\left[\boldsymbol{\eta}_j \ge x \,|\, \mathcal{A}\right] dx \tag{36}$$

By the definition of $\boldsymbol{\eta}_j$ we have for any $x : \varepsilon/2 \le x \le 1$:

$$\mathbf{Pr}\left[\boldsymbol{\eta}_j \ge x \,|\, \mathcal{A}\right] \le \mathbf{Pr}\left[\boldsymbol{d}_j \le (\varepsilon/x) \text{ and } \mathbf{V}^{(j)} \text{ is compatible with } \mathcal{A} \,|\, \mathcal{A}\right].$$

It follows from (35) that $\mathbf{Pr}\,[\boldsymbol{\eta}_j \geq x \,|\, \mathcal{A}] \leq (\varepsilon/x)\log^5 n + O_h(1/d)$. Continuing from (36):

$$\mathbf{E}\big[\boldsymbol{\eta}_j \,|\, \mathcal{A}\big] \leq (\varepsilon/2) + \int_{\varepsilon/2}^{1} \Big((\varepsilon/x)\log^5 n + O_h(1/d)\Big)\, dx = O_h\big(1/d + \varepsilon \log^6 n\big),$$

since $\varepsilon = n^{4/h-1/2}$. As a consequence, we have for any $j \in [2 : h+1]$,

$$\mathbf{E}\,[\boldsymbol{\eta}_1 \ldots \boldsymbol{\eta}_j] = \mathbf{E}\Big[\boldsymbol{\eta}_1 \cdots \boldsymbol{\eta}_{j-1} \cdot \mathbf{E}\,\big[\boldsymbol{\eta}_j \,|\, \mathbf{V}^{(1)}, \ldots, \mathbf{V}^{(j-1)}\big]\Big] = O_h\big(1/d + \varepsilon \log^6 n\big) \cdot \mathbf{E}\,[\boldsymbol{\eta}_1 \cdot \ldots \boldsymbol{\eta}_{j-1}].$$

This finishes the proof of the lemma. □

6.1 Proof of Proposition 6.3

Recall that $\mathbf{Q}^{(i)}$ denotes the following random variable that is very close to $\mathbf{R}_{-i}$:

$$\mathbf{Q}^{(i)} = \sum_{j=1}^{i} \boldsymbol{v}_j \mathcal{X}^{(j)} + \sum_{j=i+1}^{n} \boldsymbol{u}_j \mathcal{X}^{(j)}.$$

To prove Proposition 6.3, it suffices to show that

$$\mathbf{Pr}\,\big[(\mathbf{Q}^{(i)})|_I \in \mathcal{B}_I^*\big] \leq O\left(\prod_{j=2}^{h+1} \eta_j\right) + O\left(\frac{1}{n^h}\right), \tag{37}$$

where $\mathcal{B}_I^*$ denotes the origin-centered $(\#I)$-dimensional box $[-2\varepsilon, 2\varepsilon]^{\#I}$. This is because the entry-by-entry difference between $\mathbf{Q}^{(i)}$ and $\mathbf{R}_{-i}$ is at most $\beta\delta$, so we just need to make the box $\mathcal{B}_I^*$ bigger than the original box $\mathcal{B}_I$ in Proposition 6.3 (since $2\varepsilon > \varepsilon + 2\beta\delta$).

We prove (37) in the rest of the section. The claim is trivial if there is a $j \geq 2$ such that $V^{(j)}$ is incompatible with $\{V^{(1)}, \ldots, V^{(j-1)}\}$: When this happens the LHS of (21) can be upper bounded by $1/n^{\omega(1)}$ using Lemma 5.5.

Assume from now on that $V^{(j)}$ is compatible with $\{V^{(1)}, \ldots, V^{(j-1)}\}$ for all $j \in [2 : h+1]$. Let L denote the set of $j \in [2 : h+1]$ such that $\eta_j < 1$. Then we have

$$\prod_{j=2}^{h+1} \eta_j = \prod_{j \in L} \eta_j.$$

When $L = \emptyset$, (37) is trivial since the product of η_j's is 1. From now on, we assume that $t = |L| > 1$ and let $L = \{j_1, \ldots, j_t\}$, with $j_1 < \cdots < j_t$. For each $i \in [t]$, let

$$\gamma_i = d_{\ell_2}\big(V^{(j_i)}, \text{span}\{V^{(1)}, V^{(j_1)}, \ldots, V^{(j_{i-1})}\}\big),$$

with $\gamma_1 = d_{\ell_2}(V^{(j_1)}, \text{span}\{V^{(1)}\})$ when $i = 1$ (note that $j_1 \geq 2$). Using

$$\gamma_i = d_{\ell_2}\big(V^{(j_i)}, \text{span}\{V^{(1)}, V^{(j_1)}, \ldots, V^{(j_{i-1})}\}\big) \geq d_{\ell_2}\big(V^{(j_i)}, \text{span}\{V^{(1)}, V^{(2)} \ldots, V^{(j_{i-1})}\}\big) = d_{j_i},$$

we have

$$\prod_{i=1}^{t} \gamma_i \geq \prod_{i=1}^{t} d_{j_i} = \prod_{j \in L} \frac{\varepsilon}{\eta_j} = \varepsilon^t \cdot \prod_{j=2}^{h+1} \frac{1}{\eta_j} > 0. \tag{38}$$

Let A denote the $(t+1) \times n$ matrix whose row vectors are $V^{(1)}, V^{(j_1)}, \ldots, V^{(j_t)}$. Then

Lemma 6.4. *Matrix A has full rank $t+1$, and*

$$\det\left(AA^T\right) \geq \left(\prod_{i=1}^{t} \gamma_i\right)^2 .$$

Proof. It follows directly from (38) that A has full rank.

Next we exhibit a series of transformations $U_1, \ldots, U_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ with the following properties: (1) $\det(U_k) = 1$ for all $k \in [t+1]$; and (2) for each $k \in [t+1]$, the off-diagonal entries of the $k \times k$ principal minor of matrix

$$(U_1 \cdot \ldots \cdot U_k) \cdot A \cdot A^T \cdot (U_1 \cdot \ldots \cdot U_k) \tag{39}$$

are all zero and the i-th diagonal entry is at least γ_i^2 for all $i \in [k]$. Taking $k = t+1$ and recalling that $\det(A \cdot B) = \det(A) \cdot \det(B)$, this gives the claim.

To exhibit these matrices, we simply take U_k to be the lower-triangular matrix corresponding to the k-th step of the Gram-Schmidt orthogonalization of the first k rows of A. This matrix has determinant 1, and after its action, (1) the (k,k) entry of the matrix (39) becomes at least γ_k^2 (by definition of γ_k); and (2) the off-diagonal entries of the $k \times k$ principal minor are 0 as claimed (by the nature of Gram-Schmidt orthogonalization). □

To prove (37) it suffices to show that

$$\mathbf{Pr}\left[(\mathbf{Q}^{(i)})|_{\{1\}\cup L} \in \mathcal{B}'\right] = O\left(\prod_{j=2}^{h+1} \eta_j\right) + O\left(\frac{1}{n^h}\right), \tag{40}$$

where $\mathcal{B}' = [-2\varepsilon, 2\varepsilon]^{t+1}$. Below we let $\mathbf{Q}$ denote $(\mathbf{Q}^{(i)})|_{\{1\}\cup L}$ for convenience.

For this purpose, we let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ denote independent Gaussian random variables $\mathcal{N}(\mu(\ell), 1)$ with the same first $\ell = h^3$ moments as both $\boldsymbol{u}$ and $\boldsymbol{v}$ (recall the first paragraph of Section 4.3). Let

$$\mathbf{G} = \sum_{j=1}^{n} \boldsymbol{w}_j A^{(j)},$$

where $A^{(j)}$ denotes the jth column of A. We prove (40) using the following lemma:

Lemma 6.5. *Let $\eta = \prod_{j=2}^{h+1} \eta_j$. Then the two random variables $\mathbf{Q}^{(i)}$ and $\mathbf{G}$ satisfy*

$$\mathbf{Pr}[\mathbf{G} \in \mathcal{B}'] \leq O(\varepsilon\eta) \quad \textit{and} \quad \left|\mathbf{Pr}[\mathbf{Q} \in \mathcal{B}'] - \mathbf{Pr}[\mathbf{G} \in \mathcal{B}']\right| < O(\varepsilon\eta) + O\big(1/n^h\big).$$

Proposition 6.3 then follows (note that we gave out a factor of ε in the first term for free).

Proof of Lemma 6.5. For the first part, we calculate the covariance matrix of $\mathbf{G}$. Let $i_1, i_2 \in [t+1]$:

$$\mathbf{Cov}\left[\boldsymbol{w}_j A^{(j)}\right]_{i_1,i_2} = \mathbf{E}\big[(\boldsymbol{w}_j A^{(j)}_{i_1} - \mu A^{(j)}_{i_1})(\boldsymbol{w}_j A^{(j)}_{i_2} - \mu A^{(j)}_{i_2})\big] = A^{(j)}_{i_1} A^{(j)}_{i_2}.$$

So we have $\mathbf{Cov}[\mathbf{G}]_{i_1,i_2} = \sum_j A^{(j)}_{i_1} A^{(j)}_{i_2}$ and thus, $\mathbf{Cov}[\mathbf{G}] = AA^T$. The first part of the lemma then follows directly from Lemma 6.4 and the definition of (the density of) multidimensional Gaussian distributions.

For the second part we apply Proposition 4.2 and the Lindeberg method over $\mathbf{Q}^{(i)}$ and $\mathbf{G}$, with

$$\mathcal{A} = \mathcal{B}' = [-2\varepsilon, 2\varepsilon]^{t+1}, \quad \mathcal{A}_{in} = [-2\varepsilon + \xi, 2\varepsilon - \xi]^{t+1}, \quad \text{and} \quad \mathcal{A}_{out} = [-2\varepsilon - \xi, 2\varepsilon + \xi]^{t+1},$$

for some parameter $\xi : 0 < \xi < 2\varepsilon$ to be specified later. We use the following two mollifiers:

Proposition 6.6. *For all $\varepsilon, \xi > 0$ with $\xi < 2\varepsilon$, there exist two smooth functions $\Psi_{in}, \Psi_{out} : \mathbb{R}^{t+1} \to [0,1]$ with the following properties:*

1. *$\Psi_{in}(X) = 0$ for all $X \notin \mathcal{A}$ and $\Psi_{in}(X) = 1$ for all $X \in \mathcal{A}_{in}$.*
2. *$\Psi_{out}(X) = 0$ for all $X \notin \mathcal{A}_{out}$ and $\Psi_{out}(X) = 1$ for all $X \in \mathcal{A}$.*
3. *For any multi-index $J \in \mathbb{N}^{t+1}$ such that $|J| = k$, $\|\Psi_{in}^{(J)}\|_\infty, \|\Psi_{out}^{(J)}\|_\infty \le \alpha(k) \cdot (1/\xi)^k$.*

Proof. Let $\Phi_\xi : \mathbf{R} \to [0,1]$ denote the smooth function given in Claim 4.5 (note that we replaced ε with ξ in Claim 4.5). Let $\Phi_{in}, \Phi_{out} : \mathbf{R} \to [0,1]$ denote the following two smooth functions:

$$\Phi_{in}(x) = \begin{cases} \Phi_\xi(-x + 2\varepsilon) & \text{when } x \ge 0 \\ \Phi_\xi(x + 2\varepsilon) & \text{when } x < 0 \end{cases} \quad \text{and} \quad \Phi_{out}(x) = \begin{cases} \Phi_\xi(-x + 2\varepsilon + \xi) & \text{when } x \ge 0 \\ \Phi_\xi(x + 2\varepsilon + \xi) & \text{when } x < 0 \end{cases}$$

Using Φ_{in} and Φ_{out}, we let

$$\Psi_{in}(X) = \prod_{j \in [t+1]} \Phi_{in}(X_j) \quad \text{and} \quad \Psi_{out}(X) = \prod_{j \in [t+1]} \Phi_{out}(X_j).$$

The three conditions on Ψ_{in} and Ψ_{out} follow from a proof similar to that of Proposition 4.4. □

Next from Proposition 4.2, we have

$$\big|\mathbf{Pr}[\mathbf{Q} \in \mathcal{A}] - \mathbf{Pr}[\mathbf{G} \in \mathcal{A}]\big| \le \max\big\{\big|\mathbf{E}[\Psi_{in}(\mathbf{Q})] - \mathbf{E}[\Psi_{in}(\mathbf{G})]\big|, \big|\mathbf{E}[\Psi_{out}(\mathbf{Q})] - \mathbf{E}[\Psi_{out}(\mathbf{G})]\big|\big\} \\ + \max\big\{\mathbf{Pr}[\mathbf{G} \in \mathcal{A}_{out} \setminus \mathcal{A}], \mathbf{Pr}[\mathbf{G} \in \mathcal{A} \setminus \mathcal{A}_{in}]\big\}.$$

As $\mathbf{G}$ is a Gaussian distribution with $\mathbf{Cov}[\mathbf{G}] = AA^T$, we have

$$\max\big\{\mathbf{Pr}[\mathbf{G} \in \mathcal{A}_{out} \setminus \mathcal{A}], \mathbf{Pr}[\mathbf{G} \in \mathcal{A} \setminus \mathcal{A}_{in}]\big\} \le \mathbf{Pr}[\mathbf{G} \in \mathcal{A}_{out}] = O\left(\frac{(4\varepsilon + 2\xi)^{t+1}}{\varepsilon^t} \cdot \prod_{j=2}^{h+1} \eta_j\right),$$

where we have used (38), Lemma 6.4, and the definition of the density of multidimensional Gaussian distributions. To bound $|\mathbf{E}[\Psi_{in}(\mathbf{Q})] - \mathbf{E}[\Psi_{in}(\mathbf{G})]|$, we apply Lindeberg's method again and follow the same argument as in 4.2 but this time match *all the first ℓ moments*, which gives us that

$$\big|\mathbf{E}[\Psi_{in}(\mathbf{Q})] - \mathbf{E}[\Psi_{in}(\mathbf{G})]\big| \le n \cdot \frac{O_\ell(1)}{\xi^{\ell+1}} \cdot \frac{1}{n^{(\ell+1)/2}}.$$

The same bound also holds for $|\mathbf{E}[\Psi_{out}(\mathbf{Q})] - \mathbf{E}[\Psi_{out}(\mathbf{G})]|$ by the same argument.

Combining all these inequalities, we have

$$\big|\mathbf{Pr}[\mathbf{Q} \in \mathcal{A}] - \mathbf{Pr}[\mathbf{G} \in \mathcal{A}]\big| \le O\left(\frac{(4\varepsilon + 2\xi)^{t+1}}{\varepsilon^t} \cdot \prod_{j=2}^{h+1} \eta_j\right) + n \cdot \frac{O_h(1)}{\xi^{\ell+1}} \cdot \frac{1}{n^{(\ell+1)/2}}.$$

Setting $\xi = n^{2/h^2 - 1/2} < \varepsilon$, we have

$$n \cdot \frac{O_h(1)}{\xi^{\ell+1}} \cdot \frac{1}{n^{(\ell+1)/2}} = O_h\left(\frac{1}{n^{2h-1}}\right) = O_h\left(\frac{1}{n^h}\right) \quad \text{and} \quad \frac{(4\varepsilon + 2\xi)^{t+1}}{\varepsilon^t} < 6^{t+1} \cdot \varepsilon = O_h(\varepsilon).$$

This finishes the proof of the lemma. □

7 Putting all the pieces together

Finally we put all the pieces together and prove our main theorem.

Proof of Theorem 1. Fix a $c > 0$. Recall that $h = h(c) \geq 5/c$, $\ell = h^3$ and $\varepsilon = n^{4/h-1/2}$.

Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ denote independent random variables with the same distribution as $\boldsymbol{u}$ and $\boldsymbol{v}$ as given in Proposition 3.1 and 3.2, respectively, with parameters ℓ and $\mu = \mu(\ell)$. Using Theorem 9 in Appendix B, a function drawn from $\mathcal{D}_{no}$ is $\kappa(c)$-far from monotone, with probability $1 - o_n(1)$, for some constant distance parameter $\kappa(c)$ that depends on c only. Thus, to prove that a non-adaptive algorithm for monotonicity testing requires $\Omega(n^{1/2-c})$ queries, it suffices to bound

$$d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) \leq 0.1, \quad \text{for all query sets } \mathcal{X} \in \{\pm 1/\sqrt{n}\}^n \text{ with } |\mathcal{X}| \leq n^{1/2-c},$$

where $\mathbf{S} = \sum_j \boldsymbol{u}_j \mathcal{X}^{(j)}$ and $\mathbf{T} = \sum_j \boldsymbol{v}_j \mathcal{X}^{(j)}$.

Let $\mathcal{X}$ denote a query set with size at most $n^{1/2-c}$. It follows from Lemma 5.7 that there exists a scattered query set $\mathcal{X}^* \subseteq \mathcal{X}$ such that $d = |\mathcal{X}^*| \leq |\mathcal{X}| \leq n^{1/2-c}$ and

$$d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) \leq d_{\mathrm{UO}}(\mathbf{S}^*, \mathbf{T}^*) + 0.01,$$

where $\mathbf{S}^* = \sum \boldsymbol{u}_j \mathcal{X}^{*(j)}$ and $\mathbf{T}^* = \sum \boldsymbol{v}_j \mathcal{X}^{*(j)}$. Let

$$\mathbf{Q}^{*(i)} = \sum_{j=1}^{i} \boldsymbol{v}_j \mathcal{X}^{*(j)} + \sum_{j=i+1}^{n} \boldsymbol{u}_j \mathcal{X}^{*(j)},$$

Combining (18), (19), and Lemma 6.1, we have

$$\big| \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{*(i-1)})] - \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{Q}^{*(i)})] \big| \leq \frac{O_h(1)}{n^{(h+1)/2}} \cdot \frac{1}{\varepsilon^{h+1}} \cdot \Big(d^{h+1} \cdot \big(1/d + \varepsilon \log^6 n\big)^h \Big),$$

and hence as in Section 4.2, summing over all $i \in [n]$ gives that

$$|\, \mathbf{E}[\Psi_{\mathcal{O}}(\mathbf{S}^*) - \Psi_{\mathcal{O}}(\mathbf{T}^*)| = \frac{O_h(1)}{n^{(h-1)/2} \cdot \varepsilon^{h+1}}.$$

Since $d \leq n^{1/2-c}$, we have $d\varepsilon \ll 1/\log^6 n$. By Proposition 4.1 (and the 1-dimensional Berry-Esseen inequality (Theorem 3) together with a union bound across the d dimensions), we have

$$d_{\mathrm{UO}}(\mathbf{S}^*, \mathbf{T}^*) \leq O(d\varepsilon) + O(d/\sqrt{n}) + \frac{O_h(1)}{n^{(h-1)/2} \cdot \varepsilon^{h+1}} \cdot d = o(1).$$

It follows that $d_{\mathrm{UO}}(\mathbf{S}, \mathbf{T}) < 0.1$. This finishes the proof of the theorem. □

References

[AC06] N. Ailon and B. Chazelle. Information theory in property testing and monotonicity testing in higher dimension. *Information and Computation*, 204:1704–1717, 2006. 1.1

[Akh65] Naum Akhiezer. *The Classical Moment Problem.* Hafner, New York, 1965. 3.2

[BCGSM12] Jop Briët, Sourav Chakraborty, David García-Soriano, and Arie Matsliah. Monotonicity testing and shortest-path routing on the cube. *Combinatorica*, 32(1):35–53, 2012. 1.1

[Ben03] Vidmantas Bentkus. On the dependence of the Berry–Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003. 4.1, 4.2

[CS13] Deeparnab Chakrabarty and C. Seshadhri. A $o(n)$ monotonicity tester for boolean functions over the hypercube. In *ACM Symposium on Theory of Computing*, pages 411–418, 2013. 1.1

[CST14] Xi Chen, Rocco A. Servedio, and Li-Yang Tan. New algorithms and lower bounds for testing monotonicity. To appear in FOCS 2014, 2014. (document), 1.1, 1.1, 1.2, 1.2, 4, B

[DDS13] Anindya De, Ilias Diakonikolas, and Rocco A. Servedio. A Robust Khintchine Inequality, and Algorithms for Computing Optimal Constants in Fourier Analysis and High-Dimensional Geometry. In *Proceedings of the 40th Annual International Colloquium on Automata, Languages and Programming*, pages 376–387, 2013. Full version at arxiv:1207.2229. B

[DGL$^+$99] Yevgeniy Dodis, Oded Goldreich, Eric Lehman, Sofya Raskhodnikova, Dana Ron, and Alex Samorodnitsky. Improved testing algorithms for monotonocity. In *Proceedings of RANDOM*, pages 97–108, 1999. 1.1

[Fel68] William Feller. *An introduction to probability theory and its applications*, volume 1. Wiley, 3rd edition, 1968. 2

[FLN$^+$02] Eldar Fischer, Eric Lehman, Ilan Newman, Sofya Raskhodnikova, Ronitt Rubinfeld, and Alex Samorodnitsky. Monotonicity testing over general poset domains. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pages 474–483, 2002. 1.1

[GGL$^+$00] Oded Goldreich, Shafi Goldwasser, Eric Lehman, Dana Ron, and Alex Samordinsky. Testing monotonicity. *Combinatorica*, 20(3):301–337, 2000. 1.1

[GGLR98] Oded Goldreich, Shafi Goldwasser, Eric Lehman, and Dana Ron. Testing monotonicity. In *IEEE Symposium on Foundations of Computer Science*, pages 426–435, 1998. 1.1

[GOWZ10] Parikshit Gopalan, Ryan O'Donnell, Yi Wu, and David Zuckerman. Fooling functions of halfspaces under product distributions. In *Proceedings of the 25th Annual IEEE Conference on Computational Complexity*, pages 223–234, 2010. 4, 4.2

[HK08] Shirley Halevy and Eyal Kushilevitz. Testing monotonicity over graph products. *Random Struct. Algorithms*, 33(1):44–67, 2008. 1.1

[KNW10] Daniel Kane, Jelani Nelson, and David Woodruff. On the exact space complexity of sketching and streaming small norms. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1161–1178, 2010. A

[MORS10] Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco Servedio. Testing halfspaces. *SIAM Journal on Computing*, 39(5):2004–2047, 2010. B

[Mos08] Elchanan Mossel. Gaussian bounds for noise correlation of functions and tight analysis of Long Codes. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 156–165, 2008. 4.2

[Odl88] Andrew Odlyzko. On subspaces spanned by random selections of ±1 vectors. *J. Combinatorial Theory A*, 47:124–133, 1988. 5.1

[Sch50] L. Schläfli. Gesammelte mathematische abhandlugen. *Band 1, Birkhauser, Basel*, 1850. 2.1

[VV11] Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *ACM Symposium on Theory of Computing*, pages 685–694, 2011. 1.2, 1, 1.3, 4

A Standard mollifier construction

In this section we prove Claim 4.5. We begin with the following fact (see [KNW10] for a reference).

Fact A.1. *There is a smooth function $b : \mathbb{R} \to [0,1]$ such that*

(i) If $|x| > 1$, then $b(x) = 0$.

(ii) For all $\ell > 0$, $\|b^{(\ell)}\|_\infty \le e \cdot 32^\ell \cdot \ell! \cdot \ell^{2\ell+2}$.

(iii) $\int_{-\infty}^{\infty} b(x)\,dx = 1$.

Note that the bound on $\|b^{(\ell)}\|_\infty$ from (ii) above is $2^{-\ell-1} \cdot \alpha(\ell)$. We now restate Claim 4.5.

Claim. *For all $\varepsilon > 0$, there is a smooth function $\Phi_\varepsilon : \mathbb{R} \to [0,1]$ which satisfies:*

(1) If $x < 0$, then $\Phi_\varepsilon(x) = 0$.

(2) If $x > \varepsilon$, then $\Phi_\varepsilon(x) = 1$.

(3) $\|\Phi_\varepsilon^{(k)}\|_\infty \le \alpha(k)/\varepsilon^k$.

Proof. First, we define the function $b_\varepsilon : \mathbb{R} \to [0, 2/\varepsilon]$ as

$$b_\varepsilon(x) = \frac{2}{\varepsilon} \cdot b\left(\frac{2x}{\varepsilon}\right).$$

Observe that as a consequence, we have that b_ε is smooth; $b_\varepsilon(x) = 0$ if $|x| > \varepsilon/2$; $\int_{-\infty}^{\infty} b_\varepsilon(x)\,dx = 1$.

Further, taking the kth derivative of b_ε, we have

$$\frac{d^k b_\varepsilon(x)}{dx^k} = \frac{2^{k+1}}{\varepsilon^{k+1}} \cdot \left.\frac{d^k b(y)}{dy^k}\right|_{y=2x/\varepsilon}$$

As a result, we get that $\|b_\varepsilon^{(k)}\|_\infty \le \alpha(k)/\varepsilon^{k+1}$. Let us define $g : \mathbb{R} \to \{0,1\}$ as

$$g(x) = \begin{cases} 1 & \text{if } x > \varepsilon/2 \\ 0 & \text{otherwise.} \end{cases}$$

We define $\Phi_\varepsilon = b_\varepsilon * g$. Since $b_\varepsilon \in \mathcal{C}^\infty$ we have that $\Phi_\varepsilon \in \mathcal{C}^\infty$. To see that conditions (1) and (2) of Claim 4.5 hold, we note that

$$\Phi_\varepsilon(x) = \int_{-\infty}^{\infty} b_\varepsilon(y) \cdot g(x-y) \cdot dy = \int_{-\varepsilon/2}^{\varepsilon/2} b_\varepsilon(y) \cdot g(x-y) \cdot dy.$$

Note that if $x < 0$, then $g(x-y) \neq 0$ implies that $y < -\varepsilon/2$. However, for $y < -\varepsilon/2$, $b_\varepsilon(y) = 0$. This proves (1). If $x > \varepsilon$, then for all $|y| \le \varepsilon/2$, $g(x-y) = 1$. Using the fact that $b_\varepsilon(x)$ is a density, we get (2). Thus, it only remains to prove (3). For any $x \in \mathbb{R}$, we have

$$\Phi_\varepsilon^{(k)}(x) = \frac{d^k \Phi_\varepsilon(x)}{dx^k} = \left(b_\varepsilon^{(k)} * g\right)(x) = \int_{-\infty}^{\infty} b_\varepsilon^{(k)}(y) \cdot g(x-y) \cdot dy = \int_{-\varepsilon/2}^{\varepsilon/2} b_\varepsilon^{(k)}(y) \cdot g(x-y) \cdot dy.$$

Since $\|g\|_\infty = 1$ and the interval length of the integration is ε, we get that

$$\left\|\Phi_\varepsilon^{(k)}\right\|_\infty \le \varepsilon \cdot \left\|b_\varepsilon^{(k)}\right\|_\infty \le \alpha(k)/\varepsilon^k.$$

This completes the proof of the claim. □

B Distance to monotonicity for functions from $\mathcal{D}_{no}$

In this section we prove the following theorem:

Theorem 9. *Let $\mathcal{D}_{no}$ be the distribution over functions $\boldsymbol{f}(X) = \mathrm{sign}(\boldsymbol{v}_1 X_1 + \cdots + \boldsymbol{v}_n X_n)$ where each $\boldsymbol{v}_i$ is distributed according to $\boldsymbol{v}$ given in Proposition 3.2 with $\ell = h^3, h = h(c)$ as described in Section 3. Then with probability $1 - o_n(1)$ over a draw of $\boldsymbol{f}$ from $\mathcal{D}_{no}$, the function $\boldsymbol{f}$ is $\Omega_c(1)$-far from monotone.*

As noted in Section 3, this theorem can be proved using the methods of [CST14]; for the sake of completeness we give an alternate proof here.

Our proof uses the following claims; in all of them, the hidden constants will depend on c. We will need the notion of τ-regular LTFs, which we define below.

Definition 10. An LTF $f = \mathrm{sign}(v_1 X_1 + \cdots + v_n X_n)$ is said to be *τ-regular* if $|v_i|/\sqrt{\sum_{i=1}^n v_i^2} \leq \tau$ for all $i \in [n]$.

Note that as defined above, the notion of regularity refers to a representation of an LTF and not the LTF *per se*. For this section, we will blur this distinction and refer to an LTF being τ-regular as long as it has a τ-regular representation.

Claim B.1. *We have $\mathbf{Pr}_{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n}[|\{i : \boldsymbol{v}_i < 0\}| = \Omega(n)] = 1 - o_n(1)$. As a consequence, we also have*

$$\mathbf{Pr}_{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n}\left[\sum_{i=1}^n v_i^2 \cdot \mathbf{1}[v_i < 0] = \Omega(n)\right] = 1 - o_n(1).$$

Proof. The first equation follows from item (1) in Proposition 3.2 and an application of Chernoff bound. The second equation is immediate from the first equation and the fact that the support of $\boldsymbol{v}$ is bounded (and independent of n). □

Claim B.2. *A function $\boldsymbol{f} \sim \mathcal{D}_{no}$ is $O(1/\sqrt{n})$-regular with probability $1 - o_n(1)$.*

Proof. This follows from the first part of Claim B.1 and the fact that the support of $\boldsymbol{v}$ is bounded (and independent of n). □

Thus far we have been implicitly assuming the domain to be $\{-1,1\}^n$, but we may also consider the domain $\mathbb{R}^n$. We recall the standard definition of the degree-1 Hermite coefficient of a function $f : \mathbb{R}^n \to \mathbb{R}$ given by an index $i \in [n]$:

$$\tilde{f}(i) = \mathbf{E}_{\mathbf{X} \sim \mathcal{N}^n(0,1)}[f(\mathbf{X}) \cdot \mathbf{X}_i].$$

We recall the following fact that is proved in [MORS10] (Proposition 25).

Fact B.3. *For an LTF $f = \mathrm{sign}(v_1 X_1 + \ldots + v_n X_n)$, we have*

$$\tilde{f}(i) = \sqrt{\frac{2}{\pi}} \cdot \frac{v_i}{\sqrt{\sum_{i=1}^n v_i^2}}.$$

We also recall the following theorem from [DDS13] (Theorem 57).

Theorem. *If f as defined above is τ-regular, then*

$$\sum_{i=1}^{n}(\tilde{f}(i) - \widehat{f}(i))^2 = O(\tau^{1/6}).$$

Combining the above theorem with Fact B.3 and Claim B.2, we get that

$$\sum_{i:v_i<0} \widehat{f}(i)^2 = \frac{2}{\pi} \cdot \frac{\sum_{i:v_i<0} v_i^2}{\sum_i v_i^2} - o_n(1).$$

Combining with the second part of Claim B.1, we have

$$\sum_{i:v_i<0} \widehat{f}(i)^2 = \Omega(1). \tag{41}$$

Proof of Theorem 9. Let $g : \{-1,1\}^n \to \{-1,1\}$ be any monotone function. As is well known, for any $i \in [n]$ we have $\widehat{g}(i) \geq 0$. We thus have

$$\begin{aligned}\mathbf{Pr}_{x\in\{-1,1\}^n}[f(x) \neq g(x)] = \frac{1}{4} \cdot \mathbf{E}[(f(x) - g(x))^2] &\geq \frac{1}{4} \cdot \left(\sum_{i=1}^{n}(\widehat{f}(i) - \widehat{g}(i))^2\right) \\ &\geq \sum_{i:v_i<0} (\widehat{f}(i))^2 = \Omega(1).\end{aligned}$$

Here the first inequality follows by Parseval's identity while the last one uses (41). □

C Determinant of $B^{(\ell)}$

Recall that $B^{(\ell)}$ is an $r \times r$ square matrix with $r = (\ell+1)/2$, whose (i,j)th entry is $(2(i+j)-3)!!$. Here we prove by induction on odd ℓ (as $B^{(\ell)}$ is only defined over odd ℓ) that

$$\det(B^{(\ell)}) = \prod_{j \text{ odd},\, j\in[\ell]} j!. \tag{42}$$

The base case of $\ell = 1$ is trivial. Now assume for induction that the equation holds for $\ell - 2$. Given $B^{(\ell)}$, we perform the following sequence of linear transformations:

For each j from r down to 2, subtract $\big[(2(r+j)-3) \times \text{column } (j-1)\big]$ from column j.

Let A denote the new $r \times r$ matrix. Note that $\det(A) = \det(B^{(\ell)})$. Then it is easy to verify that the last row of A is all zero except the $(r,1)$th entry, which is $(2r-1)!! = \ell!!$; the (i,j)th entry of A, $i \in [r-1]$ and $j \in [2:r]$, is

$$(2(i+j)-3)!! - (2(i+j-1)-3)!! \cdot (2(r+j)-3) = (2(i+j-1)-3)!! \cdot (2i-2r),$$

which is $(2i-2r)$ times the $(i,j-1)$th entry of $B^{(\ell-2)}$.

This implies that the upper right $[r-1] \times [2:r]$ submatrix of A is $B^{(\ell-2)}$ after scaling the rows by $-(2r-2), -(2r-4), \ldots, -2$, respectively. As a result, we have

$$\det(B^{(\ell)}) = \det(A) = (-1)^{r+1} \cdot \ell!! \cdot \prod_{i\in[r-1]} (2i-2r) \cdot \det(B^{(\ell-2)}) = \ell! \cdot \det(B^{(\ell-2)}).$$

We obtain (42) after plugging in the inductive hypothesis for $B^{(\ell-2)}$. This finishes the induction.