Testing Noisy Linear Functions for Sparsity

Xue Chen
xue.chen1@northwestern.edu
Northwestern University
Evanston, Illinois, USA

Anindya De*
anindyad@cis.upenn.edu
University of Pennsylvania
Philadelphia, Pennsylvania, USA

Rocco A. Servedio†
rocco@cs.columbia.edu
Columbia University
New York, New York, USA

ABSTRACT

We consider the following basic inference problem: there is an unknown high-dimensional vector $w \in \mathbb{R}^n$, and an algorithm is given access to labeled pairs (x, y) where $x \in \mathbb{R}^n$ is a measurement and $y = w \cdot x +$ noise. What is the complexity of deciding whether the target vector w is (approximately) k-sparse? The recovery analogue of this problem — given the promise that w is sparse, find or approximate the vector w — is the famous *sparse recovery* problem, with a rich body of work in signal processing, statistics, and computer science.

We study the decision version of this problem (i.e. deciding whether the unknown w is k-sparse) from the vantage point of *property testing*. Our focus is on answering the following high-level question: when is it possible to efficiently *test* whether the unknown target vector w is sparse versus far-from-sparse using a number of samples which is *completely independent* of the dimension n? We consider the natural setting in which x is drawn from an i.i.d. product distribution $\mathcal{D}$ over $\mathbb{R}^n$ and the noise process is independent of the input x. As our main result, we give a general algorithm which solves the above-described testing problem using a number of samples which is completely independent of the ambient dimension n, as long as $\mathcal{D}$ is not a Gaussian. In fact, our algorithm is *fully noise tolerant*, in the sense that for an arbitrary w, it approximately computes the distance of w to the closest k-sparse vector. To complement this algorithmic result, we show that weakening any of our conditions makes it information-theoretically impossible for *any* algorithm to solve the testing problem with fewer than essentially $\log n$ samples. Thus our conditions essentially characterize when it is possible to test noisy linear functions for sparsity with constant sample complexity.

Our algorithmic approach is based on relating the cumulants of the output distribution (i.e. of y) with elementary power sum symmetric polynomials in w and using the latter to measure the sparsity of w. This approach crucially relies on a theorem of Marcinkiewicz from probability theory. In fact, to obtain effective sample complexity bounds with our approach, we prove a new finitary version of Marcinkiewicz's theorem. This involves extending the complex analytic arguments used in the original proof with results about the distribution of zeros of entire functions.

CCS CONCEPTS

• **Theory of computation → Streaming, sublinear and near linear time algorithms**.

KEYWORDS

Property testing, sparse recovery, cumulants.

ACM Reference Format:
Xue Chen, Anindya De, and Rocco A. Servedio. 2020. Testing Noisy Linear Functions for Sparsity. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC '20), June 22–26, 2020, Chicago, IL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3357713.3384239

1 INTRODUCTION

This paper addresses a basic data analysis problem from the perspective of *property testing*. To motivate our work, we begin by considering the following simple and fundamental inference problem: For independent uniform strings $\boldsymbol{x} \sim \{-1, 1\}^n$ and Gaussian noise $\boldsymbol{\eta} \sim N(0, 0.1)$, an algorithm gets access to labeled samples of the form $(\boldsymbol{x}, \boldsymbol{y})$ where $\boldsymbol{y} = w \cdot \boldsymbol{x} + \boldsymbol{\eta}$ and w is some fixed but unknown unit vector in $\mathbb{R}^n$. The task of recovering w from these noisy samples is an instance of the standard *linear regression* problem, which is of course very well studied in computer science, econometrics, and statistics (see e.g. [23, 28] or many other references). As is well known, $\Theta(n)$ samples are both necessary and sufficient to recover w (to within a small constant error), and the ordinary least squares algorithm is a computationally efficient algorithm which achieves this sample complexity.

Now suppose that the algorithm is promised that w is *k-sparse*, i.e. it has only k non-zero entries. In this case, the influential line of work on *compressive sensing* shows that much better sample complexities and running times can be achieved. In particular, the breakthrough work of Candes, Romberg and Tao [17] shows that using just $m = O(k \log n)$ samples and running in time $\text{poly}(m, n)$, it is possible to (approximately) recover the k-sparse vector w. Observe that when k is small (like a constant), this is an exponential improvement over the sample complexity achieved by standard linear regression. We further note that by results such as [1, 37], the bound of $O(k \log n)$ samples is essentially tight, and that compressive sensing algorithms are applicable for more general choices of the distribution of $\boldsymbol{x}$ and the noise $\boldsymbol{\eta}$ (see the survey by Candes [14]).

In this paper we consider a natural *decision* analogue of the above problem: the algorithm has access to the same type of $(\boldsymbol{x}, \boldsymbol{y} =$

*Supported by NSF grant CCF-1926872 and CCF-1910534.
†Supported by NSF grants CCF-1814873, IIS-1838154, CCF-1563155, and by the Simons Collaboration on Algorithms and Geometry.

STOC '20, June 22–26, 2020, Chicago, IL, USA

ACM ISBN 978-1-4503-6979-4/20/06…$15.00
https://doi.org/10.1145/3357713.3384239

$w \cdot \boldsymbol{x} + \boldsymbol{\eta}$) samples as above, but it is *not* promised that the target vector w is k-sparse. Rather, the task of the algorithm now is to distinguish between the cases that (i) the target vector w is k-sparse, versus (ii) the target vector w is ε-far from every k-sparse vector w' (for some appropriate notion of "far"). Using algorithms from compressive sensing, it is straightforward to obtain an algorithm with $m = O(k \log n)$ sample complexity and $\mathrm{poly}(m, n)$ runtime. But can one do much better? In particular, it is *a priori* conceivable that there is an algorithm for this decision problem[1] with sample complexity *completely independent of the ambient dimension n*. Do such "ultra-efficient" algorithms in fact exist?

As a corollary of our main algorithmic result, we give an affirmative answer to this question. Our result implies that in the above setting, it is indeed possible to distinguish between w which (i) is k-sparse, versus (ii) is ε-far in ℓ_2 distance from all k-sparse vectors w', with an $m = O_{k,\varepsilon}(1)$ sample complexity that is completely independent of n. In fact, we achieve much more: our algorithm can handle a broad range of distributions of $\boldsymbol{x}$, and in essentially the same sample complexity we can approximate the distance from w to the closest k-sparse vector. Thus we can essentially determine the "fit" of the best k-sparse vector using only $m = O_{k,\varepsilon}(1)$ samples. Moreover, the running time of our algorithm is $\mathrm{poly}(m)$ if it is allowed to skip the reading of $\boldsymbol{x}$ in every sample.

1.1 Motivation: Property Testing

Before describing our main results in more detail, we recall a line of work on *property testing of functions* which strongly motivates our study. In the standard property testing framework, an algorithm is given access to an unknown function f via an oracle O. For a property $\mathcal{P}$ of functions, the goal of a property testing algorithm for $\mathcal{P}$ is to make as few queries to O as possible and distinguish (with success probability, say, 9/10) between the cases that (i) the function f has the property $\mathcal{P}$, versus (ii) the function f is at least ε-far in Hamming distance from every function g with property $\mathcal{P}$. As a well-known example of this framework, the seminal work of [10] showed that when $\mathcal{P}$ is the property of being $\mathrm{GF}(2)$-linear, f is any function from $\mathrm{GF}^n(2)$ to $\mathrm{GF}(2)$, and the oracle O is a black-box oracle for f, then there is an algorithm with query complexity $O(1/\varepsilon)$. We refer the reader to books and surveys such as [25, 26, 38, 39] which give an overview of the nearly three decades of work in this area.

An often-sought-after "gold standard" for property testing algorithms, that can (perhaps surprisingly) be achieved for many problems, is an algorithm with *constant query complexity*, i.e. a query complexity that only depends on the error parameter ε and is completely independent of the ambient dimension n. This, for example, is the case with $\mathrm{GF}(2)$-linearity testing [10], low-degree testing [24], junta testing [20], and other problems. Indeed, there are grand conjectures (and partial results towards them) which seek to characterize all such properties $\mathcal{P}$ which can be tested with a constant number of queries to a black-box oracle (see e.g. [5, 6, 30]).

In this spirit, we explore the question of whether (and when), given noisy labeled samples of the form $(\boldsymbol{x}, \boldsymbol{y})$ where $\boldsymbol{y} = w \cdot \boldsymbol{x} + \boldsymbol{\eta}$, we can test k-sparsity of w with a number of samples that only depends on k and ε, and is independent of n. Before describing our precise model, we point out an important difference between our model and much work on property testing. In the standard model of property testing of functions described above, it is usually assumed that the algorithm can make black-box queries to the unknown function; in contrast, in our model, the algorithm only has "passive" access to random samples. Obtaining dimension-independent guarantees when given only sample access can be quite challenging; for example, the sample complexity of testing $\mathrm{GF}(2)$ linearity in this model is $\Theta(n)$ samples [27] whereas as stated above only $O(1/\varepsilon)$ queries are required by the [10] result. We refer the reader to [18, 27, 31] for some property testing results in the "sample-based" model.

1.2 The Problem We Consider

In order to describe the algorithmic problem that we consider in more detail, let us define the notion of *distance to k-sparsity*. Given a nonzero vector $w \in \mathbb{R}^n$, we define its distance to k-sparsity to be

$$\mathrm{dist}(w, k\text{-sparse}) := \min_{w' \in \mathbb{R}^n:\ w' \text{ is } k\text{-sparse}} \frac{\|w - w'\|_2}{\|w\|_2}; \qquad (1)$$

this is equivalent to the fraction of the 2-norm of w that comes from the coordinates that are not among the k largest-magnitude ones. Note that when w is a unit vector, then $\mathrm{dist}(w, k\text{-sparse})$ is the same as the ℓ_2 distance between w and the closest k-sparse vector.

Basic model: We are now ready to describe our model. We are given access to independent labeled examples of the form $(\boldsymbol{x}, \boldsymbol{y})$ where $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}$. In each such labeled example $\boldsymbol{x}$ is drawn from some distribution $\mathcal{D}$ over $\mathbb{R}^n$ and the label value $\boldsymbol{y}$ is a noise-corrupted version of $w \cdot \boldsymbol{x}$ for some unknown target vector $w \in \mathbb{R}^n$. In particular, $\boldsymbol{y} = w \cdot \boldsymbol{x} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is drawn from some noise distribution (which is independent of $\boldsymbol{x}$). The goal is to distinguish between the following two cases: (i) w is a k-sparse vector (meaning that it has at most k nonzero coordinates), versus (ii) w is *ε-far* from being k-sparse (meaning that $\mathrm{dist}(w, k\text{-sparse}) \geq \varepsilon$). Thus, we are considering a promise problem, or equivalently any output is okay in the intermediate case in which w is not k-sparse but is ε-close to being k-sparse. We refer to this problem as *(non-robust) k-sparsity testing*.

Our algorithms will in fact solve a robust version of this problem: in the same model as above, for any given $\varepsilon > 0$, our algorithms will approximate the value of $\mathrm{dist}(w, k\text{-sparse})$ to within an additive $\pm\varepsilon$. We refer to this problem as *noise tolerant k-sparsity testing* (see Parnas, Ron and Rubinfeld [36]); it is immediate that any algorithm for this noise-tolerant version immediately implies an algorithm with the same complexity for non-robust k-sparsity testing. In fact, while our main algorithmic result is for the noise tolerant problem, our lower bounds (which we describe later) are for the non-robust version (which *a fortiori* makes them applicable to the noise-tolerant version).

Our desideratum: constant-sample testability. As is the case for similar-in-spirit property testing problems such as k-junta testing [7, 13, 21], we view k as a parameter which is fixed relative to n, and our main goal is to obtain a *constant-sample* tester, i.e. a testing algorithm for which the number of samples used is $O_{k,\varepsilon}(1)$

[1] This is in contrast with the recovery problem, as shown by lower bounds such as [1, 37] mentioned above.

completely independent of n. As stressed earlier (unlike the junta testing problem or many other problems studied in Boolean function property testing), our testing algorithms are *not* allowed to "actively" make queries — their only source of information about w is access to the i.i.d. samples $(\boldsymbol{x}, \boldsymbol{y})$ that are generated as described above.

1.3 Our Algorithmic Results

Informally speaking, our main positive result says that for a broad class of input distributions $\mathcal{D}$, if the parameters of the noise are provided then there is a testing algorithm with $O_{k,\varepsilon}(1)$ sample complexity independent of n. Here is a qualitative statement of our main result (Theorem 5.1 gives a more precise version). We start with a description of the algorithmic guarantee for non-robust k-sparsity testing.

Theorem 1.1 (Qualitative statement of main result). *Fix any random variable $\boldsymbol{X}$ over $\mathbb{R}$ which has variance 1, finite moments of every order[2], and is not Gaussian (i.e. its total variation distance from every Gaussian is nonzero). For any n, let $\mathcal{D}$ be the product distribution over $\mathbb{R}^n$ whose marginals are each distributed according to $\boldsymbol{X}$, i.e. $\mathcal{D} \equiv \boldsymbol{X}^n$. Let $\boldsymbol{\eta}$ be a random variable corresponding to a noise distribution over $\mathbb{R}$ which is such that all its moments are finite.*

Then there is an algorithm (depending on $\mathcal{D}$ and $\boldsymbol{\eta}$) with the following property: for any $w \in \mathbb{R}^n$ with[3] $1/C \le \|w\|_2 \le C$, given k, ε, and access to independent samples $(\boldsymbol{x}, \boldsymbol{y} = w \cdot \boldsymbol{x} + \boldsymbol{\eta})$ where each $\boldsymbol{x} \sim \mathcal{D}$,

- *if w is k-sparse then with probability $9/10$ the algorithm outputs "k-sparse"; and*
- *if w is ε-far from k-sparse then with probability $9/10$ the algorithm outputs "far from k-sparse."*

*The number m of samples used by the algorithm depends only on $C, k, \varepsilon, \boldsymbol{X}$ and $\boldsymbol{\eta}$; in particular, it is independent of n. We will refer to such an algorithm as an ε-*tester for k-sparsity under $\mathcal{D}$ and $\boldsymbol{\eta}$ with sample complexity m.

Tolerant testing: As mentioned earlier, our algorithmic guarantees are in fact, much stronger. Namely, under the same conditions on $\mathcal{D}$ and $\boldsymbol{\eta}$ as above, the algorithm in Theorem 1.1, with high probability, in fact computes $\mathrm{dist}(w, k\text{-sparse})$ to an additive $\pm\varepsilon$. Thus for $\mathcal{D}$ and $\boldsymbol{\eta}$ as above, this shows that noise tolerant k-sparsity testing can be done with a constant number of samples.

Remark 1.2 (Explicit bounds and sharper quantitative bounds for "benign" distributions)**.** Theorem 1.1 shows that for every non-Gaussian random variable $\boldsymbol{X}$ the corresponding testing problem has a constant-sample algorithm, but it does not give a uniform upper bound on sample complexity that holds for all non-Gaussian distributions. (Indeed, no such uniform upper bound on sample complexity can exist; see Remark 2.2 for an elaboration of this point.) However, if the background random variable $\boldsymbol{X}$ is supported on a bounded set, say $[-\mathrm{B}, \mathrm{B}]$, then it is in fact possible to get an explicit uniform upper bound on the sample complexity (which is a tower of height $O(k)$). We do this by proving a new finitary version of a theorem due to J. Marcinkiewicz [33] from probability theory. This involves extending the complex analytic arguments used in the original proof; prior to this work, to the best of our knowledge no finitary analogue of the Marcinkiewicz theorem was known [12, 29, 35]. We give this proof in Section 6.

Going beyond Theorem 1.1, we show that for a large class of "benign" distributions (which includes the uniform distribution over $[0, 1]$, any product distribution over $\{-1, 1\}$, and many others), a different and simpler algorithm provides a uniform upper bound on sample complexity, which is roughly $(k/\varepsilon)^{O(k)}$. (See the full version for a detailed statement and proof of this result.)

1.4 Lower Bounds: Qualitative Optimality of Our Algorithmic Results

1.4.1 On the Role of Noise and Its Independence of the Data Points. We begin by addressing the role of noise in our model. Without noise corrupting the labels, when the background random variable $\boldsymbol{X}$ is continuous, even the recovery problem will admit a simple algorithm which uses only $k + 1$ samples (see the full version for an elaboration on this point). Thus, all of our positive results are for settings in which the labels are corrupted by noise. On the other hand, some of our lower bounds are for problem variants in which the labels are noise-free; this of course only makes the corresponding lower bounds stronger.

Secondly, our model (described in Section 1.2) requires that the distribution of the noise $\boldsymbol{\eta}$ is independent of the distribution of $\boldsymbol{x}$. It is easy to see that if the noise process corrupting the label $\boldsymbol{y}$ of a labeled example $(\boldsymbol{x}, \boldsymbol{y})$ is allowed to depend on $\boldsymbol{x}$, then it is possible for the noise to perfectly simulate k-sparsity when the target vector is far from k-sparse or vice versa. In this situation no algorithm, even with infinite sample complexity, can succeed in testing k-sparsity. Thus throughout this work we assume that the noise $\boldsymbol{\eta}$ in each labeled example is independent of the example $\boldsymbol{x}$.

1.4.2 Necessity of the Conditions in Our Algorithmic Result. There are three main requirements in the conditions of Theorem 1.1 which may give pause to the reader. First, the distribution $\mathcal{D}$ must be an i.i.d. product distribution: the n coordinate marginal distributions are not only independent, they are identically distributed according to some single univariate random variable $\boldsymbol{X}$. Second, certain parameters (various cumulants) of the noise distribution must be provided to the testing algorithm. And finally, the underlying random variable $\boldsymbol{X}$ is not allowed to be a Gaussian distribution.

While these may seem like restrictive requirements, it turns out that each one is in fact *necessary* for constant-sample testability. We give three different lower bounds which show, roughly speaking, that if any of these requirements is relaxed then finite-sample testability with no dependence on n is information-theoretically impossible — in fact, in each case the testing problem becomes essentially as difficult as the sparse recovery problem, requiring $\tilde{\Omega}(\log n)$ samples. In this work, we always use the notation "$\tilde{\Omega}(\cdot)$" to hide factors polylogarithmic in its argument. So $\tilde{\Omega}(\log n)$ means $\Omega(\frac{\log n}{\mathrm{poly}(\log\log n)})$.

Our first lower bound shows that even if $\mathcal{D}$ is allowed to be a product distribution in which half the coordinates are one simple

[2] It will be clear from our proofs that having finite moments of all orders is a stronger condition than our algorithm actually requires; we state this stronger condition here for simplicity of exposition.

[3] It may be helpful to think of C as being a large absolute constant, but we establish our results for general C.

integer-valued distribution (a Poisson distribution) and the other half are a different simple integer-valued distribution (a Poisson distribution with a different parameter), then at least $\Omega(\frac{\log n}{\log\log n})$ samples may be required. This lower bound holds even if no noise is allowed. The proof is given in the full version:

Theorem 1.3 ($\mathcal{D}$ must be i.i.d.). *Let $\mathcal{D}$ be the product distribution $(\mathrm{Poi}(1))^{n/2} \times (\mathrm{Poi}(100))^{n/2}$. Then even if there is no noise (i.e. the noise distribution $\boldsymbol{\eta}$ is identically zero), any algorithm which is an $(\varepsilon = 0.99)$-tester for 1-sparsity under $\mathcal{D}$ must have sample complexity $m = \Omega(\frac{\log n}{\log\log n})$.*

Our second lower bound shows that even if only two "known" possibilities are allowed for the noise distribution, then for $\mathcal{D} = \boldsymbol{X}^n$ where $\boldsymbol{X}$ is a simple "known" integer-valued underlying univariate random variable, at least $\Omega(\frac{\log n}{\log\log n})$ many samples may be required. The proof is given in the full version:

Theorem 1.4 (The noise distribution $\boldsymbol{\eta}$ must be known). *Let $\mathcal{D}$ be the i.i.d. product distribution $\mathcal{D} = (\mathrm{Poi}(1))^n$. Suppose that the noise distribution $\boldsymbol{\eta}$ is unknown to the testing algorithm but is promised to be either $\mathrm{Poi}(1)$ or $\mathrm{Poi}(100)$. Then any $(\varepsilon = 0.99)$-tester for 1-sparsity under $\mathcal{D}$ and the unknown noise distribution $\boldsymbol{\eta} \in \{\mathrm{Poi}(1), \mathrm{Poi}(100)\}$ must have sample complexity $m = \Omega(\frac{\log n}{\log\log n})$.*

Finally, our third (and most technically involved) lower bound says that if the underlying univariate random variable $\boldsymbol{X}$ is allowed to be a Gaussian, then even if the noise is Gaussian at least $\Omega(\log n)$ samples are required. The proof is given in the full version:

Theorem 1.5 ($\mathcal{D}$ cannot be a Gaussian). *Let $\mathcal{D}$ be the standard $N(0,1)^n$ n-dimensional Gaussian distribution and let $\boldsymbol{\eta}$ be distributed as $N(0, c^2)$ where $c > 0$ is any constant. Then the sample complexity of any $(\varepsilon = 0.99)$-tester for 1-sparsity under $\mathcal{D}$ and $\boldsymbol{\eta}$ is $\Omega(\log n)$.*

1.5 Related Work

We view this paper as lying at the confluence of several strands of research in theoretical computer science. As mentioned earlier, a strong motivation for our algorithmic desiderata comes from property testing. In particular, our k-sparsity testing question is in some sense akin to the well-studied problem of *junta testing*, i.e., distinguishing between functions $f : \{\pm1\}^n \to \{\pm1\}$ which depend on at most k coordinates versus those which are ε-far from every such function. There is a very rich line of work on junta testing, see e.g. [7–9, 13, 19, 22] and other works. However, we note that all these papers (and other junta testing papers of which we are aware) assume query access to the unknown function f, whereas in our work we only assume a much weaker form of access, namely noisy labeled random samples. Some other relevant works in the property testing literature are the aforementioned works [18, 27, 31] (see also [2]), which give algorithmc property testing results in the "sample-based" model, and [4], which like our work considers testing with respect to various L_p distances, including the L_2 distance (similar to our work).

A second strand of work is from compressive sensing. Here the results of [17] and related works such as [15, 16] (as well as many other papers) give computationally efficient algorithms to (approximately) recover a sparse vector w given labeled samples of the form $\{(\boldsymbol{x}^{(i)}, w \cdot \boldsymbol{x}^{(i)} + \boldsymbol{\eta})\}_{i=1}^T$ with sample complexity $T = O(k \log n)$. On one hand, such a sample complexity does not meet our core algorithmic desideratum of being independent of n. On the other hand, the algorithmic guarantee in [17] holds as long as the matrix formed by $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}$ satisfies the so-called *restricted isometry property* (see [14] for more details), which is a significantly more general condition than ours. It is natural to wonder if an analogue of Theorem 1.1 can be obtained if $\mathcal{D}$ satisfies the weaker condition of being such that randomly drawn samples from $\mathcal{D}$ satisfy the restricted isometry property with high probability. The answer to this question is negative; in particular, Theorem 1.3 gives an example of a distribution $\mathcal{D}$ for which $\tilde{\Omega}(\log n)$ samples are necessary for testing k-sparsity, but it is easy to show that randomly drawn samples from this distribution satisfy the restricted isometry property with high probability.

Finally, another related line of work is given by Kong and Valiant [32], who considered a setting in which an algorithm gets labeled samples of the form $(\boldsymbol{x}, \boldsymbol{y} = w \cdot \boldsymbol{x} + \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is an *unknown* distribution independent of $\boldsymbol{x}$ and w is a general (non-sparse) n-dimensional vector. The task of the algorithm is to estimate the variance of $\boldsymbol{\eta}$ or equivalently, $\|w\|_2$; they view such a result as *estimating how much of the data, i.e., $\boldsymbol{y}$, is explained by the linear part $w \cdot \boldsymbol{x}$*. While learning w itself requires $\Theta(n)$ samples (essentially the same as linear regression), their main result is that $\|w\|_2$ can be estimated with a sublinear number of samples. In particular, if the distribution of $\boldsymbol{x}$ is isotropic, then the sample complexity required for this is only $O(\sqrt{n})$. In light of Theorem 1.1 and the results of [32], it is natural to ask whether there is a non-trivial estimator for noise in our setting when the target vector w is assumed to be k-sparse. However, Theorem 1.4 essentially answers this in the negative, showing that if the magnitude of the noise is unknown, then any estimator must require $\tilde{\Omega}(\log n)$ samples even for 1-linearity testing. On the other hand, $O(\log n)$ samples suffice for recovering the target w (and hence the magnitude of the noise) when k is a constant.

2 OUR TECHNIQUES AND A DETAILED OVERVIEW OF OUR RESULTS

2.1 Our Algorithmic Techniques: Analysis Based on Cumulants

Both of our algorithms for testing sparsity make essential use of the *cumulants* of the one-dimensional coordinate marginal random variable $\boldsymbol{X}$. For any integer $\ell \geq 0$ and any real random variable $\boldsymbol{X}$, the ℓ-th cumulant of $\boldsymbol{X}$, denoted $\kappa_\ell(\boldsymbol{X})$, is defined in terms of the first ℓ moments of $\boldsymbol{X}$, and, like the moments of $\boldsymbol{X}$, it can be estimated using independent draws from $\boldsymbol{X}$ (see Definition 3.1 for a formal definition of cumulants.) However, cumulants enjoy a number of attractive properties which are not shared by moments and which are crucial for our analysis.

There are two key properties, both very simple. First, cumulants are *additive* for independent random variables:

$$\text{If } \boldsymbol{X}, \boldsymbol{Y} \text{ are independent, then } \kappa_\ell(\boldsymbol{X} + \boldsymbol{Y}) = \kappa_\ell(\boldsymbol{X}) + \kappa_\ell(\boldsymbol{Y}).$$

Second, cumulants are *homogeneous*:

$$\text{For all } c \in \mathbb{R}, \text{ it holds that } \kappa_\ell(c\boldsymbol{X}) = c^\ell \cdot \kappa_\ell(\boldsymbol{X}).$$

We now explain the key idea of why additivity and homogeneity of cumulants are useful for the algorithmic problem we consider. These properties directly imply that if a distribution $\mathcal{D}$ over $\mathbb{R}^n$ has coordinate marginals that are i.i.d. according to a random variable X, then for $\boldsymbol{x} \sim \mathcal{D}$ and $\boldsymbol{y} = w \cdot \boldsymbol{x} + \boldsymbol{\eta}$, we have that

$$\kappa_\ell(\boldsymbol{y}) = \kappa_\ell(\boldsymbol{\eta}) + \kappa_\ell(X) \cdot \sum_{i=1}^{n} w_i^\ell.$$

It follows that if the ℓ-th cumulants of $\boldsymbol{\eta}$ and of X are known and the ℓ-th cumulant $\kappa_\ell(X)$ of X is not too small, then from an estimate of $\kappa_\ell(\boldsymbol{y})$ (which can be obtained from samples) it is possible to obtain an estimate of the power sum $\sum_{i=1}^n w_i^\ell$. By doing this for k suitable different (even) values of ℓ, provided that the cumulants $\kappa_\ell(X)$ are not too small, it is possible to estimate the magnitudes of the k largest-magnitude coordinates of w. These estimates can be shown to yield the desired information about whether or not w is (close to) k-sparse.

The argument sketched in the previous paragraph explains, at least at an intuitive level, why it is possible to test for k-sparsity *if the random variable X has k nonzero cumulants*. But why will every non-Gaussian random variable X (as described in Theorem 1.1) satisfy this property, and why does Theorem 1.1 exclude Gaussian distributions? The second of these questions has a very simple answer so we address it first: it is well known that for any normal distribution $X \sim N(\mu, \sigma^2)$, the first two cumulants are $\kappa_1(X) = \mu$, $\kappa_2(X) = \sigma^2$, and all other cumulants are zero. It follows that indeed our algorithmic approach cannot be carried out for normal distributions.[4] The answer to the first question comes from a deep result in probability theory due to J. Marcinkiewicz:

Theorem 2.1 (Marcinkiewicz's theorem [11, 33, 34]). *If X is a random variable that has a finite number of nonzero cumulants, then X must be a normal random variable (and X has at most two nonzero cumulants).*

It follows that if X is not a normal distribution, then it must have infinitely many nonzero cumulants, and hence the algorithmic approach sketched above can be made to work for testing k-sparsity under X^n. Details of the estimation procedure and of the analysis of the overall general algorithm are provided in Section 4 and Section 5 respectively.

2.2 A Structural Result on Cumulants: Nonzero Cumulants Cannot Be "Spaced Far Apart"

As described above, our main positive result on testing for sparsity under a product distribution X^n uses a sequence of orders $i_1, i_2, \ldots, i_k$ such that the corresponding cumulants $\kappa_{i_j}(X)$ are all nonzero. Since the running time of our algorithm depends directly on i_k, it is natural to ask how large is this value. Recall that Marcinkiewicz's theorem ensures that for any non-Gaussian distribution there indeed must exist nonzero cumulants of infinitely many orders $i_1, i_2, \ldots$, but it gives no information about how far apart these orders may need to be. Thus we are motivated to investigate the following question: given a real random variable X, how large can the gap in orders be between consecutive nonzero cumulants? This is a natural question which, prior to our work, seems to have been completely unexplored.

In Section 6 we give the first result along these lines, by giving an explicit upper bound on the gap between nonzero cumulants for random variables with *bounded support* (see Theorem 6.1). This theorem establishes that for any real random variable X with unit variance and support bounded in $[-\mathrm{B}, \mathrm{B}]$, given any positive integer ℓ there must be a value $j \in [\ell+1, (4\mathrm{B})^{O(\ell)}]$ such that the j-th cumulant $\kappa_j(X)$ has magnitude at least $\left|\kappa_j(X)\right| \geq 2^{-(4\mathrm{B})^{O(\ell)}}$. Like the proof of Marcinkiewicz's theorem, the proof of our Theorem 6.1 uses complex analytic arguments, specifically results about the distribution of zeros of entire functions, the Hadamard factorization of entire functions and the Hadamard Three-Circle Theorem.

2.3 A More Efficient Algorithm for "Nice" Distributions

In addition to the general positive result described above, we also give a refined result, showing that a significantly better sample complexity can be achieved for distributions which are "nice" in the sense that they have $k+1$ consecutive even cumulants $\kappa_2, \kappa_4, \ldots, \kappa_{2k+2}$ that are all (noticeably) nonzero. This is achieved via a different algorithm; like the previously described general algorithm, it uses (estimates of) the power sums $\sum_{i=1}^n w_i^\ell$, but it uses these power sums in a different way, by exploiting some basic properties of symmetric polynomials. The first $k+1$ power sums $\sum_{i=1}^n w_i^2$, $\sum_{i=1}^n w_i^4, \ldots$ are used to estimate the $(k+1)$-st elementary symmetric polynomial $\sum_{1 \leq i_1 < i_2 < \cdots < i_{k+1} \leq n} w_{i_1} w_{i_2} \cdots w_{i_{k+1}}$. The value of this polynomial will clearly be zero if w is k-sparse, and it can be shown that it will be "noticeably far from nonzero" if w is far from k-sparse. These ideas can be converted into a testing algorithm; see the full version for details.

2.4 Our Lower Bounds and Lower Bound Techniques

The lower bounds of Theorem 1.3 and Theorem 1.4 both crucially exploit the well known additivity property of the Poisson distribution: for $a, b > 0$, we have that $\mathrm{Poi}(a) + \mathrm{Poi}(b) = \mathrm{Poi}(a+b)$. To see why this is useful for lower bounds, let us explain the high-level idea that underlies Theorem 1.3. For intuition, first imagine that rather than receiving pairs $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}$, instead the testing algorithm is only given the output value $\boldsymbol{y}$ from each pair. Then by the additivity of the Poisson distribution, it would be information-theoretically impossible to distinguish between (i) the case in which $\boldsymbol{y}$ is a sum of 100 coordinates each of which is distributed as $\mathrm{Poi}(1)$ (and hence the target vector w is 0.99-far from being 1-sparse), versus (ii) the case in which $\boldsymbol{y}$ is a single coordinate distributed as $\mathrm{Poi}(100)$ (and hence the target vector w is 1-sparse). Of course, in our actual testing scenario things are not so simple because the testing algorithm does receive the coordinates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of each example $(\boldsymbol{x}, \boldsymbol{y})$ along with the value of $\boldsymbol{y}$, and this provides additional useful information. Our proof establishes that this additional information is essentially useless unless $\tilde{\Omega}(\log n)$ samples are provided. Roughly speaking, this is because with $n/2$ coordinates distributed as $\mathrm{Poi}(1)$ and $n/2$ coordinates distributed as $\mathrm{Poi}(100)$, there are "too many possibilities" of each sort ((i) and (ii) above) for the $\boldsymbol{x}$'s to

[4] Recall that by our lower bound Theorem 1.5, this is not a failing of our particular algorithm sketched above but an inherent difficulty in the testing problem. Theorem 1.5 shows that no algorithm can test k-sparsity with a sample complexity that is $o(\log n)$ when the underlying distribution is normal.

provide useful distinguishing information until this many samples have been received.

The lower bound of Theorem 1.4 is based on similar ideas. Now all n coordinates are identically distributed as Poi(1), but the noise may be distributed either as Poi(1) or as Poi(100). As above, if only the output values $\boldsymbol{y} = w \cdot \boldsymbol{x} + \boldsymbol{\eta}$ were available to the tester, it would be impossible to distinguish between (i′) the target vector w is 1-sparse and the noise is Poi(100), versus (ii′) the target vector is 100-sparse and the noise is Poi(1), since in both cases the distribution of $\boldsymbol{y}$ is Poi(101). The formal proof is by a reduction to Theorem 1.3.

Finally, we turn to the lower bound of Theorem 1.5, which states that $\Omega(\log n)$ samples are required for the testing problem if the distribution $\mathcal{D}$ is $N(0,1)^n$ and the noise distribution $\boldsymbol{\eta}$ is normally distributed as $N(0, c^2)$. The high level idea is that it is difficult to distinguish between the following two distributions over pairs $(\boldsymbol{x}, \boldsymbol{y})$:

- First distribution (no-distribution): in each draw of $(\boldsymbol{x}, \boldsymbol{y})$ from the no-distribution, each $\boldsymbol{x}_j$ is an independent $N(0,1)$ random variable, and $\boldsymbol{y}$ is an $N(0, 1+c^2)$ normal random variable which is completely independent of all of the $\boldsymbol{x}_j$'s;
- Second distribution (yes-distribution): there is a fixed but unknown uniform random coordinate $\boldsymbol{i} \in [n]$, and in each draw of $(\boldsymbol{x}, \boldsymbol{y})$ from the yes-distribution, each $\boldsymbol{x}_j$ is an independent $N(0,1)$ random variable and $\boldsymbol{y} = \boldsymbol{x}_{\boldsymbol{i}} + N(0, c^2)$.

Similar to the first paragraph of this subsection, since the sum of a draw from $N(0,1)$ plus an independent draw from $N(0, c^2)$ is a draw from $N(0, 1+c^2)$, if only the output value $\boldsymbol{y}$ from each pair were given to a tester then it would be information-theoretically impossible to distinguish between the two distributions described above. And similar to the discussion in that paragraph, the idea that animates our lower bound proof here is that the additional information (the $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$-coordinates of each sample) available to the testing algorithm is essentially useless unless $\Omega(\log n)$ samples are provided. As before, roughly speaking, this is because there are "too many possibilities" (for which coordinate might be the unknown hidden $\boldsymbol{i} \sim [n]$ in the second distribution) for the $\boldsymbol{x}$-components of the samples to provide useful distinguishing information until $\Omega(\log n)$ many samples have been received. The formal argument uses Bayes' rule to analyze the optimal distinguishing algorithm (corresponding to a maximum likelihood approach) and employs the Berry-Esseen theorem to make these intuitions precise.

Remark 2.2. We note here that we also give a quantitative refinement of Theorem 1.5. Since for a Gaussian random variable $\boldsymbol{X}$ all cumulants $\kappa_\ell(\boldsymbol{X})$, $\ell > 2$, are zero, we may informally view Theorem 1.5 as saying that if the cumulants of $\boldsymbol{X}$ are zero then the number of samples required to test for sparsity under $\boldsymbol{X}^n$ may be arbitrarily large (going to infinity as n does). This intuitively suggests that if the cumulants of $\boldsymbol{X}$ are "small" then "many" samples should be required to test for sparsity under $\boldsymbol{X}^n$. In the full version we make this intuition precise: building on Theorem 1.5, we show (roughly speaking) that if the cumulants of a random variable $\boldsymbol{X}$ are at most γ, then at least $1/\gamma$ samples are required for testing sparsity under $\boldsymbol{X}^n$ and Gaussian noise. See the full version for a precise statement and proof.

2.5 Directions for Future Work

Our results suggest a number of directions for future work; we touch on a few of these below.

Within the sparsity testing framework that this paper considers, it would be interesting to gain a more quantitatively precise understanding of the sample complexity required to test sparsity. A natural specific question here is the following: let $\boldsymbol{X}$ be a simple random variable such as $\boldsymbol{X}$ = uniform on $\{-1, 1\}$ or $\boldsymbol{X}$ = uniform on $[0, 1]$. For these specific distributions, what is the optimal dependence on k for the k-sparsity testing question that we have considered? It would be interesting to determine whether or not an exponential dependence on k is required.

Another natural quantitative question arises from our results in Section 6. Theorem 6.1 implies an explicit "tower-type" upper bound on the minimum value i_k such that a random variable $\boldsymbol{X}$ as above must have at least k nonzero cumulants in $\{1, \ldots, i_k\}$. It would be interesting to obtain sharper quantitative bounds or bounds that hold under relaxed conditions on the random variable $\boldsymbol{X}$.

Finally, another intriguing potential direction is to look beyond sparsity and attempt to identify other contexts in which sparsity is testable with a constant sample complexity independent of n. A concrete first goal along these lines is to investigate the sparsity testing question when $(\boldsymbol{x}, \boldsymbol{y})$ is distributed as $\boldsymbol{y} = \phi(w \cdot \boldsymbol{x}) + \text{noise}$ for various natural transfer functions ϕ such as the probit function or the logistic function.

2.6 Notational Conventions

Given a vector $w \in \mathbb{R}^n$ we write $\|w\|_\ell$ to denote the ℓ-norm of w, i.e. $\|w\|_\ell = \left(\sum_{i=1}^n w_i^\ell\right)^{1/\ell}$. For a nonzero vector $w \in \mathbb{R}^n$ where $n > k$, the vector w's *distance from being k-sparse* is

$$\operatorname{dist}(w, k\text{-sparse}) := \min_{w' \in \mathbb{R}^n:\ w' \text{ is } k\text{-sparse}} \frac{\|w - w'\|_2}{\|w\|_2}.$$

Equivalently, if the entries of w are sorted by magnitude so that $|w_{i_1}| \geq \cdots \geq |w_{i_n}|$, the distance of w from being k-sparse is

$$\frac{\sqrt{w_{i_{k+1}}^2 + \cdots + w_{i_n}^2}}{\|w\|_2}.$$

For a random variable Z, we write $m_\ell(Z)$ to denote its ℓth raw moment, i.e., $\mathbf{E}[Z^\ell]$.

3 PRELIMINARIES: FACTS ABOUT CUMULANTS

In this section we recall some basic facts about cumulants which we will use extensively.

Definition 3.1. The *cumulants* of $\boldsymbol{X}$ are defined by the cumulant generating function $K(t)$, which is the natural logarithm of the moment generating function $M(t) = \mathbf{E}[e^{t\boldsymbol{X}}]$:

$$K(t) = \ln \mathbf{E}[e^{t\boldsymbol{X}}].$$

Equivalently, $e^{K(t)} = \mathbf{E}[e^{t\boldsymbol{X}}]$. For $\ell > 0$ the cumulants of $\boldsymbol{X}$, which are denoted $\kappa_\ell(\boldsymbol{X})$, are the coefficients in the Taylor expansion of

the cumulant generating function about the origin:

$$K(t) = \sum_{\ell=1}^{\infty} \kappa_\ell(\boldsymbol{X}) \frac{t^\ell}{\ell!}.$$

Equivalently, $\kappa_\ell(\boldsymbol{X}) = K^{(\ell)}(0)$.

One useful property of cumulants is additivity for independent random variables, which follows as an easy consequence of the definition:

Fact 3.2. *If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent random variables then $\kappa_\ell(\boldsymbol{X}+\boldsymbol{Y}) = \kappa_\ell(\boldsymbol{X}) + \kappa_\ell(\boldsymbol{Y})$.*

Corollary 3.3. *For any random variable $\boldsymbol{X}$, the value of $\kappa_\ell(\boldsymbol{X} - \boldsymbol{X})$ is zero when ℓ is odd and is $2 \cdot \kappa_\ell(\boldsymbol{X})$ when ℓ is even.*

Another useful property is ℓ-th order homogeneity of the ℓ-th cumulant:

Fact 3.4. *For any $c \in \mathbb{R}$ and any $\ell \in \mathbb{N}$, we have $\kappa_\ell(c\boldsymbol{X}) = c^\ell \kappa_\ell(\boldsymbol{X})$.*

Looking ahead, all of our algorithms will work by estimating cumulants of the real random variable $\boldsymbol{y}$ which is distributed as $\boldsymbol{y} = w \cdot \boldsymbol{x} + \boldsymbol{\eta}$ where $\boldsymbol{x} \sim \boldsymbol{X}^n$ and $\boldsymbol{\eta}$ is independently drawn from a noise distribution. By Fact 3.2 and Corollary 3.3, we can (and do) assume throughout the analysis of our algorithms that $\boldsymbol{X}$ and $\boldsymbol{\eta}$ are both symmetric distributions. This is because we can combine two independent draws $(\boldsymbol{x}_1, \boldsymbol{y}_1)$ and $(\boldsymbol{x}_2, \boldsymbol{y}_2)$ with $\boldsymbol{y}_i = w \cdot \boldsymbol{x}_i + \boldsymbol{\eta}$ into one draw $((\boldsymbol{x}_1 - \boldsymbol{x}_2)/\sqrt{2}, (\boldsymbol{y}_1 - \boldsymbol{y}_2)/\sqrt{2})$, such that the new marginal distribution $\frac{\boldsymbol{X}-\boldsymbol{X}}{\sqrt{2}}$ and noise distribution $\frac{\boldsymbol{\eta}-\boldsymbol{\eta}}{\sqrt{2}}$ are both symmetric and have the same variances as before combining.

Let $m_\ell(\boldsymbol{X})$ denote the ℓth moment $\mathbf{E}[\boldsymbol{X}^\ell]$ of a random variable $\boldsymbol{X}$. There is a one-to-one mapping between the first n moments and the first n cumulants which can be derived by relating coefficients in the Taylor series expansions of the cumulant and moment generating functions [3]:

Fact 3.5. *Let $\boldsymbol{X}$ be a random variable with mean zero. Then*

$$\kappa_\ell(\boldsymbol{X}) = m_\ell(\boldsymbol{X}) - \sum_{j=1}^{\ell-1} \binom{\ell-1}{j-1} \kappa_j(\boldsymbol{X}) \cdot m_{\ell-j}(\boldsymbol{X}). \tag{2}$$

Cumulants can be expressed in terms of moments and vice-versa:

$$m_\ell(\boldsymbol{X}) = \sum_{k=1}^{\ell} B_{\ell,k}\Big(\kappa_1(\boldsymbol{X}), \ldots, \kappa_{\ell-k+1}(\boldsymbol{X})\Big) \tag{3}$$

and

$$\kappa_\ell(\boldsymbol{X}) = \sum_{k=1}^{\ell} (-1)^{k-1}(k-1)! B_{\ell,k}\Big(m_1(\boldsymbol{X}), \ldots, m_{\ell-k+1}(\boldsymbol{X})\Big), \tag{4}$$

where $B_{\ell,k}$ are incomplete Bell polynomials,

$$B_{\ell,k}(x_1, \ldots, x_{\ell-k+1}) = \sum \frac{\ell!}{j_1! \cdots j_{\ell-k+1}!} \left(\frac{x_1}{1}\right)^{j_1} \cdots \left(\frac{x_{\ell-k+1}}{(\ell-k+1)!}\right)^{j_{\ell-k+1}},$$

whose summation is over all non-negative sequences $(j_1, \ldots, j_{\ell-k+1})$ that satisfy

$$j_1 + \cdots + j_{\ell-k+1} = k \text{ and } j_1 + 2j_2 + \cdots + (\ell-k+1)j_{\ell-k+1} = \ell.$$

Equation (2) can be used to give a upper bound on $\kappa_\ell(\boldsymbol{X})$ in terms of the moments of $\boldsymbol{X}$:

Claim 3.6. *For any random variable $\boldsymbol{X}$ with mean zero and any even ℓ, we have $|\kappa_\ell(\boldsymbol{X})| \le m_\ell(\boldsymbol{X}) \cdot e^\ell \cdot \ell!$.*

Remark 3.7. When $\boldsymbol{X}$ is the random variable that is uniform over $\{0, 1, \ldots, C\}$, the ℓ-th cumulant is $\kappa_\ell(\boldsymbol{X}) = \frac{\mathrm{Bern}(\ell)}{\ell} \cdot (C^\ell - 1)$ where $\mathrm{Bern}(\ell)$ is the Bernoulli number of order ℓ which has an asymptotic growth as $(\frac{\ell/2}{\pi e})^\ell$ [41]. This simple example shows that the dominant $\ell!$ term in Claim 3.6 is essentially best possible.

We defer the proof of Claim 3.6 to the full version.

4 ESTIMATING MOMENTS OF THE WEIGHT VECTOR w USING MOMENTS AND CUMULANTS

Throughout this section $\boldsymbol{X}$ will denote a real random variable with mean zero, unit variance, and finite moments of all orders, and $w \in \mathbb{R}^n$ will be a vector that is promised to have $\|w\|_2 \in [1/C, C]$. The main result of this section is the following theorem, which shows that it is possible to estimate norms of the vector w given access to noisy samples of the form $(\boldsymbol{x}, \boldsymbol{y} = w \cdot \boldsymbol{x} + \boldsymbol{\eta})$ where $\boldsymbol{x} \sim \boldsymbol{X}^n$:

Theorem 4.1. *Let $\boldsymbol{X}$ be a symmetric real-valued random variable with variance 1 and finite moments of all orders, and let $\boldsymbol{\eta}$ be a symmetric real-valued random variable with finite moments of all orders. There is an algorithm (depending on $\boldsymbol{X}$ and $\boldsymbol{\eta}$)[5] with the following property: Let $w \in \mathbb{R}^n$ be any (unknown) vector with $\|w\|_2 \in [1/C, C]$. Given any $\varepsilon, \delta > 0$ and any even integer ℓ such that $|\kappa_\ell(\boldsymbol{X})| \ge \tau$, the algorithm takes as input $m = \mathrm{poly}(\ell!, m_{2\ell}(\boldsymbol{X}) + m_{2\ell}(\boldsymbol{\eta}), 1/(\delta\varepsilon), 1/\tau, C^\ell)$ many independent random samples where each $\boldsymbol{x}^{(i)} \sim \boldsymbol{X}^n$ and each $\boldsymbol{z}^{(i)} = w \cdot \boldsymbol{x}^{(i)} + \boldsymbol{\eta}$. It outputs an estimate M_ℓ of $\sum_{i=1}^n w_i^\ell$, the ℓ-th power of the ℓ-norm of the vector w, which with probability at least $1 - \delta$ satisfies*

$$\left|M_\ell - \|w\|_\ell^\ell\right| \le \varepsilon.$$

We will also use the following result on estimating the moments of $w \cdot \boldsymbol{X} + \boldsymbol{\eta}$:

Lemma 4.2. *Let $\boldsymbol{X}$ be a symmetric real-valued random variable with mean zero, variance 1, and finite moments of all orders, and let $\boldsymbol{\eta}$ be a symmetric real-valued random variable with finite moments of all orders.*

There is an algorithm (depending on $\boldsymbol{X}$ and $\boldsymbol{\eta}$) with the following property: Let $w \in \mathbb{R}^n$ be any (unknown) vector with $\|w\|_2 \in [1/C, C]$. Given any ε and δ and any even integer ℓ, the algorithm takes as input $m = \mathrm{poly}(\ell!, m_{2\ell}(\boldsymbol{X}) + m_{2\ell}(\boldsymbol{\eta}), 1/(\delta\varepsilon), C^\ell)$ many independent random samples where each $\boldsymbol{x}^{(i)} \sim \boldsymbol{X}^n$ and each $\boldsymbol{z}^{(i)} = w \cdot \boldsymbol{x}^{(i)} + \boldsymbol{\eta}$. It outputs an estimate $\widetilde{m}_\ell(\boldsymbol{Z})$, which with probability at least $1 - \delta$ satisfies

$$\left|\widetilde{m}_\ell(\boldsymbol{Z}) - \mathbf{E}[|w \cdot \boldsymbol{x}^{(i)} + \boldsymbol{\eta}|^\ell]\right| \le \varepsilon.$$

We defer the proof of Theorem 4.1 and Lemma 4.2 to the full version.

[5] As will be clear from the proof, the algorithm only needs to "know" $\boldsymbol{X}$ and $\boldsymbol{\eta}$ in the sense of having sufficiently accurate estimates of certain cumulants.

5 GENERAL TESTING ALGORITHM: PROOF OF THEOREM 1.1

The main result of this section is Theorem 5.1, which is a more precise version of Theorem 1.1. Roughly speaking, it says that there is a constant-sample tolerant tester for k-sparsity for any non-Gaussian distribution.

Theorem 5.1 (Detailed statement of main result: tolerant tester for non-Gaussian distributions). *Fix any real random variable $\boldsymbol{X}$ which has variance one and finite moments of every order, and is not a Gaussian distribution (i.e. its total variation distance from every Gaussian distribution is nonzero). Let $\boldsymbol{\eta}$ be any real random variable with finite moments of every order.*

There is a tolerant testing algorithm with the following properties: Let $0 \le c < s \le 1$ be any given completeness and soundness parameters, and let w be any vector (unknown to the algorithm) with $1/C \le \|w\|_2 \le C$. The algorithm is given c, s, ε, k, C and access to independent samples $(\boldsymbol{x}, \boldsymbol{y} = w \cdot \boldsymbol{x} + \boldsymbol{\eta})$ where each $\boldsymbol{x} \sim \boldsymbol{X}^n$. Its sample complexity is

$$m = \mathrm{poly}\Big(\ell_1!, m_{2\ell_1}(\boldsymbol{X}) + m_{2\ell_1}(\boldsymbol{\eta}), 1/\delta_1^{\ell_1}, 1/\tau, C^{\ell_1}\Big),$$

where $\tau = \min_{i\in[k]}\{|\kappa_{\ell_i}(\boldsymbol{X})|\}$ and $\{\ell_i\}_{i\in[k]}, \{\delta_i\}_{i\in[k]}$ are as defined below. The algorithm satisfies the following:

- *if $\mathrm{dist}(w, k\text{-sparse}) \le c$ then with probability at least $9/10$ the algorithm outputs "yes;" and*
- *if $\mathrm{dist}(w, k\text{-sparse}) \ge s$ then with probability at least $9/10$ the algorithm outputs "no."*

Furthermore, if the random variable $\boldsymbol{X}$ is supported in $[-\mathrm{B}, \mathrm{B}]$ for some constant B, then the sample complexity of the tolerant tester (as a function of k) is bounded by a tower function of height $O(k)$.

Remark 5.2. The running time of our algorithm is $\mathrm{poly}(m)$ (independent of n) if our algorithm is allowed to obtain $\boldsymbol{y}$ directly and skip the reading of $\boldsymbol{x}$. The same remark holds for the more efficient tester for "nice" distributions given in the full version.

We begin by stating the algorithm:

(1) First, recall that as stated earlier, we may assume that $\boldsymbol{X}$ and $\boldsymbol{\eta}$ are both symmetric. We rescale all samples by a factor of C so that $\|w\|_2 \in [1/C^2, 1]$ in our subsequent analysis.
We fix $\varepsilon = \frac{s^2-c^2}{2C^4}$ and apply Lemma 4.2 with $\ell = 2$ to obtain an estimate s_2 of $\sum_{i=1}^n w_i^2 = \|w\|_2^2 = \mathrm{E}[|w \cdot \boldsymbol{X}^n + \boldsymbol{\eta}|^2] - \mathrm{E}[|\boldsymbol{\eta}|^2]$ that is accurate to within additive error $\varepsilon/4$ (with probability 0.99).

(2) Set a sequence of error parameters $\delta_1 < \delta_2 < \ldots < \delta_k$ and natural numbers (orders of cumulants) $\ell_1 > \ell_2 > \cdots > \ell_k$ with the following properties:
   (a) $\delta_k = \varepsilon/(12k)$ and $\ell_k \ge 100/\delta_k^3$ is even ;
   (b) For $i = k-1, k-2, \ldots, 1$, $\delta_i = (\delta_{i+1}/5\ell_{i+1})^{\ell_{i+1}}/(2k)$ and $\ell_i \ge 100/\delta_i^3$ is even;
   (c) For each $i \in [1, k]$ the ℓ_i-th cumulant $\kappa_{\ell_i}(\boldsymbol{X})$ of $\boldsymbol{X}$ is nonzero.

(3) For $j = 1, \ldots, k$: run the algorithm of Theorem 4.1 to obtain an estimate M_{ℓ_j} which satisfies $\left|M_{\ell_j} - \|w\|_{\ell_j}^{\ell_j}\right| \le (\delta_j/5\ell_j)^{\ell_j}/(2k)$ with failure probability at most $1/(20k)$. Set

$$\widetilde{w}_j = \min\left\{1, \left|M_{\ell_j} - \sum_{i=1}^{j-1} \widetilde{w}_i^{\ell_j}\right|^{1/\ell_j}\right\}.$$

   (The intuition is that at the j-th iteration of this step, the algorithm computes an estimate $\widetilde{w}_j$ of the magnitude of the j-th largest magnitude coordinate in the weight vector w.)

(4) If $\sum_{i=1}^k \widetilde{w}_i^2 < (1 - \frac{s^2-c^2}{2}) \cdot s_2$, output "No," and otherwise output "Yes."

A remark is in order regarding condition 2(c) above. Recall that by Marcinkiewicz's theorem [11, 34], since $\boldsymbol{X}$ is not a Gaussian distribution it must have infinitely many nonzero cumulants. (This is where we use the assumption that $\boldsymbol{X}$ is not Gaussian; indeed if $\boldsymbol{X}$ were Gaussian then τ as defined in the theorem statement would be zero.) Hence a sequence of orders $\ell_1 > \cdots > \ell_k$ satisfying conditions 2(a), 2(b) and 2(c) must indeed always exist.

To analyze the algorithm we will use the following lemma, which shows that a good estimate of $\|w\|_\ell^\ell$ yields a good estimate of $\|w\|_\infty$:

Lemma 5.3. *Given any vector w with $\|w\|_2^2 \le 1$ and $\delta > 0$, let $\ell \ge 100/\delta^3$ be even and let M_ℓ satisfy $\left|M_\ell - \|w\|_\ell^\ell\right| \le (\frac{\delta}{5})^\ell/2$. Then $\left|M_\ell^{1/\ell} - \|w\|_\infty\right| \le \delta$.*

We defer the proof of Lemma 5.3 to Section 5.1 and use it to prove Theorem 5.1.

Proof of Theorem 5.1. Without loss of generality we assume that the coordinates of w satisfy $w_1 \ge w_2 \ge \cdots \ge w_n \ge 0$. We use induction to prove that $|\widetilde{w}_j - w_j| \le \delta_j/\ell_j$ for all $j = 1, \ldots, k$.

For the base case $j = 1$, we have that the difference between M_{ℓ_1} and $\|w\|_{\ell_1}^{\ell_1}$ has magnitude at most $(\delta_1/5\ell_1)^{\ell_1}/2k$, so we can apply Lemma 5.3. The value of $\tilde{w}_1$ as defined in Step 3 is $M_{\ell_1}^{1/\ell_1}$, so Lemma 5.3 gives that $|\widetilde{w}_1 - w_1| \le \delta_1/\ell_1$.

For the inductive step, we assume that the claimed bound holds for all $\widetilde{w}_1, \ldots, \widetilde{w}_{j-1}$, and we will apply Lemma 5.3 to bound the distance between w_j and $\widetilde{w}_j$. We bound the error between $M_{\ell_j} - \sum_{i=1}^{j-1} \widetilde{w}_i^{\ell_j}$ and $\|w\|_{\ell_j}^{\ell_j} - \sum_{i=1}^{j-1} w_i^{\ell_j}$ by

$$\begin{aligned}
&|M_{\ell_j} - \|w\|_{\ell_j}^{\ell_j}| + \sum_{i=1}^{j-1} |\widetilde{w}_i^{\ell_j} - w_i^{\ell_j}| \\
\le &(\delta_j/5\ell_j)^{\ell_j}/2k + \sum_{i=1}^{j-1} |\widetilde{w}_i - w_i| \cdot \ell_j \qquad \text{(using } 0 \le \widetilde{w}_i, w_i \le 1\text{)} \\
\le &(\delta_j/5\ell_j)^{\ell_j}/2k + \sum_{i=1}^{j-1} (\delta_i/\ell_i) \cdot \ell_j \\
\le &(\delta_j/5\ell_j)^{\ell_j}/2k + \sum_{i=1}^{j-1} \delta_i \le (\delta_j/5\ell_j)^{\ell_j}/2.
\end{aligned}$$

In the third step of our algorithm, we use $M_{\ell_j} - \sum_{i=1}^{j-1} \widetilde{w}_i^{\ell_j}$ as an estimation of $\|(w_j, w_{j+1}, \ldots, w_n)\|_{\ell_j}^{\ell_j}$. The above calculation shows the error of this estimation is $(\delta_j/5\ell_j)^{\ell_j}/2$. Thus applying Lemma 5.3 to $\left|M_{\ell_j} - \sum_{i=1}^{j-1} \widetilde{w}_i^{\ell_j}\right|^{1/\ell_j}$ and the vector $(w_j, w_{j+1}, \ldots, w_n)$ with its

"δ" parameter being δ_j/ℓ_j, we get that $|\widetilde{w}_j - w_j| \le \delta_j/\ell_j$. This concludes the inductive proof.

With this upper bound on each $|\widetilde{w}_j - w_j|$ in hand, we can infer that

$$\sum_{i=1}^{k} |\widetilde{w}_i^2 - w_i^2| = \sum_{i=1}^{k} |\widetilde{w}_i - w_i| \cdot |\widetilde{w}_i + w_i| \le \sum_{i=1}^{k} (\delta_i/\ell_i) \cdot 3 \le \varepsilon/4, \quad (5)$$

where the first inequality uses $\|w\|_2 \le 1$ and the closeness of each $\widetilde{w}_i$ to w_i to upper bound $|\widetilde{w}_i + w_i| \le 3$.

We now use Equation (5) to establish correctness of our algorithm. To do this, we first consider the "yes" case in which $\mathrm{dist}(w, k\text{-sparse}) \le c$. In this case, we have that $\sum_{i=1}^{k} \widetilde{w}_i^2 \ge \sum_{i=1}^{k} w_i^2 - \varepsilon/4 \ge (1-c^2) \cdot \|w\|_2^2 - \varepsilon/4$. Since $s_2 = \|w\|_2^2 \pm \varepsilon/4$, we have

$$\sum_{i=1}^{k} |\widetilde{w}_i|^2 \ge (1-c^2)(s_2 - \varepsilon/4) - \varepsilon/4 \ge (1-c^2) \cdot s_2 - \varepsilon/2$$

Furthermore, since $\|w\|_2 \in [1/C^2, 1]$ shows $\|w\|_2^2 \in [1/C^4, 1]$, given $\varepsilon = \frac{s^2-c^2}{2C^4}$, we have

$$s_2 \ge \|w\|_2^2 - \varepsilon/4 \ge 1/C^4 - 1/(8C^4) \ge 2/(3C^4) > \varepsilon/(s^2-c^2)$$

and we can simplify our lower bound on $\sum_{i=1}^{k} \widetilde{w}_i^2$ to

$$(1-c^2)s_2 - \varepsilon/2 > (1-c^2)s_2 - \frac{s^2-c^2}{2} \cdot s_2 = \left(1 - \frac{s^2-c^2}{2}\right) \cdot s_2,$$

from which we see that the algorithm is correct in the "yes"-case.

Similarly, in the "NO" case, we have $\sum_{i=1}^{k} \widetilde{w}_i^2 < \left(1 - \frac{s^2-c^2}{2}\right) \cdot s_2$. This proves the assertions made in the two bulleted statements of the theorem.

Finally, when $\boldsymbol{X}$ is supported in $[-\mathrm{B}, \mathrm{B}]$, we apply Theorem 6.1 to upper bound ℓ_i: given any ℓ_{i+1} and δ_{i+1}, for $t = 100k^3/(\delta_{i+1}/5\ell_{i+1})^{3\ell_{i+1}}$, there always exists $\ell_i \in [t, (4\mathrm{B})^{O(t)}]$ with $\kappa_{\ell_i}(X) \ge 2^{-(4\mathrm{B})^{O(t)}}$. Thus τ is also lower bounded by $2^{-(4\mathrm{B})^{O(\ell_1)}}$.

□

5.1 Proof of Lemma 5.3

For convenience we assume throughout this subsection that $w_1 \ge w_2 \ge \cdots \ge w_n \ge 0$ in the vector w.

Fact 5.4. *If $\|w\|_2 \le 1$, then $\|w\|_\ell^\ell$ is always between w_1^ℓ and $w_1^{\ell-2}$ for any $\ell \ge 3$.*

Proof. $w_1^\ell \le \|w\|_\ell^\ell = \sum_{i=1}^{n} w_i^\ell \le w_1^{\ell-2} \sum_{i=1}^{n} w_i^2 \le w_1^{\ell-2}$. □

Proof of Lemma 5.3. Recall that by assumption we have $w_1 = \|w\|_\infty \le 1$. Let θ denote $M_\ell^{1/\ell}$ and $\Delta \le (\delta/5)^\ell/2$ denote the error such that $M_\ell = \|w\|_\ell^\ell \pm \Delta$. We consider two cases based on the size of w_1:

(1) The first case is that $w_1 \le \delta/5$. In this case we upper bound θ by

$$\begin{aligned}
&(\|w\|_\ell^\ell + \Delta)^{1/\ell} \\
\le &(w_1^{\ell-2} + \Delta)^{1/\ell} \quad \text{(using the upper bound from Fact 5.4 on } \|w\|_\ell^\ell) \\
\le &((\delta/5)^{\ell-2} + (\delta/5)^\ell/2)^{1/\ell} \\
\le &2^{1/\ell} \cdot (\delta/5)^{(\ell-2)/\ell} \\
\le &2^{1/\ell} \cdot (5/\delta)^{2/\ell} \cdot \delta/5 \qquad \text{(using the fact } \ell = 100/\delta^3) \\
\le &2\delta/5.
\end{aligned}$$

So we have that $|\theta - w_1| \le 2\delta/5 + w_1$, which is at most $3\delta/5$ by the assumption of w_1 and the Lemma.

(2) The second case is that $w_1 > \delta/5$. In this case we first bound $w_1 - \theta$ by

$$\begin{aligned}
w_1 - (\|w\|_\ell^\ell - \Delta)^{1/\ell} &\le w_1 - (w_1^\ell - \Delta)^{1/\ell} \\
&\quad \text{(using the lower bound from Fact 5.4 on } \|w\|_\ell^\ell) \\
&= w_1 - w_1(1 - \frac{\Delta}{w_1^\ell})^{1/\ell} \\
&\le w_1 - w_1(1 - 2\frac{\Delta}{\ell \cdot w_1^\ell}) \\
&\quad \text{(using } (1-x)^{1/\ell} \ge 1 - 2x/\ell \text{ when } x \le 1/2) \\
&= 2w_1 \cdot \frac{\Delta}{\ell \cdot w_1^\ell}.
\end{aligned}$$

Then we bound $\theta - w_1$ by

$$\begin{aligned}
&(\|w\|_\ell^\ell + \Delta)^{1/\ell} - w_1 \\
&\le (w_1^{\ell-2} + \Delta)^{1/\ell} - w_1 \\
&\quad \text{(using the upper bound from Fact 5.4 on } \|w\|_\ell^\ell) \\
&\le (w_1^{\ell-2} + \Delta)^{1/\ell} - (w_1^\ell + \Delta)^{1/\ell} + (w_1^\ell + \Delta)^{1/\ell} - w_1 \\
&\le (w_1^\ell + \Delta)^{1/\ell} \cdot \left(\left(\frac{w_1^{\ell-2} + \Delta}{w_1^\ell + \Delta}\right)^{1/\ell} - 1\right) + w_1\left(1 + \frac{\Delta}{w_1^\ell}\right)^{1/\ell} - w_1 \\
&\le (w_1^\ell + \Delta)^{1/\ell} \cdot \left(\left(1 + \frac{w_1^{\ell-2}(1-w_1^2)}{w_1^\ell + \Delta}\right)^{1/\ell} - 1\right) \\
&\qquad + w_1\left(1 + \frac{\Delta}{\ell \cdot w_1^\ell}\right) - w_1 \qquad \text{(using } (1+x)^{1/\ell} \le 1 + x/\ell) \\
&\le (w_1^\ell + \Delta)^{1/\ell} \cdot \frac{w_1^{\ell-2}}{\ell(w_1^\ell + \Delta)} + w_1 \frac{\Delta}{\ell \cdot w_1^\ell}. \\
&\qquad \text{(using } (1+x)^{1/\ell} \le 1 + x/\ell \text{ again)}
\end{aligned}$$

We combine the above two bounds to get that

$$|\theta - w_1| \le (w_1^\ell + \Delta)^{1/\ell} \cdot \frac{w_1^{\ell-2}}{\ell(w_1^\ell + \Delta)} + 3w_1 \frac{\Delta}{\ell \cdot w_1^\ell}.$$

Plugging in our bounds on w_1 and $\Delta \le (\delta/5)^\ell/2$ into this inequality, this is at most

$$\begin{aligned}
&(w_1^\ell + \Delta)^{1/\ell} \cdot \frac{1}{\ell \cdot w_1^2} + 3\frac{\Delta}{\ell \cdot w_1^{\ell-1}} \\
&\le (w_1 + \Delta^{1/\ell}) \cdot \frac{1}{\ell \cdot w_1^2} + 3\frac{\Delta}{\ell w_1^{\ell-1}} \\
&\qquad \text{(using } (x+y)^{1/\ell} \le x^{1/\ell} + y^{1/\ell}\text{)} \\
&\le \frac{1}{\ell \cdot \delta/5} + \frac{\delta/5}{\ell \cdot (\delta/5)^2} + \frac{3 \cdot (\delta/5)^\ell/2}{\ell \cdot (\delta/5)^{\ell-1}} \\
&\qquad \text{(since in this case } w_1 \ge \delta/5\text{)} \\
&\le \frac{\delta}{2}. \qquad \text{(using } \ell \ge 100/\delta^3\text{)}
\end{aligned}$$

□

6 BOUNDING THE GAP BETWEEN NON-ZERO CUMULANTS

The result of Marcinkiewicz (Theorem 2.1) shows that any non-Gaussian random variable $\boldsymbol{X}$ has an infinite number of non-zero cumulants. However, this result is not constructive and leaves open two obvious questions:

(1) Suppose $\kappa_\ell(\boldsymbol{X}) \neq 0$. What can we say about the quantity

$$\arg\min_{\ell' > \ell} \kappa_{\ell'}(\boldsymbol{X}) \neq 0?$$

In other words, how many consecutive zero cumulants can $\boldsymbol{X}$ have following the non-zero cumulant $\kappa_\ell(\boldsymbol{X})$?

(2) Merely having a non-zero cumulant $\kappa_{\ell'}(\boldsymbol{X})$ is not sufficient for us; since our results depend on the magnitude of the non-zero cumulants, we would also like a lower bound on the magnitude of $\kappa_{\ell'}(\boldsymbol{X})$ (where ℓ' is as defined above). Can we get such a lower bound on $\kappa_{\ell'}(\boldsymbol{X})$?

The main result of this section is to give an *effective* answer to both these questions when the random variable $\boldsymbol{X}$ has bounded support. To the best of our knowledge (and based on conversations with experts [12, 29, 35]), previously no such effective bound was known for gaps between non-zero cumulants.

Before stating our result, we note that for any real random variable $\boldsymbol{X}$ the random variable $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{X}'$ (where $\boldsymbol{X}'$ is an independent copy of $\boldsymbol{X}$) is (i) symmetric and (ii) has $\kappa_\ell(\boldsymbol{Y}) = (1 + (-1)^\ell)\kappa_\ell(\boldsymbol{X})$. Thus for the purposes of this section, it suffices to restrict our attention to symmetric random variables and even-numbered cumulants.

Theorem 6.1. *Given any ℓ and any symmetric random variable $\boldsymbol{X}$ with unit variance and support $[-\mathrm{B}, \mathrm{B}]$, where $\mathrm{B} \ge 1$, there exists $\ell' = (\Theta(1) \cdot \mathrm{B}^4 \log \mathrm{B})^\ell$ such that $\left|\kappa_j(\boldsymbol{X})\right| \ge 2^{-\ell'}$ for some $j \in (\ell, \ell']$.*

Before delving into the formal proof of this theorem, we give a high-level overview. Recall that the cumulant generating function and moment generating function of $\boldsymbol{X}$ are defined respectively as

$$K_{\boldsymbol{X}}(z) = \sum_{j \ge 1} \frac{\kappa_j(\boldsymbol{X})}{j!} z^j; \quad M_{\boldsymbol{X}}(z) = \mathbf{E}[e^{z\boldsymbol{X}}].$$

The first main ingredient (Claim 6.2) is that the function $M_{\boldsymbol{X}}(z)$ has a root in the complex disc of radius $O(\mathrm{B}^3)$ centered at the origin. The proof of this is somewhat involved and uses a range of ingredients such as bounding the number of zeros of entire functions and the Hadamard factorization theorem.

Now, suppose it were the case that $|\kappa_j(\boldsymbol{X})| \le 2^{-\ell'}$ for all $j \in (\ell, \ell']$ for a sufficiently large ℓ'. We consider the "truncated" function $P_\ell(z)$

$$P_\ell(z) = \sum_{j=1}^{\ell} \frac{\kappa_j(\boldsymbol{X})}{j!} z^j.$$

Observe that while $K_{\boldsymbol{X}}(z)$ is not necessarily well defined everywhere, (i) it is easy to show that it is well defined in the open disc of radius $1/(e\mathrm{B})$ (call this set $\mathcal{B}$); (ii) the function $P_\ell(z)$ is an entire function. Further, since $\kappa_j(\boldsymbol{X})$ is assumed to have very small magnitude for all $j \in (\ell, \ell']$, it is not difficult to show that $P_\ell(z)$ and $K_{\boldsymbol{X}}(z)$ are close to each other in $\mathcal{B}$. Using $e^{K_{\boldsymbol{X}}(z)} = M_{\boldsymbol{X}}(z)$ in $\mathcal{B}$ (since both are well-defined), we infer that $e^{P_\ell(z)}$ and $M_{\boldsymbol{X}}(z)$ are also close to each other in $\mathcal{B}$. In other words, the function $h(z) := e^{P_\ell(z) - M_{\boldsymbol{X}}(z)}$ is close to zero in $\mathcal{B}$.

Finally, we observe that $e^{P_\ell(z)}$ has no zeros in $\mathbb{C}$ and in fact, we can show that it has relatively large magnitude within a ball of radius $O(\mathrm{B}^3)$. Using the first ingredient that $M_{\boldsymbol{X}}(z)$ has a zero in this disc, we derive that the maximum of $|h(z)|$ is large in a disc of radius $O(\mathrm{B}^3)$. However, since h is an entire function, once ℓ' is sufficiently large, this contradicts the fact that $h(z)$ is close to zero in $\mathcal{B}$ (this uses Hadamard's three circle theorem). This finishes the proof.

Proof of Theorem 6.1. Towards a contradiction, fix $\zeta = 2^{-\ell'}$ and let us assume that $|\kappa_j(\boldsymbol{X})| < \zeta$ for $j \in (\ell, \ell']$. Let us consider the moment generating function $M_{\boldsymbol{X}} : \mathbb{C} \to \mathbb{C}$ defined by $M_{\boldsymbol{X}}(z) = \mathbf{E}[e^{z\boldsymbol{X}}]$. From the fact that the random variable $\boldsymbol{X}$ is bounded in $[-\mathrm{B}, \mathrm{B}]$, it follows that the function $M_{\boldsymbol{X}}$ is an entire function (i.e., holomorphic over all of $\mathbb{C}$). Next, consider the cumulant generating function $K_{\boldsymbol{X}} : \mathbb{C} \to \mathbb{C}$ defined as

$$K_{\boldsymbol{X}}(z) = \sum_{j \ge 1} \frac{\kappa_j(\boldsymbol{X})}{j!} z^j.$$

From Claim 3.6, we know that $|\kappa_j(\boldsymbol{X})| \le \mathrm{B}^j \cdot e^j \cdot j!$. Define the open disc $\mathcal{B} = \{z : |z| < 1/(e\mathrm{B})\}$ and observe that the right hand side series is absolutely convergent in $\mathcal{B}$ and hence $K_{\boldsymbol{X}}$ is holomorphic in $\mathcal{B}$. We recall from the definition of cumulants that for $z \in \mathcal{B}$, $e^{K_{\boldsymbol{X}}(z)} = M_{\boldsymbol{X}}(z)$.

We will need the following claim about the roots of $M_{\boldsymbol{X}}$:

Claim 6.2. *For any symmetric random variable $\boldsymbol{X}$ with unit variance and support $[-\mathrm{B}, \mathrm{B}]$ where $\mathrm{B} \ge 1$, there exists z_0 with $|z_0| \le 200\mathrm{B}^3$ such that $M_{\boldsymbol{X}}(z_0) = \mathbf{E}[e^{z_0 \boldsymbol{X}}] = 0$.*

We defer the proof of Claim 6.2 to Section 6.1. Let us define $P_\ell(z)$ to be the polynomial obtained by truncating the cumulant generating function Taylor series expansion to degree ℓ, so $P_\ell(z) = \sum_{1 \le j \le \ell} \frac{\kappa_j(\boldsymbol{X})}{j!} z^j$. We now define the function $g : \mathbb{C} \to \mathbb{C}$ as

$$g(z) = e^{P_\ell(z)} - \mathbf{E}[e^{z\boldsymbol{X}}]. \tag{6}$$

Observe that g is an entire function. The following claim lower bounds the magnitude of g on the point z_0 defined above:

Claim 6.3. *Let z_0 be the complex number satisfying $\mathbf{E}[e^{z_0 X}] = 0$ in Claim 6.2. Then we have*

$$|g(z_0)| \geq e^{-2\cdot(200e)^\ell \cdot \mathrm{B}^{4\ell}}.$$

Proof. We have

$$\begin{aligned}|P_\ell(z_0)| &\leq \sum_{j=1}^{\ell} \frac{|\kappa_j(\boldsymbol{X})|}{j!}|z_0|^j \\ &\leq \sum_{j=1}^{\ell} e^j \cdot \mathrm{B}^j \cdot |z_0|^j \leq 2(e\mathrm{B})^\ell \cdot (200\mathrm{B}^3)^\ell.\end{aligned}$$

The first inequality is just a triangle inequality whereas the second inequality uses Claim 3.6. Since $\mathbf{E}[e^{z_0 X}] = 0$, we get that $|g(z_0)| \geq e^{-|P_\ell(z_0)|} \geq e^{-2\cdot(200e)^\ell \cdot \mathrm{B}^{4\ell}}$. □

We now recall the Hadamard three-circle theorem.

Theorem 6.4 (Hadamard three-circle theorem). *Let $0 < r_1 < r_2 < r_3$ and let h be an analytic function on the annulus $\{z \in \mathbb{R} : |z| \in [r_1, r_3]\}$. Let $M_h(r)$ denote the maximum of $h(z)$ on the circle $|z| = r$. Then,*

$$\ln \frac{r_3}{r_1} \ln M_h(r_2) \leq \ln \frac{r_3}{r_2} \ln M_h(r_1) + \ln \frac{r_2}{r_1} \ln M_h(r_3).$$

We are now ready to finish the proof of Theorem 6.1. The proof uses the following claim:

Claim 6.5. *There is a point z_* satisfying*

$$|z_*| \leq \frac{1}{2e\mathrm{B}} \quad \text{and } |g(z_*)| \geq e^{-12\left((400e)^\ell \cdot \mathrm{B}^{4\ell} \cdot \ln(400e\mathrm{B})\right)}. \tag{7}$$

Proof. Recall from Claim 6.2 that the point z_0 satisfies $|z_0| \leq 200\mathrm{B}^3$ and $M_{\boldsymbol{X}}(z_0) = \mathbf{E}[e^{z_0 X}] = 0$. There are two cases:

(1) If $|z_0| \leq \frac{1}{2e\mathrm{B}}$: In this case, we set $z_* = z_0$. By definition, it satisfies the first condition in (7) and using Claim 6.3, it satisfies the second condition.
(2) If $|z_0| > \frac{1}{2e\mathrm{B}}$: Set $r_1 = \frac{1}{2e\mathrm{B}}$, $r_2 = |z_0|$ and $r_3 = er_2$. For g as defined in (6), from Claim 6.3, we have

$$M_g(r_2) \geq |g(z_0)| = |e^{P_\ell(z_0)} - \mathbf{E}[e^{z_0 X}]| \geq e^{-2\cdot(200e)^\ell \cdot \mathrm{B}^{4\ell}}. \tag{8}$$

On the other hand, consider any point z_3 such that $|z_3| = r_3$. We have that

$$\begin{aligned}|g(z_3)| &\leq |e^{P_\ell(z_3)}| + |\mathbf{E}[e^{z_3 X}]| \\ &\leq e^{\sum_{j=1}^{\ell}(e\mathrm{B})^j \cdot |z_3|^j} + \int_{-\mathrm{B}}^{\mathrm{B}} \mathbf{Pr}[\boldsymbol{X} = x] \cdot e^{|z_3|\cdot x}\mathrm{d}x \\ &\leq e^{\sum_{j=1}^{\ell}(e\mathrm{B})^j \cdot r_3^j} + \int_{-\mathrm{B}}^{\mathrm{B}} \mathbf{Pr}[\boldsymbol{X} = x] \cdot e^{r_3 \cdot x}\mathrm{d}x \leq e^{2(e\mathrm{B})^\ell \cdot r_3^\ell},\end{aligned}$$

and hence

$$|M_g(r_3)| \leq e^{2(e\mathrm{B})^\ell \cdot r_3^\ell}. \tag{9}$$

Now observe that the function g defined in (6) is an entire function. Consequently, using $r_3/r_2 = e$, we can apply the Hadamard Three Circle Theorem to g to obtain

$$\begin{aligned}&\ln M_g(r_1) \\ &\geq \ln \frac{r_3}{r_1} \ln M_g(r_2) - \ln \frac{r_2}{r_1} \ln M_g(r_3) \\ &\geq -\left(2\cdot(200e)^\ell \cdot \mathrm{B}^{4\ell}\right) \ln \frac{r_3}{r_1} - \left(2(e\mathrm{B})^\ell \cdot (er_2)^\ell\right) \ln \frac{r_2}{r_1} \\ &\qquad \text{(applying (8), (9))} \\ &\geq -\left(3(400e)^\ell \cdot \mathrm{B}^{4\ell} \ln(400e\mathrm{B}^4)\right).\end{aligned}$$

The last inequality uses that $r_2 \leq 200\mathrm{B}^3$ (from Claim 6.2). This implies the existence of a point z_* satisfying (7) and concludes the proof of Claim 6.5. □

Continuing with the proof of Theorem 6.1, observe that the Taylor expansion for $K_{\boldsymbol{X}}(z)$ (at $z = 0$) converges absolutely in $\mathcal{B}$. Thus $K_{\boldsymbol{X}}(z)$ is holomorphic in $\mathcal{B}$ and is given by its Taylor expansion. Since $z_* \in \mathcal{B}$, recalling our initial assumption that the $(\ell+1)$-th through ℓ'-th cumulants all have magnitude at most ζ, we have that

$$\begin{aligned}|K_{\boldsymbol{X}}(z_*) - P_\ell(z_*)| &\leq \sum_{j=\ell+1}^{\ell'} \frac{\zeta \cdot |z_*|^j}{j!} + \sum_{j>\ell'} \frac{|\kappa_j(\boldsymbol{X})| \cdot |z_*|^j}{j!} \\ &\leq \sum_{j=\ell+1}^{\ell'} \frac{\zeta}{j! \cdot (2e\mathrm{B})^j} + \sum_{j>\ell'} \frac{\mathrm{B}^j \cdot e^j \cdot j!}{(2e\mathrm{B})^j \cdot j!} \\ &\qquad \text{(using Claim 3.6 and (7))} \\ &\leq \frac{2\zeta}{\ell! \cdot (2e\mathrm{B})^\ell} + 2^{-\ell'} \leq 2^{-\ell'+1}. \end{aligned} \tag{10}$$

Since $K_{\boldsymbol{X}}(z)$ is holomorphic in $\mathcal{B}$, so is $M_{\boldsymbol{X}}(z) = e^{K_X(z)}$, and we have that

$$\begin{aligned}\left|M_{\boldsymbol{X}}(z_*) - e^{P_\ell(z_*)}\right| &= \left|e^{K_X(z_*)} - e^{P_\ell(z_*)}\right| \\ &= \left|e^{P_\ell(z_*)}\right| \cdot \left|e^{K_X(z_*) - P_\ell(z_*)} - 1\right| \\ &\leq \left|e^{P_\ell(z_*)}\right| \cdot 2^{-\ell'+2}, \end{aligned} \tag{11}$$

where the last inequality is by Equation (10). However, applying Claim 3.6 and recalling that $|z_*| \leq 1/(2e\mathrm{B})$, we also have

$$\left|P_\ell(z_*)\right| \leq \sum_{j=1}^{\ell} \frac{|z_*|^j \kappa_j(\boldsymbol{X})}{j!} \leq \sum_{j=1}^{\ell} \frac{|z_*|^j \cdot \mathrm{B}^j \cdot e^j \cdot j!}{j!} \leq 1.$$

Plugging this back into (11), we get

$$\left|M_{\boldsymbol{X}}(z_*) - e^{P_\ell(z_*)}\right| \leq 4e \cdot 2^{-\ell'}.$$

However, recalling that $g(z) = M_{\boldsymbol{X}}(z) - e^{P_\ell(z)}$, this contradicts (7) with room to spare provided that, say,

$$\ell' > 50\left((400e)^\ell \cdot \mathrm{B}^{4\ell} \cdot \ln(400e\mathrm{B})\right).$$

This finishes the proof of Theorem 6.1 □

6.1 Proof of Claim 6.2

We start by showing that the function $M_{\boldsymbol{X}}(z)$ must necessarily decay along the line $\{z : \mathrm{Re}(z) = 0\}$ close to the origin.

Claim 6.6. *For the symmetric random variable $\boldsymbol{X}$, which has unit variance and is supported on $[-\mathsf{B}, \mathsf{B}]$ where $\mathsf{B} \geq 1$, there exists $\alpha_* \in \mathbb{R}$ such that $|\alpha_*| \leq 3$, $M_{\boldsymbol{X}}(i\alpha_*) \in \mathbb{R}$ and $|M_{\boldsymbol{X}}(i\alpha_*)| \leq 1 - \frac{1}{2\mathsf{B}^2}$.*

Proof. First of all, note that by symmetry of $\boldsymbol{X}$, $M_{\boldsymbol{X}}(i\alpha) = \mathbf{E}[e^{i\alpha \boldsymbol{X}}]$ is necessarily real-valued for any $\alpha \in \mathbb{R}$. Next, choose $F > 1$ (we will fix its exact value soon). We have

$$\begin{aligned}
&\int_{\alpha=-F}^{F} M_{\boldsymbol{X}}(2\pi i\alpha) d\alpha \\
&= \int_{\alpha=-F}^{F} \int_{x=-\mathsf{B}}^{\mathsf{B}} \mathbf{Pr}[\boldsymbol{X} = x] \cdot e^{2\pi i\alpha x} dx d\alpha \\
&= \int_{x=-\mathsf{B}}^{\mathsf{B}} \mathbf{Pr}[\boldsymbol{X} = x] \int_{\alpha=-F}^{F} e^{2\pi i\alpha x} d\alpha dx \\
&\leq \int_{|x|\leq 1/2} 2\,\mathbf{Pr}[\boldsymbol{X} = x] \cdot F dx + \\
&\quad \int_{|x|>1/2} \mathbf{Pr}[\boldsymbol{X} = x] \cdot \frac{\sin(2\pi F x)}{\pi x} dx. \qquad (12)
\end{aligned}$$

Now, observe that

$$1 = \mathbf{E}[|\boldsymbol{X}|^2] \leq \mathbf{Pr}[|\boldsymbol{X}| > 1/2] \cdot \mathsf{B}^2 + (1 - \mathbf{Pr}[|\boldsymbol{X}| > 1/2]) \cdot \frac{1}{4}.$$

Thus, we obtain

$$\mathbf{Pr}[|\boldsymbol{X}| > 1/2] \geq \frac{3}{4\left(\mathsf{B}^2 - \frac{1}{4}\right)} \geq \frac{3}{4\mathsf{B}^2}. \qquad (13)$$

Likewise, observe that $\sin(2\pi F x)/(\pi x)$ always has magnitude at most $2/\pi$ for $|x| > 1/2$. Plugging (13) and this back into (12), and using $F > 1$, we have that

$$\int_{\alpha=-F}^{F} M_{\boldsymbol{X}}(2\pi i\alpha) d\alpha \leq \left(1 - \frac{3}{4\mathsf{B}^2}\right) \cdot 2F + \frac{3}{4\mathsf{B}^2} \cdot \frac{2}{\pi}.$$

This implies that there is a point $\alpha_* \in [-F, F]$ such that

$$M_{\boldsymbol{X}}(2\pi i\alpha_*) \leq 1 - \frac{3}{4\mathsf{B}^2} + \frac{3}{4\mathsf{B}^2 \pi F}.$$

Plugging in $F = 3$, we get the claim. □

Observe that $M_{\boldsymbol{X}}(z)$ is an entire function and thus is well defined on all of $\mathbb{C}$. The next lemma bounds the number of zeros of $M_{\boldsymbol{X}}(z)$ in a ball of radius R. This is essentially the same as the first part of Theorem 2.1 in [40], though the bound given there is asymptotic whereas we need a precise quantitative bound.

Claim 6.7. *For $M_{\boldsymbol{X}}(z)$ as defined above and $r > 0$, let $n(R)$ denote the number of zeros of $M_{\boldsymbol{X}}(z)$ contained in the ball $\{|z| : |z| \leq R\}$ (counting multiplicities). Then $n(R) \leq e\mathsf{B}R$.*

Before proceeding with the proof of Claim 6.7, we recall a useful ingredient, namely, Jensen's formula (see Theorem 1.1, Section 5 in [40]):

Theorem 6.8 (Jensen's formula). *Let h be an analytic function in a region of $\mathbb{C}$ which contains the closed disc $\mathbf{D} = \{z : |z| \leq R\}$. Suppose $h(0) \neq 0$ and h does not have zeros on the boundary $\partial\mathbf{D} = \{z : |z| = R\}$. Then*

$$\int_0^1 \ln\left|h(Re^{2\pi i t})\right| \mathrm{d}t = \ln|h(0)| + \sum_{z:|z|<R,\ h(z)=0} \ln\frac{R}{|z|},$$

where the summation on the right hand side counts the roots of h with multiplicity.

Proof of Claim 6.7. First since $M_{\boldsymbol{X}}(z)$ is an entire function, its zeros are isolated. Thus, by perturbing R infinitesimally, we can assume that $M_{\boldsymbol{X}}(z)$ has no zeros on $\partial\mathbf{D}$. Further, an immediate consequence of the Jensen's formula is that the number of zeros of an analytic function in $\mathbf{D}$ must be finite. To see this, let $R' > R$ and apply Jensen's formula on the cicle of radius R'. By Jensen's formula, it follows that $\sum_{z:|z|<R',\ h(z)=0} \ln\frac{R'}{|z|}$ is finite which implies that the number of zeros in $\mathbf{D}$ has to be finite. For any radius R_*, let us now enumerate the zeros of $M_{\boldsymbol{X}}(z)$ that lie within the disc $\{z \in \mathbb{R} : |z| \leq R_*\}$ as $z_1, \ldots, z_{n(R_*)}$ such that $|z_1| \leq |z_2| \ldots \leq |z_{n(R_*)}|$. Then,

$$\begin{aligned}
&\sum_{i=1}^{n(R_*)} \ln\frac{R_*}{|z_i|} \\
&= \sum_{i=1}^{n(R_*)-1} i \cdot \ln\frac{|z_{i+1}|}{|z_i|} + n(R_*) \cdot \ln\frac{R_*}{|z_{n(R_*)}|} \\
&= \int_0^{R_*} n(r)\frac{\mathrm{d}r}{r}. \qquad (14)
\end{aligned}$$

The last equality simply follows by observing that since $n(r)$ is finite in $[0, R_*]$, hence we can split the integral on the right hand side at the points of discontinuity of $n(r)$. Next, we have

$$\begin{aligned}
n(R) &\leq n(R) \int_R^{eR} \frac{\mathrm{d}r}{r} \leq \int_R^{eR} n(r)\frac{\mathrm{d}r}{r} \\
&\leq \int_0^{eR} n(r)\frac{\mathrm{d}r}{r} = \sum_{i=1}^{n(eR)} \ln\frac{eR}{|z_i|}. \qquad (15)
\end{aligned}$$

In the above, the first three inequalities follow by definition while the last equality is an application of (14) with $R_* = eR$. Finally, by definition, $M_{\boldsymbol{X}}(0) = 1$ and $\ln|M_{\boldsymbol{X}}(z)| \leq \mathsf{B}|z|$. Using these two facts with (15) and Theorem 6.8, we get

$$n(R) \leq \sum_{i=1}^{n(eR)} \ln\frac{eR}{|z_i|} = \int_0^1 \ln\left|M_{\boldsymbol{X}}(eRe^{2\pi i t})\right| \mathrm{d}t \leq e\mathsf{B}R.$$

This finishes the proof of Claim 6.7. □

Corollary 6.9. *Let $M_{\boldsymbol{X}}(z) = \mathbf{E}[e^{z\boldsymbol{X}}]$ as defined earlier, and let $\alpha > 1$, $R_* > 0$ be such that $M_{\boldsymbol{X}}(z)$ has no roots in the ball $\{z : |z| \leq R_*\}$. Then,*

$$\sum_{z:M_{\boldsymbol{X}}(z)=0} \frac{1}{|z|^\alpha} \leq \frac{\alpha(\alpha-1)\cdot e\mathsf{B}}{R_*^{\alpha-1}}.$$

Proof. It follows from Claim 6.7 that the number of roots of $M_{\boldsymbol{X}}(z)$ is countable. Let us enumerate these roots as $z_1, z_2, \ldots$. We have that

$$\begin{aligned}
\sum_{z:M_{\boldsymbol{X}}(z)=0} \frac{1}{|z|^\alpha} &= \sum_i \frac{1}{|z_i|^\alpha} = \sum_i \alpha \int_{|z_i|}^{\infty} \frac{1}{r^{1+\alpha}} \mathrm{d}r \\
&= \alpha \int_0^{\infty} \frac{n(r)}{r^{1+\alpha}} \mathrm{d}r.
\end{aligned}$$

From the assumption $n(r) = 0$ for all $r \le R_*$, we can use Claim 6.7 to upper bound the right hand side as

$$\alpha \int_{R_*}^{\infty} \frac{e\mathrm{B}r}{r^{1+\alpha}}\mathrm{d}r = \alpha \cdot e\mathrm{B} \int_{R_*}^{\infty} \frac{1}{r^{\alpha}}\mathrm{d}r \le \frac{\alpha(\alpha-1)e\mathrm{B}}{R_*^{\alpha-1}}. \qquad \square$$

The last ingredient we will need to prove Claim 6.2 is the Hadamard factorization theorem (see Theorem 5.1, Section 5 in [40]):

Theorem 6.10. *Let h be an entire function that is of order 1 (i.e., $\log|h(z)| = O(|z|)$). If $h(0) \neq 0$, then there exist $A, A' \in \mathbb{R}$ such that*

$$h(z) = e^{Az+A'} \prod_{n\ge 1} \left(1 - \frac{z}{z_n}\right) e^{\frac{z}{z_n}},$$

where $z_1, z_2, \ldots$ are the roots of $h(z)$.

Proof of Claim 6.2. As $M_{\boldsymbol{X}}(z)$ is an entire function of order one, we can use Theorem 6.10 to express it as

$$M_{\boldsymbol{X}}(z) = e^{Az+A'} \prod_{n\ge 1} \left(1 - \frac{z}{z_n}\right) e^{\frac{z}{z_n}},$$

where $z_1, z_2, \ldots$ are the roots of $M_{\boldsymbol{X}}(z)$. We first recall that $M_{\boldsymbol{X}}(0) = 1$, and hence $A' = 0$. We next observe that since $M_{\boldsymbol{X}}(z)$ is a symmetric function, if z_n is a root then so is $-z_n$ (and with the same multiplicity). Together with the symmetry of $M_{\boldsymbol{x}}(z)$, this implies that the coefficient A of z appearing in the exponent is also zero. Next, we observe that $M_{\boldsymbol{X}}(z)$ cannot have any root on the real line. Thus, if we define $\Omega_1 = \{z : \mathrm{Re}(z) > 0\}$, then the right hand side of the above equation simplifies to

$$M_{\boldsymbol{X}}(z) = \prod_{z_i \in \Omega_1 : M_{\boldsymbol{X}}(z_i) = 0} \left(1 - \frac{z^2}{z_i^2}\right).$$

Now, suppose that $M_{\boldsymbol{X}}(z)$ does not have any zeros in a ball of radius R_* around the origin. Then, for any z such that $|z| \le R_*$, the above gives that

$$\begin{aligned} |M_{\boldsymbol{X}}(z)| &\ge \prod_{z_i \in \Omega_1 : M_{\boldsymbol{X}}(z_i) = 0} \left(1 - \frac{|z|^2}{|z_i|^2}\right) \\ &\ge 1 - \sum_{z_i \in \Omega_1 : M_{\boldsymbol{X}}(z_i) = 0} \frac{|z|^2}{|z_i|^2}. \end{aligned}$$

Applying Corollary 6.9 (with $\alpha = 2$), we have that

$$|M_{\boldsymbol{X}}(z)| \ge 1 - \frac{2e\mathrm{B}|z|^2}{R_*}.$$

Choosing $R_* = 72e\mathrm{B}^3$, we get that $|M_{\boldsymbol{X}}(z)| \ge 1 - |z|^2/36\mathrm{B}^2$ for all $|z| \le 72e\mathrm{B}^3$. In particular, for all $|z| \le 3$, $M_{\boldsymbol{X}}(z) \ge 1 - 1/(4\mathrm{B}^2)$. This contradicts Claim 6.6. Thus, $M_{\boldsymbol{X}}(z)$ has a root of magnitude at most $72e\mathrm{B}^3 \le 200\mathrm{B}^3$. $\square$

REFERENCES

[1] Khanh Do Ba, Piotr Indyk, Eric Price, and David P Woodruff. 2010. Lower bounds for sparse recovery. In *Proceedings of the twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*. 1190–1197.

[2] Maria-Florina Balcan, Eric Blais, Avrim Bum, and Liu Yang. 2012. Active property testing. In *Symposium on Foundations of Computer Science, FOCS 2012*. 21–30.

[3] O.E. Barndorff-Nielsen. 1988. *Cumulants. In: Parametric Statistical Models and Likelihood. Lecture Notes in Statistics, vol 50.* Springer, New York.

[4] Piotr Berman, Sofya Raskhodnikova, and Grigory Yaroslavtsev. 2014. L_p-testing. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, David B. Shmoys (Ed.). ACM, 164–173.

[5] Arnab Bhattacharyya, Eldar Fischer, and Shachar Lovett. 2013. Testing low complexity affine-invariant properties. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 1337–1355.

[6] Arnab Bhattacharyya, Elena Grigorescu, and Asaf Shapira. 2015. A unified framework for testing linear-invariant properties. *Random Struct. Algorithms* 46, 2 (2015), 232–260.

[7] Eric Blais. 2009. Testing juntas nearly optimally. In *Proc. 41st Annual ACM Symposium on Theory of Computing (STOC)*. 151–158.

[8] E. Blais. 2010. Testing Juntas: A Brief Survey. In *Property Testing - Current Research and Surveys*. 32–40.

[9] Eric Blais, Clément L Canonne, Talya Eden, Amit Levi, and Dana Ron. 2019. Tolerant junta testing and the connection to submodular optimization and function isomorphism. *ACM Transactions on Computation Theory (TOCT)* 11, 4 (2019), 24.

[10] M. Blum, M. Luby, and R. Rubinfeld. 1993. Self-testing/correcting with applications to numerical problems. *J. Comp. Sys. Sci.* 47 (1993), 549–595. Earlier version in STOC'90.

[11] Wlodzimierz Bryc. 2005. Normal Distribution: characterizations with applications. https://homepages.uc.edu/~brycwz/probab/charakt/charakt.pdf

[12] W. Bryc. 2019. Personal communication.

[13] Nader H. Bshouty. 2019. Almost Optimal Distribution-Free Junta Testing. In *34th Computational Complexity Conference (CCC)*. 2:1–2:13.

[14] E. Candes. 2006. Compressive sampling. In *Proc. International Congress of Mathematicians, Madrid, Spain, Aug. 2006.*

[15] Emmanuel Candes and Justin Romberg. 2007. Sparsity and incoherence in compressive sampling. *Inverse problems* 23, 3 (2007), 969.

[16] Emmanuel J Candes. 2008. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique* 346, 9-10 (2008), 589–592.

[17] Emmanuel J Candes, Justin K Romberg, and Terence Tao. 2006. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics* 59, 8 (2006), 1207–1223.

[18] Xi Chen, Adam Freilich, Rocco A Servedio, and Timothy Sun. 2017. Sample-Based High-Dimensional Convexity Testing. In *APPROX/RANDOM 2017*.

[19] H. Chockler and D. Gutfreund. 2004. A lower bound for testing juntas. *Inform. Process. Lett.* 90, 6 (2004), 301–305.

[20] Eldar Fischer, Guy Kindler, Dana Ron, Shmuel Safra, and Alex Samorodnitsky. 2004. Testing juntas. *Journal of Computer and System Sciences* 68, 4 (2004), 753–787.

[21] E. Fischer, G. Kindler, D. Ron, S. Safra, and A. Samorodnitsky. 2004. Testing juntas. *J. Computer & System Sciences* 68, 4 (2004), 753–787.

[22] E. Fischer, G. Kindler, D. Ron, S. Safra, and A. Samorodnitsky. 2004. Testing juntas. *Journal of Computer & System Sciences* 68 (2004), 753–787.

[23] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

[24] Peter Gemmell, Richard Lipton, Ronitt Rubinfeld, Madhu Sudan, and Avi Wigderson. 1991. Self-testing/correcting for polynomials and for approximate functions. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*. ACM, 33–42.

[25] O. Goldreich (Ed.). 2010. *Property Testing: Current Research and Surveys*. Springer. LNCS 6390.

[26] O. Goldreich. 2017. *Introduction to Property Testing*. Cambridge University Press.

[27] Oded Goldreich and Dana Ron. 2016. On Sample-Based Testers. *ACM Transactions on Computation Theory (TOCT)* 8, 2 (2016), 7.

[28] William Greene. 2003. *Econometric analysis*. Pearson Education.

[29] S. Janson. 2019. Personal communication.

[30] T. Kaufman and M. Sudan. 2008. Algebraic property testing: the role of invariance. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)*. 403–412.

[31] Michael Kearns and Dana Ron. 2000. Testing problems with sublearning sample complexity. *Journal of Computer and System Sciences* 61, 3 (2000), 428–456.

[32] Weihao Kong and Gregory Valiant. 2018. Estimating learnability in the sublinear data regime. In *Advances in Neural Information Processing Systems*. 5455–5464.

[33] J. Marcinkiewicz. 1939. Sur une propriete de la loi de Gauss. *Mathematische Zeitschrift* 44 (1939), 612–618.

[34] Jozef Marcinkiewicz and Antoni Zygmund. 1964. *Collected papers*. Wydawnictwo Naukowe PWN, Poland.

[35] J. Neeman. 2019. Personal communication.

[36] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. 2006. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.* 72, 6 (2006), 1012–1042.

[37] Eric Price and David P. Woodruff. 2012. Applications of the Shannon-Hartley theorem to data streams and sparse recovery. In *Proceedings of the 2012 IEEE International Symposium on Information Theory, ISIT 2012, Cambridge, MA, USA, July 1-6, 2012*. 2446–2450.

[38] D. Ron. 2008. Property Testing: A Learning Theory Perspective. *Foundations and Trends in Machine Learning* 1, 3 (2008), 307–402.

[39] D. Ron. 2010. Algorithmic and Analysis Techniques in Property Testing. *Foundations and Trends in Theoretical Computer Science* 5 (2010), 73–205. Issue 2.

[40] Elias Stein and Rami Shakarchi. 2003. *Complex Analysis*. Princeton University Press, Princeton, New Jersey.

[41] Josh Swanson. 2017. Asymptotic Normality and Combinatorial Statistics. http://www.math.ucsd.edu/~jswanson/talks/2017_asymptotic_normality.pdf