Random Restrictions of High Dimensional Distributions and Uniformity Testing with Subcube Conditioning*

Clément L. Canonne† Xi Chen‡ Gautam Kamath§ Amit Levi¶ Erik Waingarten‖

Abstract

We give a nearly-optimal algorithm for testing uniformity of distributions supported on $\{-1,1\}^n$, which makes $\widetilde{O}(\sqrt{n}/\varepsilon^2)$ many queries to a subcube conditional sampling oracle (Bhattacharyya and Chakraborty (2018)). The key technical component is a natural notion of random restrictions for distributions on $\{-1,1\}^n$, and a quantitative analysis of how such a restriction affects the mean vector of the distribution. Along the way, we consider the problem of *mean testing* with independent samples and provide a nearly-optimal algorithm.

1 Introduction

The focus of this paper is high-dimensional distribution testing. The algorithmic problem is the following: we are granted oracle access (the type of which we will specify shortly) to a probability distribution p on $\Sigma = \{-1,1\}^n$, and must distinguish with probability at least $2/3$ between the case where p is the uniform distribution, and that where p is ε-far from uniform in total variation distance. The classical works of distribution testing [GGR96, GR00, BFR$^+$00] study the above question in the standard statistical setting, where the oracle provides independent samples from p. In this case, the hallmark results are an algorithm and a matching lower bound, showing that $\Theta(\sqrt{|\Sigma|}/\varepsilon^2)$ independent samples are necessary and sufficient for testing uniformity [Pan08, VV14]. When studying distributions supported on high-dimensional domains, unfortunately, this implies that the complexity of sample-optimal algorithms scales exponentially with the dimension, effectively making the problem intractable. To circumvent this issue, recent work has proceeded by either restricting the class of input distributions (e.g., restricting p to be a product distribution; see Section 1.3), or by allowing stronger oracle access. We take the latter approach, and consider an oracle access which is particularly well-suited to the high-dimensional structure: the *subcube conditional query model.*

Subcube conditional query access, first suggested in [CRS15] and studied in [BC18], allows algorithms to specify a subcube of the high-dimensional domain and request a sample from the distribution *conditioned* on the sample lying in the subcube specified — equivalently, to request samples after fixing some of their variables. The operation is akin to the notion of *restrictions* in the analysis of Boolean functions. Specifically, we identify the distribution by its probability mass function $p\colon \{-1,1\}^n \to \mathbb{R}_+$. An algorithm may then specify a subcube by a string $\rho \in \{-1,1,*\}^n$, where $*$'s denote free variables and non-$*$'s denote the values of restricted variables. Calling the oracle on such a ρ results in a sample from the distribution $p_{|\rho}$ (now supported on $\{-1,1\}^{\mathrm{stars}(\rho)}$) given by restricting the function $p_{|\rho}\colon \{-1,1\}^{\mathrm{stars}(\rho)} \to \mathbb{R}_+$ and re-normalizing it, so that it represents a distribution $p_{|\rho}$.[1]

Our main results are two-fold: (i) We define a natural notion of random restrictions for high-dimensional distributions and analyze the behavior of the mean vector of a distribution under such random restrictions (see Theorem 1.2 in Section 1.1); (ii) Leveraging this analysis, we obtain a nearly-optimal algorithm for testing uniformity over $\{-1,1\}^n$ with subcube conditioning. As stated below, subcube conditioning allows us to go from $2^{n/2}/\varepsilon^2$ sample complexity to $\sqrt{n}/\varepsilon^2$:

*The full version of this paper, with all proofs is available on https://arxiv.org/abs/1911.07357.

†IBM Research. ccanonne@cs.columbia.edu. Supported by a Goldstine Postdoctoral Fellowship. Part of this work was conducted while a Motwani Postdoctoral Fellow at Stanford University.

‡Columbia University. xichen@cs.columbia.edu. Supported by NSF IIS-1838154 and NSF CCF-1703925.

§Cheriton School of Computer Science, University of Waterloo. g@csail.mit.edu. Supported by an NSERC Discovery grant and a University of Waterloo startup grant. Part of this work was conducted while the author was supported by a Microsoft Research Fellowship, as part of the Simons-Berkeley Research Fellowship program, and while visiting Microsoft Research, Redmond.

¶Cheriton School of Computer Science, University of Waterloo. amit.levi@uwaterloo.ca. Research supported by the David R. Cheriton Graduate Scholarship.

‖Stanford University and the Simons Institute, eaw@cs.columbia.edu. This material is based upon work supported by the National Science Foundation under Award No. 2002201.

[1] When conditioning on a subcube with zero support, one may consider models where the oracle returns a uniform sample [CFGM16] or outputs "error" [CRS15]. We note that our algorithm will never run into this scenario.

Theorem 1.1. *(Main Result: Uniformity Testing) There exists an algorithm which, given subcube conditional query access to a distribution p supported on $\{-1,1\}^n$ and a distance parameter $\varepsilon \in (0,1)$, makes $\tilde{O}(\sqrt{n}/\varepsilon^2)$ queries and can distinguish with probability at least $2/3$ between the case when p is uniform, and when p is ε-far from uniform in total variation distance.*

Theorem 1.1 is tight up to poly-logarithmic factors. Indeed, as observed in [BC18], the sample complexity lower bound of $\Omega(\sqrt{n}/\varepsilon^2)$ of [CDKS17, DDK18] for testing uniformity of *product* distributions carries over to subcube conditional sampling.[2] Our result shows that, with subcube conditional queries, testing uniformity over arbitrary distributions is no harder than that over the much more restricted class of product ones.

Comparison with [BC18]. Theorem 1.1 improves the upper bound of $\tilde{O}(n^2/\varepsilon^2)$ of [BC18] for uniformity testing with subcube conditional queries, bringing it to the sublinear regime. The algorithm of [BC18] is based on a chain rule that, roughly speaking, bounds the mean of an *individual* coordinate of a distribution after a random restriction. In contrast, our algorithm applies new machinery (Theorem 1.2) developed to analyze the mean *vector* (its ℓ_2-norm, in particular) after a random restriction. Along the way, we study the *mean testing* problem, a natural variant of uniformity testing for high-dimensional distributions, and obtain optimal bounds for this question in the standard sampling model.

While our bounds for uniformity testing are quantitatively stronger, we note that the algorithm of [BC18] works for the problem of testing against any known distribution and over any product domains. Extending our results to these settings is an interesting direction for future work.

Comparison with [CJLW20]. In a simultaneous submission, [CJLW20] leverages techniques developed in the current paper for analyzing random restrictions and mean testing to study the *learning and testing of k-junta distributions* (uniformity testing can be viewed as the case when $k=0$). A set of new algorithmic primitives of independent interest is developed in [CJLW20] to deal with k-junta distributions, and a substantial component of [CJLW20] is in (nearly-optimal) lower bounds for learning and testing k-junta distributions. The algorithmic results of [CJLW20] demonstrate the potential of techniques developed in the current paper for attacking broader learning and testing problems with subcube conditioning.

[2] The reason is that, for product distributions, the coordinates are already independent, and therefore conditioning on subcubes does not grant any additional power.

1.1 Technical Ingredients

We start by reviewing the work of [CDKS17] for testing uniformity over product distributions.

Product Distributions and Mean Distance. The simplest class of distributions on $\{-1,1\}^n$ is arguably the class of *product distributions*, where all coordinates are independent. This setting was studied in [CDKS17], and is particularly nice to analyze due to the relation between the total variation distance between distributions and the ℓ_2 distance between their mean vectors. Specifically, let p be a product distribution supported on $\{-1,1\}^n$, and $\mu(p) \in [-1,1]^n$ be its *mean vector*,

$$\mu(p) = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim p}[\boldsymbol{x}] \in [-1,1]^n.$$

It is not hard to show that if p is ε-far from uniform in total variation distance, then $\|\mu(p)\|_2 \gtrsim \varepsilon$. Hence, for product distributions, large total variation distance to uniformity implies large mean vector in ℓ_2 norm. Given this fact, Canonne, Diakonikolas, Kane, and Stewart [CDKS17] design an algorithm based on estimating the norm of the mean vector, and show that $O(\sqrt{n}/\varepsilon^2)$ many samples from product distributions suffice to test uniformity.[3]

However, the relationship observed between distance to uniformity in total variation and *mean distance* (i.e., the ℓ_2 norm of the mean vector) for product distributions is not true in general. A simple example is the uniform distribution supported on just two vectors $\{x,-x\}$, which is very far from uniform yet has mean vector 0. Towards relating these two notions for general distributions, we define our notion of random restriction.

Random Restrictions. For any $\sigma \in [0,1]$, we write $\mathcal{S}_\sigma$ for the distribution supported on subsets of $[n]$ given by letting $\mathbf{S} \sim \mathcal{S}_\sigma$ include each index $i \in \mathbf{S}$ independently with probability σ. Given any distribution p supported on $\{-1,1\}^n$, let $\mathcal{D}_\sigma(p)$ be the distribution, supported on $\{-1,1,*\}^n$, of *random restrictions* of p. In order to sample a random restriction $\boldsymbol{\rho} \sim \mathcal{D}_\sigma(p)$, we sample a set $\mathbf{S} \sim \mathcal{S}_\sigma$ and a sample $\boldsymbol{x} \sim p$; then, we let $\boldsymbol{\rho}_i$ be set according to:

$$\boldsymbol{\rho}_i = \begin{cases} * & \text{if } i \in \mathbf{S} \\ \boldsymbol{x}_i & \text{if } i \notin \mathbf{S} \end{cases}. \tag{1.1}$$

For any $\rho \in \{-1,1,*\}^n$, we denote by $p_{|\rho}$ the distribution on $\{-1,1\}^{\mathrm{stars}(\rho)}$ given by $\boldsymbol{x}_{\mathrm{stars}(\rho)}$ where $\boldsymbol{x}$ is drawn from p *conditioned* on every $i \notin \mathrm{stars}(\rho)$ being set to ρ_i. This defines the restriction of a distribution p. Another operation on distributions we will consider is that of

[3] In fact, they consider the more general problem of identity testing of product distributions, where they obtain analogous results.

projection; for any set $S \subset [n]$, we write $\overline{S} = [n] \setminus S$, and the distribution $p_{\overline{S}}$ supported on $\{-1,1\}^{\overline{S}}$ is given by letting $\boldsymbol{y} \sim p_{\overline{S}}$ be $\boldsymbol{y} = \boldsymbol{x}_{\overline{S}}$ for $\boldsymbol{x} \sim p$.

Let $d_{\mathrm{TV}}(p, \mathcal{U})$ denote the total variation distance between p and the uniform distribution of the same dimension. At a high level, our main technical result shows that *the mean distance of random restrictions is implied by total variation distance of random projections.*

THEOREM 1.2. *(Informal version; see Theorem 1.4) Let p be any distribution over $\{-1,1\}^n$. Then,*

$$\mathop{\mathbf{E}}_{\boldsymbol{\rho}\sim\mathcal{D}_\sigma(p)}\left[\|\mu(p_{|\boldsymbol{\rho}})\|_2\right] \geq \sigma \cdot \mathop{\mathbf{E}}_{\mathbf{S}\sim\mathcal{S}_\sigma}\left[d_{\mathrm{TV}}(p_{\overline{\mathbf{S}}}, \mathcal{U})\right]. \tag{1.2}$$

Although the above differs from Theorem 1.4 in certain respects (the inequality in Theorem 1.4 incurs additional poly-logarithmic factors as well as a small additive error), (1.2) captures the key relationship between the total variation distance and the mean distance that we leverage, and will provide intuition for the introduction.

Mean Testing. The above discussion naturally leads to the following problem, which we refer to as *mean testing*. Given sample access to a distribution p supported on $\{-1,1\}^n$, we seek to distinguish with probability at least $2/3$ between the case where p is uniform and that where p has a large mean vector, i.e., $\|\mu(p)\|_2 \geq \varepsilon\sqrt{n}$. When p is assumed to be a product distribution, [CDKS17, DDK18] showed that for $\varepsilon \leq 1/\sqrt{n}$ the sample complexity of the problem is $\Theta(1/(\varepsilon^2\sqrt{n}))$.[4] The idea is that the empirical mean of a product distribution will have norm concentrated around its true value, and thus it suffices to look at this empirical estimate. In our case, however, additional care is needed, as there can be arbitrary correlations between coordinates of a sample and this concentration does not hold in general. Nevertheless, we present an algorithm for mean testing which is optimal up to a triply-logarithmic small loss in the sample complexity.

THEOREM 1.3. (MEAN TESTING) *There exists an algorithm which given*

$$O\left(\max\left\{\frac{1}{\varepsilon^2\sqrt{n}}, \frac{1}{\varepsilon}\right\}\right)$$

i.i.d. samples from an arbitrary distribution p on $\{-1,1\}^n$ and a parameter $\varepsilon \in (0,1]$ can distinguish with probability at least $2/3$ between (i) p is the uniform distribution, and (ii) $\|\mu(p)\|_2 \geq \varepsilon\sqrt{n}$.

[4]This is implicit in [CDKS17, Theorem 4.1]. As stated, their result is suited for distinguishing between the mean vector having norm 0 or at least ε', where $\varepsilon' \in (0,1]$: we re-parameterize with $\varepsilon' = \varepsilon\sqrt{n}$.

Moreover, as detailed in Section 4.1, the above immediately implies a similar sample complexity for *Gaussian* mean testing, where one is given i.i.d. samples from a multivariate normal distribution $p = \mathcal{G}(\mu, \Sigma)$ and must distinguish between $p = \mathcal{G}(0, I)$ and $\|\mu\|_2 \geq \varepsilon\sqrt{n}$.

Uniformity Testing with Subcube Conditioning. In view of the above discussion, we aim to use random restrictions and Theorem 1.3 to test uniformity with subcube conditional queries. The final technical ingredient is the following inequality, very similar to the "chain rule" of Bhattacharyya and Chakraborty [BC18] suited for uniformity testing on $\{-1,1\}^n$.

LEMMA 1.1. *Let p be a distribution supported on $\{-1,1\}^n$. Then, for any $\sigma \in [0,1]$,*

$$d_{\mathrm{TV}}(p, \mathcal{U}) \leq \mathop{\mathbf{E}}_{\mathbf{S}\sim\mathcal{S}_\sigma}\left[d_{\mathrm{TV}}(p_{\overline{\mathbf{S}}}, \mathcal{U})\right] + \mathop{\mathbf{E}}_{\boldsymbol{\rho}\sim\mathcal{D}_\sigma(p)}\left[d_{\mathrm{TV}}(p_{|\boldsymbol{\rho}}, \mathcal{U})\right].$$

This lemma naturally leads to a recursive approach for uniformity testing. Given a distribution p which is ε-far from uniform, either the total variation of random projections is large, or the total variation of random restrictions is large. In the former case, we apply (1.2), which allows us to reduce the problem to that of mean testing, and invoke Theorem 1.3. In the latter case, we take a random restriction and recurse (but on far fewer variables).

1.2 Proof Overview

We now formally state and explain the intuition behind our main theorem, relating the distance to uniformity of random projections to the mean distance after random restrictions.

THEOREM 1.4. *Let p be any distribution over $\{-1,1\}^n$ and $\sigma \in [0,1]$. Then,*

$$\mathop{\mathbf{E}}_{\boldsymbol{\rho}\sim\mathcal{D}_\sigma(p)}\left[\|\mu(p_{|\boldsymbol{\rho}})\|_2\right] \geq \frac{\sigma}{\mathrm{poly}(\log n)} \cdot$$
$$\widetilde{\Omega}\left(\mathop{\mathbf{E}}_{\mathbf{S}\sim\mathcal{S}_\sigma}[d_{\mathrm{TV}}(p_{\overline{\mathbf{S}}}, \mathcal{U})] - 2e^{-\min(\sigma,1-\sigma)n/10}\right).$$

We encourage the reader to think of applying Theorem 1.4 to a distribution p which is ε-far from uniform, and to think of the case when the parameter σ is a small constant (or inverse of a poly-logarithmic factor), and $\mathbf{E}_{\mathbf{S}\sim\mathcal{S}_\sigma}[d_{\mathrm{TV}}(p_{\overline{\mathbf{S}}}, \mathcal{U})] = \Theta(\varepsilon)$. Specifically, consider a parameter setting where $e^{-\min(\sigma,1-\sigma)n/10} = o(\varepsilon)$ so that the right-hand side of the expression in Theorem 1.4 becomes $\varepsilon\sigma/\mathrm{poly}(\log n, \log(1/\varepsilon))$.

The proof of Theorem 1.4 proceeds by proving a lemma (Lemma 3.1) which captures the behavior of random restrictions with t stars versus those with $t+1$ stars. We prove a robust version of Pisier's inequality [Pis86], an inequality which was first studied in the geometry of Banach spaces. When the projections $p_{\overline{S}}$ are far from uniform in total variation, the robust version of Pisier's inequality will help us lower bound certain quantities of a collection of directed graphs defined using $p_{\overline{S}}$ on the subcube $\{-1,1\}^{\overline{S}}$. The coordinate values of the mean vector of $p_{|\boldsymbol{\rho}}$ will depend on whether random vertices and directions induced from $\boldsymbol{\rho}$ are directed edges of one of the graphs in the collection. Hence, the lower bound on expected norm of the mean vector will follow from studying structures of directed graphs in the collection.

We first introduce the robust version of Pisier's inequality. As a warm-up and in order to show the usefulness of this inequality, we give an algorithm making $\widetilde{O}(n/\varepsilon^2)$ queries which we refer to as an *edge tester*. The reason for this name is that the algorithm samples a random restriction $\boldsymbol{\rho}$ with exactly one star, so that the distribution $p_{|\boldsymbol{\rho}}$ is supported on an edge of the hypercube. Then, the algorithm tests whether $p_{|\boldsymbol{\rho}}$ is uniform. Lastly, we provide a proof sketch for a weaker version of Theorem 1.4 which lower bounds the expectation of $\|\mu(p_{|\boldsymbol{\rho}})\|_2^2$, rather than $\|\mu(p_{|\boldsymbol{\rho}})\|_2$. The weaker inequality is insufficient for our purposes, but the proof is conceptually much simpler (since $\|\cdot\|_2^2$ is additive over coordinates and thus, we may use linearity of expectation). To prove Theorem 1.4, attempting to lower bound $\|\mu(p_{|\boldsymbol{\rho}})\|_2$ reveals further challenges which necessitate additional care.

1.2.1 A Robust Pisier's Inequality

We let $\mathcal{H}$ denote the set of undirected edges of the hypercube $\{-1,1\}^n$. For a function $f\colon \{-1,1\}^n \to \mathbb{R}$ and $i \in [n]$, we write $L_i f(x) \stackrel{\text{def}}{=} (f(x) - f(x^{(i)}))/2$. The following inequality is known as Pisier's inequality [Pis86]. We state it below in a way most closely to how it will be applied in this paper; the inequality holds in much larger generality for functions $f\colon \{-1,1\}^n \to X$ over general Banach spaces X (see, in particular, [NS02]).

Theorem 1.5. (Pisier's inequality) *Let $f\colon \{-1,1\}^n \to \mathbb{R}$ be a function with $\mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})] = 0$. Then,*

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\left[|f(\boldsymbol{x})|\right] \lesssim \log n \cdot \mathop{\mathbf{E}}_{\boldsymbol{x},\boldsymbol{y}\sim\{-1,1\}^n}\left[\left|\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{x}_i L_i f(\boldsymbol{x})\right|\right].$$

In this paper, we will need a *robust* version of Pisier's inequality in order to derive Theorem 1.4. The notion of robustness is equivalent to the notion considered in [KMS18], who proved robust (and directed) versions of Talagrand's inequality for Boolean functions, and part of our proof utilizes the robustness of the inequality in a similar way. Specifically, we consider an arbitrary orientation of the edges of the hypercube and sum the values of $L_i f(\boldsymbol{x})$ only when the edge $\{\boldsymbol{x}, \boldsymbol{x}^{(i)}\}$ is oriented from $\boldsymbol{x}$ to $\boldsymbol{x}^{(i)}$.

Theorem 1.6. *(Robust version of Pisier's inequality) Let $f\colon \{-1,1\}^n \to \mathbb{R}$ be a function satisfying $\mathbf{E}_{\boldsymbol{x}}[|f(\boldsymbol{x})|] = 0$. Let $G = (\{-1,1\}^n, E)$ be any orientation of the hypercube. Then,*

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\left[|f(\boldsymbol{x})|\right] \lesssim \log n \cdot \mathop{\mathbf{E}}_{\boldsymbol{x},\boldsymbol{y}\sim\{-1,1\}^n}\left[\left|\sum_{\substack{i\in[n]\\(\boldsymbol{x},\boldsymbol{x}^{(i)})\in E}} \boldsymbol{y}_i\boldsymbol{x}_i L_i f(\boldsymbol{x})\right|\right].$$

The proof follows the template of [NS02, Theorem 2], and checks that the necessary changes, even after considering directed edges, still give the desired inequality.

We note that Talagrand [Tal93] prove that the $\log n$ factor in Pisier's inequality (Theorem 1.5) is unnecessary for real-valued functions, but that it is for general Banach spaces. While one may follow Talagrand's proof to remove the $\log n$ factor in Theorem 1.6, it is not immediately clear whether this approach can handle arbitrary orientations.

1.2.2 Warmup: A Linear Query Algorithm

To see why Theorem 1.6 is helpful, we give a simple (albeit suboptimal) $\widetilde{O}(n/\varepsilon^2)$-query algorithm for testing uniformity. Suppose p is a distribution supported on $\{-1,1\}^n$ which is ε-far from uniform in total variation distance. We apply Theorem 1.6 to the function $f(x) \stackrel{\text{def}}{=} 2^n \cdot p(x) - 1$, where the directed graph $G = (\{-1,1\}^n, E)$ is given by letting

$$E = \Big\{(x, x^{(i)}) : \{x, x^{(i)}\} \text{ is an edge of } \{-1,1\}^n \text{, and } p(x) \geq p(x^{(i)})\Big\}.$$

Theorem 1.6 applied on f implies that[5]

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\left[|f(\boldsymbol{x})|\right] \lesssim \log n \cdot \mathop{\mathbf{E}}_{\boldsymbol{x},\boldsymbol{y}\sim\{-1,1\}^n}\left[\left|\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{x}_i\big(f(\boldsymbol{x}) - f(\boldsymbol{x}^{(i)})\big)^+\right|\right]. \tag{1.3}$$

[5]Our proof of Theorem 1.4 will rely on this strengthening of Pisier's inequality for multiple reasons, as our approach crucially uses the ability to pick any orientation of the edges. Even for this simple application one can observe that, without the strengthening, the right-hand side of (1.3) would be replaced by the same expression but without the $^+$. The discussion below (1.4) and the derivation of (1.5) explains why having the $^+$ in the expression is crucial.

Given (1.3), we start by applying the fact that the left-hand side of the inequality is at least 2ε. From there, the following three (in)equalities are obtained via Khintchine's inequality, importance sampling, and Jensen's inequality, respectively. (We also use the convention that $0/0 = 0$.)

$$
\begin{aligned}
\frac{\varepsilon}{\log n} &\lesssim \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^n}\left[\sqrt{\sum_{i=1}^{n}\left(\left(f(\boldsymbol{x})-f(\boldsymbol{x}^{(i)})\right)^+\right)^2}\right] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim p}\left[\sqrt{\sum_{i=1}^{n}\left(\frac{(f(\boldsymbol{x})-f(\boldsymbol{x}^{(i)}))^+}{1+f(\boldsymbol{x})}\right)^2}\right] \\
&\leq \left(\mathop{\mathbf{E}}_{\boldsymbol{x}\sim p}\left[\sum_{i=1}^{n}\left(\frac{(p(\boldsymbol{x})-p(\boldsymbol{x}^{(i)}))^+}{p(\boldsymbol{x})}\right)^2\right]\right)^{1/2}.
\end{aligned}
\tag{1.4}
$$

Notice that for every x and i, either $(p(x)-p(x^{(i)}))^+ = 0$ or $p(x) > p(x^{(i)}) \geq 0$. First, this makes sure that $a/0$ with $a > 0$ never occurs in (1.4). Additionally,

$$
0 \leq \frac{1}{2}\cdot\frac{(p(x)-p(x^{(i)}))^+}{p(x)} \leq \frac{|p(x)-p(x^{(i)})|}{p(x)+p(x^{(i)})}, \tag{1.5}
$$

and the quantity on the right-hand side is exactly the bias on the distribution of the ith bit of a draw of p conditioning on all $i' \neq i$ bits set according to x. In particular, a standard averaging/bucketing argument shows that there exists $\beta \geq \varepsilon^2/(n\log^2 n)$ such that

$$
\begin{aligned}
&\mathop{\mathbf{Pr}}_{\substack{\boldsymbol{x}\sim p\\ \boldsymbol{i}\sim[n]}}\left[\left(\frac{|p(\boldsymbol{x})-p(\boldsymbol{x^{(i)}})|}{p(\boldsymbol{x})+p(\boldsymbol{x^{(i)}})}\right)^2 \gtrsim \frac{\varepsilon^2}{\beta\cdot n\log^2 n\log(n/\varepsilon)}\right] \\
&\geq \beta.
\end{aligned}
\tag{1.6}
$$

The above lower bound suggests the following "edge tester:" for all $h \in \{0, \ldots, O(\log(n/\varepsilon))\}$ where $2^{-h} \gtrsim \varepsilon^2/(n\log^2 n)$, independently sample $O(2^h \log(n/\varepsilon))$ pairs $(\boldsymbol{x}, \boldsymbol{i})$ ($\boldsymbol{x}$ from p, and $\boldsymbol{i} \sim [n]$). For each pair $(\boldsymbol{x}, \boldsymbol{i})$, we consider the distribution supported on $\{-1, 1\}$ given by sampling $\boldsymbol{y} \sim p$ *conditioned* on every $i' \neq \boldsymbol{i}$ having $\boldsymbol{y}_{i'} = \boldsymbol{x}_{i'}$, and use $\widetilde{O}(2^{-h}n/\varepsilon^2)$ queries to estimate the bias of the conditional distribution up to error $(\varepsilon\cdot(2^h/n)^{1/2})/\mathrm{polylog}(n/\varepsilon)$ with high probability. If p was uniform, every conditional distribution considered will be uniform; however, if p is ε-far from uniform in total variation distance, (1.6) implies that some setting of h will reveal a large bias with high probability. The query complexity of this algorithm is

$$
\sum_{h=0}^{O(\log(n/\varepsilon))} O\left(2^h\cdot\log\left(\frac{n}{\varepsilon}\right)\right)\cdot\widetilde{O}\left(\frac{2^{-h}n}{\varepsilon^2}\right) = \widetilde{O}\left(\frac{n}{\varepsilon^2}\right).
$$

1.2.3 A Weaker Version of Theorem 1.4

In order to highlight some of the conceptual ideas involved in proving Theorem 1.4, we sketch how one may prove the following weaker inequality, which lower bounds the expected squared ℓ_2 norm instead of the expected ℓ_2 norm:

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{\rho}\sim\mathcal{D}(t+1,p)}\left[\|\mu(p_{|\boldsymbol{\rho}})\|_2^2\right] \\
&\gtrsim \frac{1}{\log^2 n}\cdot\frac{t+1}{n-t}\cdot\mathop{\mathbf{E}}_{\substack{\mathbf{T}\subset[n]\\|\mathbf{T}|=t}}\left[d_{\mathrm{TV}}(p_{\overline{\mathbf{T}}},\mathcal{U})^2\right]
\end{aligned}
\tag{1.7}
$$

where $\mathcal{D}(t+1, p)$ denotes the distribution over random restrictions where we enforce $\mathrm{stars}(\boldsymbol{\rho}) = t+1$ (to compare this to Theorem 1.4, one should think of σ as t/n). Specifically, we sample a random set $\mathbf{S} \subset [n]$ of size $t+1$ and $\boldsymbol{x} \sim p$, then, we set $\boldsymbol{\rho}$ as in (1.1). Consider imposing a random order $\boldsymbol{\pi}\colon [t+1] \to \mathbf{S}$ and apply linearity of expectation to re-write the left-hand side of (1.7) as

$$
\sum_{i=1}^{t+1}\mathop{\mathbf{E}}_{\boldsymbol{\rho},\boldsymbol{\pi}}\left[(\mu(p_{|\boldsymbol{\rho}})_{\boldsymbol{\pi}(i)})^2\right] = (t+1)\mathop{\mathbf{E}}_{\boldsymbol{\rho},\boldsymbol{\pi}}[\mu(p_{|\boldsymbol{\rho}})^2_{\boldsymbol{\pi}(t+1)}]. \tag{1.8}
$$

Toward lower bounding (1.8), consider the following way of sampling $\boldsymbol{\rho}$ and $\boldsymbol{\pi}$. We first sample t random indices $\boldsymbol{\pi}(1), \ldots, \boldsymbol{\pi}(t)$ uniformly from $[n]$ without replacement and let $\mathbf{T} = \{\boldsymbol{\pi}(1), \ldots \boldsymbol{\pi}(t)\}$. Then, we sample $\boldsymbol{x} \sim p$. Finally, we sample $\boldsymbol{\pi}(t+1)$ uniformly from the set $\overline{\mathbf{T}} = [n] \setminus \mathbf{T}$. We may write $\mathbf{S} = \mathbf{T} \cup \{\boldsymbol{\pi}(t+1)\}$ and similarly have $\boldsymbol{\rho}$ be set according to (1.1). Consider the function $\boldsymbol{\ell}\colon \{-1,1\}^{\overline{\mathbf{T}}} \to [-1, \infty)$ which is given by

$$
\boldsymbol{\ell}(y) = 2^{|\overline{\mathbf{T}}|}\mathop{\mathbf{Pr}}_{\boldsymbol{y}\sim p_{\overline{\mathbf{T}}}}[\boldsymbol{y} = y] - 1. \tag{1.9}
$$

We notice that $\boldsymbol{\ell}$ has mean zero (since $p_{\overline{\mathbf{T}}}$ is a probability distribution), similarly to Section 1.2.2, we orient the edges of the hypercube $\{-1,1\}^{\overline{\mathbf{T}}}$ from the endpoint with higher value of $\boldsymbol{\ell}$ to lower value of $\boldsymbol{\ell}$. Applying Theorem 1.6 to $f = \boldsymbol{\ell}$ and this orientation of the hypercube edges, the left-hand side is exactly $2d_{\mathrm{TV}}(p_{\overline{\mathbf{T}}}, \mathcal{U})$. Further, we write the random variable $\boldsymbol{z} = \boldsymbol{x}_{\overline{\mathbf{T}}}$, which is distributed exactly as $\boldsymbol{z} \sim p_{\overline{\mathbf{T}}}$, and the random variable $\boldsymbol{j} = \boldsymbol{\pi}(t+1)$, which is distributed uniformly from $\overline{\mathbf{T}}$. We have, similarly to (1.4) but with $f = \boldsymbol{\ell}$ rather than p,

$$
\begin{aligned}
&\frac{d_{\mathrm{TV}}(p_{\overline{\mathbf{T}}},\mathcal{U})}{\log n} \\
&\lesssim \left(|\overline{\mathbf{T}}|\cdot\mathop{\mathbf{E}}_{\substack{\boldsymbol{z}\sim p_{\overline{\mathbf{T}}}\\ \boldsymbol{j}\sim\overline{\mathbf{T}}}}\left[\left(\frac{(\boldsymbol{\ell}(\boldsymbol{z})-\boldsymbol{\ell}(\boldsymbol{z}^{(\boldsymbol{j})}))^+}{1+\boldsymbol{\ell}(\boldsymbol{z})}\right)^2\right]\right)^{1/2}.
\end{aligned}
\tag{1.10}
$$

One crucial observation is that the inner value of the expectation in the right-hand side of (1.10) is, up to a factor of at most 4, $\mu(p_{|\boldsymbol{\rho}})^2_{\boldsymbol{\pi}(t+1)}$, where $\boldsymbol{\rho}$ uses $\mathbf{T} \cup \{\boldsymbol{j}\}$ as stars and non-star values are set to $\boldsymbol{z}_{-\boldsymbol{j}}$. Plugging this back into (1.8) gives (1.7).

1.3 Related Work The seminal works of Goldreich, Goldwasser, and Ron [GGR96], Goldreich and Ron [GR00], and Batu, Fortnow, Rubinfeld, Smith, and White [BFR$^+$00] initiated the study of distribution testing, viewing probability distributions as a natural application for property testing (see [Gol17] for coverage of this much broader field). Since these works, distribution testing has enjoyed a wealth of study, resulting in a thorough understanding of the complexity of testing many distribution properties (see, e.g., [BFF$^+$01, BKR04, Pan08, ADJ$^+$11, BFRV11, Val11, ILR12, DDS$^+$13, CDVV14, VV17, Wag15, ADK15, BV15, DKN15, DK16, Can16, BCG17, BC17, DKW18, DGPP18], and [Rub12, Can15b, BW18, Kam18] for recent surveys). As a result, sample-optimal algorithms are known for a number of core problems.

However, the known sample complexity lower bounds, while sublinear, typically still involve a polynomial dependence on the domain size, suggesting that new models are needed for studying distribution testing on high-dimensional domains. One approach is to assume additional *structure* from the input distributions. Settings were studied where the distribution is known to be monotone [RS09], a low-degree Bayesian Network [CDKS17, DP17, ABDK18], a Markov Random Field [DDK18, GLP18, BBC$^+$19], or having some "flat" histogram structure [DKP19].

The other approach is to allow stronger oracle access. The subcube conditional sampling model, which is the focus of this work, is a variant of the general conditional sampling model particularly apt for the study of high-dimensional distributions. Bhattacharyya and Chakraborty [BC18], who initiated the systematic study of this variant, showed that for many problems of interest such as uniformity, identity, and closeness testing, subcube conditional queries enabled one to avoid the curse of dimensionality, and established sample complexity upper bounds *polynomial* in the dimension (albeit superlinear). The conditional sampling model itself, which was introduced simultaneously by Chakraborty, Fischer, Goldhirsh, and Matsliah [CFGM13, CFGM16], and Canonne, Ron, and Servedio [CRS14, CRS15], allows more general queries: namely, the algorithm may specify an arbitrary subset of the domain and request a sample conditioned on it lying in the subset. In many cases, the conditional sampling model circumvents sample-complexity lower bounds. Since its introduction, there has been significant study into the complexity of testing a number of properties of distributions under conditional samples, in both adaptive and nonadaptive settings [Can15a, FJO$^+$15, ACK15b, FLV17, SSJ17, BCG17, BC18, KT19]. Beyond distribution testing, this model of conditional sampling has found applications in group testing [ACK15a], sublinear algorithms [GTZ17], and crowdsourcing [GTZ18]. Other ways to augment the power of distribution testing algorithms include letting the algorithm query the probability density function (PDF) or cumulative distribution function (CDF) of the distribution [BDKR05, GMV06, RS09, CR14], or giving it probability-revealing samples [OS18].

1.4 Notation and Prelimaries We use boldface symbols to represent random variables, and non-boldface symbols for fixed values (potentially realizations of these random variables) — see, e.g., $\boldsymbol{\rho}$ versus ρ. Given a set $S \subseteq [n]$, we let $\mathcal{U}_S$ denote the uniform distribution over $\{-1,1\}^S$. Usually, as the support of $\mathcal{U}_S$ will be clear from the context, we will drop the subscript and simply write $\mathcal{U}$. We write $f(n) \lesssim g(n)$ if, for some $c > 0$, $f(n) \leq c \cdot g(n)$ for all $n \geq 0$ (the $\gtrsim$ symbol is defined similarly). $f(n) \asymp g(n)$ if $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$. We use the notation $\tilde{O}(f(n))$ as $O(f(n)\mathrm{polylog}(f(n)))$, and $\tilde{\Omega}(f(n))$ to denote $\Omega(f(n)/(1 + |\mathrm{polylog}(f(n))|))$. The notation $[k]$ denotes the set of integers $\{1, \ldots, k\}$.

The Frobenius norm of a matrix $M \in \mathbb{R}^{d_1 \times d_2}$ is

$$\|M\|_F = \left(\sum_{i \in [d_1]} \sum_{j \in [d_2]} M_{ij}^2 \right)^{1/2}.$$

For a string $x \in \{-1,1\}^n$, we use $x^{(i)}$ to denote the string that is identical to x but with coordinate i flipped, i.e., $x_j^{(i)} = x_j$ for all $j \neq i$, and $x_i^{(i)} = -x_i$.

We formally define the subcube conditional query access, which was suggested in Canonne, Ron, Servedio [CRS15] as an instance of the general conditional sampling oracle [CRS15, CFGM16], and first explicitly studied in Bhattacharyya and Chakraborty [BC18].

Definition 1.2. *A subcube conditional sampling (*SCOND*) oracle for a distribution p supported on $\{-1,1\}^n$ is an oracle which accepts a query subcube $\rho \in \{-1,1,*\}^n$, and outputs a sample from the distribution $\boldsymbol{x} \sim p$ conditioned on every $i \notin \mathrm{stars}(\rho)$ having $\boldsymbol{x}_i = \rho_i$. We use the convention that if the algorithm considers a restriction with zero support, the oracle outputs a uniform sample.*

2 The Algorithm

We prove our theorem for uniformity testing with subcube conditioning, restated below:

Theorem 2.1. *There exists an algorithm* SubCondUni *which, given $n \geq 1$, a subcube ora-*

cle to a distribution p over $\{-1,1\}^n$ and a distance parameter ε with $\varepsilon \in (0,1)$, has the following guarantees. The algorithm makes $\widetilde{O}(\sqrt{n}/\varepsilon^2)$ many calls to the oracle and satisfies the following two conditions:

(i) If p is uniform, then the algorithm returns accept *with probability at least $2/3$.*

(ii) If $d_{\mathrm{TV}}(p,\mathcal{U}) \geq \varepsilon$, then the algorithm returns reject *with probability at least $2/3$.*

The rest of this section is devoted to the proof of Theorem 2.1. For our convenience of working with $\log(1/\varepsilon)$ in the proof, we assume below that $\varepsilon \leq 1/2$ in the input of SubCondUni. This way $\log(c/\varepsilon)$ can be treated as $O(\log(1/\varepsilon))$ whenever $c \geq 1$ is a fixed constant. As discussed earlier in Section 1, SubCondUni (see Algorithm 1) is based on Theorem 1.4 and Lemma 1.1, which we now prove.

Let p be a distribution over $\{-1,1\}^n$ with $d_{\mathrm{TV}}(p,\mathcal{U}) \geq \varepsilon$. Let

$$\sigma \stackrel{\text{def}}{=} \sigma(\varepsilon) = \frac{1}{C_0 \cdot \log^4(16/\varepsilon)}$$

where $C_0 > 0$ is an absolute constant. (The value of C_0 is only used at the end of this section, where setting $C_0 = 10^{11}$ is good enough.) Let us further assume that n and ε together satisfy

$$e^{-\sigma n/10} \leq \varepsilon/8. \tag{2.11}$$

Violation of (2.11) implies that

$$n = O\left(\frac{1}{\sigma} \cdot \log\left(\frac{1}{\varepsilon}\right)\right) = O\left(\log^5\left(\frac{1}{\varepsilon}\right)\right) \tag{2.12}$$

and we handle this case by applying the linear-query tester described in Section 1.2.2 with query complexity $\widetilde{O}(n/\varepsilon^2) = \widetilde{O}(1/\varepsilon^2)$ using (2.12).

For the general case with (2.11) satisfied, consider a distribution p with $d_{\mathrm{TV}}(p,\mathcal{U}) \geq \varepsilon$. Lemma 1.1 implies that either $\mathbf{E}_{\mathbf{S}\sim\mathcal{S}_\sigma}\left[d_{\mathrm{TV}}(p_{\overline{\mathbf{S}}},\mathcal{U})\right] \geq \varepsilon/2$ or

$$\mathop{\mathbf{E}}_{\boldsymbol{\rho}\sim\mathcal{D}_\sigma(p)}\left[d_{\mathrm{TV}}(p_{|\boldsymbol{\rho}},\mathcal{U})\right] \geq \varepsilon/2. \tag{2.13}$$

Assuming the former and using (2.11), we have from Theorem 1.4 that (using $\sigma = 1/\mathrm{polylog}(1/\varepsilon)$)

$$\mathop{\mathbf{E}}_{\boldsymbol{\rho}\sim\mathcal{D}_\sigma(p)}\left[\|\mu(p_{|\boldsymbol{\rho}})\|_2/\sqrt{n}\right] \geq \widetilde{\Omega}\left(\frac{\varepsilon}{\sqrt{n}}\right). \tag{2.14}$$

This naturally lends itself to a recursive approach: for the general case when n and ε satisfy (2.11), we use TestMean from Theorem 1.3 to test (2.14), and recursive calls to SubCondUni to test (2.13).

Algorithm 1 SubCondUni(n,p,ε)

Require: Dimension n, oracle access to distribution p over $\{-1,1\}^n$, and parameter $\varepsilon \in (0,1/2]$

StartBaseCase ▷ Base case: violation of (2.11)
  if n and ε violate (2.11) **then**
    Run the linear tester described in Section 1.2.2 and **return** the same answer
EndBaseCase
StartMainCase ▷ General case: (2.11) satisfied
  Let $L = L(n,\varepsilon) = \widetilde{O}(\sqrt{n}/\varepsilon)$ be as defined in (2.15) to simplify (2.14)
  for $j = 1,2,\ldots,\lceil \log 2L \rceil$ **do** ▷ Test (2.14), via bucketing
    Sample $s_j = 8L\log(2L)\cdot 2^{-j}$ restrictions from $\mathcal{D}_\sigma(p)$
    for every restriction $\boldsymbol{\rho}$ sampled with $|\mathrm{stars}(\boldsymbol{\rho})| > 0$ **do**
      Run TestMean$(|\mathrm{stars}(\boldsymbol{\rho})|, p_{|\boldsymbol{\rho}}, 2^{-j})$ for $r = O(\log(n/\varepsilon))$ times
      return reject if the majority of calls return reject
  for $j = 1,2,\ldots,\lceil \log(4/\varepsilon) \rceil$ **do** ▷ Test the second part of (2.13), recursively
    Sample $s'_j = (32/\varepsilon)\log(4/\varepsilon)\cdot 2^{-j}$ restrictions from $\mathcal{D}_\sigma(p)$
    for each restriction $\boldsymbol{\rho}$ sampled satisfying $0 < |\mathrm{stars}(\boldsymbol{\rho})| \leq 2\sigma n$ **do**
      Run SubCondUni$(|\mathrm{stars}(\boldsymbol{\rho})|, p_{|\boldsymbol{\rho}}, 2^{-j})$ for $t = 100\log(16/\varepsilon)$ times
      return reject if the majority of calls return reject
EndMainCase
return accept

The description of SubCondUni is given in Algorithm 1. For convenience, we let

$$L \stackrel{\text{def}}{=} L(n,\varepsilon) = \widetilde{O}(\sqrt{n}/\varepsilon) \tag{2.15}$$

such that the right hand side of (2.14) can be replaced by $1/L$.

We are now ready to prove Theorem 2.1.

3 Proof of Theorem 1.4

Let $\mathcal{S}(t)$ be the uniform distribution supported on all subsets of $[n]$ of size t. Given a distribution p supported on $\{-1,1\}^n$, we let $\mathcal{D}(t,p)$ be the distribution supported on restrictions $\{-1,1,*\}^n$ given in a similar fashion to that of $\mathcal{D}_\sigma(p)$, except we use sets of size t; we sample $\mathbf{S} \sim \mathcal{S}(t)$ and $\boldsymbol{x} \sim p$, and we let $\boldsymbol{\rho}_i = *$ if $i \in \mathbf{S}$ and $\boldsymbol{x}_i$ otherwise. The bulk of the work goes into proving the following lemma. After the statement, we show the lemma implies Theorem 1.4.

LEMMA 3.1. *Let p be a distribution supported on $\{-1,1\}^n$, $t \in [n-1]$, and denote*

$$\alpha \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\mathbf{T} \sim \mathcal{S}(t)} \left[d_{\mathrm{TV}}(p_{\overline{\mathbf{T}}}, \mathcal{U}) \right] \geq 0.$$

Then,

$$\text{(3.16)} \qquad \mathop{\mathbf{E}}_{\boldsymbol{\rho} \sim \mathcal{D}(t,p)} \left[\|\mu(p_{|\boldsymbol{\rho}})\|_2 \right] + \mathop{\mathbf{E}}_{\boldsymbol{\rho} \sim \mathcal{D}(t+1,p)} \left[\|\mu(p_{|\boldsymbol{\rho}})\|_2 \right]$$

$$\text{(3.17)} \qquad \gtrsim \frac{t}{n} \cdot \frac{\alpha}{\log^2 n \cdot \log(n/\alpha) \cdot \log(1/\alpha)}.$$

3.1 Robust Pisier Inequality In this section, we prove the robust version of Pisier's inequality. Robustness here is equivalent to that of [KMS18] where we will consider a function $f\colon \{-1,1\}^n \to \mathbb{R}$, and we will lower bound a functional on the values of the edges of the hypercube after assigning them directions.

Formally, fix $n \in \mathbb{N}$ and let $\mathcal{H}$ be the undirected graph over the hypercube $\{-1,1\}^n$ that consists of undirected edges $\{x, x^{(i)}\}$ with $x \in \{-1,1\}^n$ and $i \in [n]$. For $i \in [n]$, recall that $L_i f\colon \{-1,1\}^n \to \mathbb{R}$ is the linear operator given by

$$L_i f(x) = \frac{f(x) - f(x^{(i)})}{2}.$$

Notice that in Theorem 1.5, every edge $\{x, x^{(i)}\}$ in $\mathcal{H}$ for a fixed value of $\boldsymbol{y}$ is counted twice in the right-hand side; once for the endpoint x and once for the endpoint $x^{(i)}$. In the robust version, we may arbitrarily choose, for each edge $\{x, x^{(i)}\}$, whether to "charge" the edge to x or to $x^{(i)}$. For this purpose, we consider an orientation G of $\mathcal{H}$ (so for each $\{x, x^{(i)}\}$ in $\mathcal{H}$, G contains either $(x, x^{(i)})$ or $(x^{(i)}, x)$), and charge an edge $\{x, x^{(i)}\}$ to x if $(x, x^{(i)})$ is in G and to $x^{(i)}$ otherwise. For convenience, we will abuse the notation in the rest of the section to use the name of a directed graph (such as G) to denote its edge set as well, since its vertex set is usually clear from the context. So we will write $(u,v) \in G$ if (u,v) is a directed edge in G.

We are now ready to state the robust version of Pisier's inequality, which generalizes Theorem 1.6.

THEOREM 3.1. (ROBUST PISIER'S INEQUALITY) *Let $f\colon \{-1,1\}^n \to \mathbb{R}$ be a function with*

$$\text{(3.18)} \qquad \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n} \left[f(\boldsymbol{x}) \right] = 0$$

and let G be an orientation of $\mathcal{H}$. Then, for any $s \in [1,\infty)$ we have

$$\left(\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n} \left[\left| f(\boldsymbol{x}) \right|^s \right] \right)^{1/s} \lesssim \log n \cdot \left(\mathop{\mathbf{E}}_{\boldsymbol{x},\boldsymbol{y} \sim \{-1,1\}^n} \left[\left| \sum_{\substack{i \in [n] \\ (\boldsymbol{x}, \boldsymbol{x}^{(i)}) \in G}} \boldsymbol{y}_i \boldsymbol{x}_i L_i f(\boldsymbol{x}) \right|^s \right] \right)^{1/s}.$$

The proof itself follows the template of [NS02, Theorem 2] and checks that the necessary changes still give the desired inequality. Before beginning with the proof, we recall some basic notions of Fourier analysis on the hypercube, and define some elements which appear in the argument. Recall that any function $f\colon \{-1,1\}^n \to \mathbb{R}$ has a unique Fourier expansion $f(x) = \sum_{S \subset [n]} \widehat{f}(S) \chi_S(x)$, where $\chi_S(x) = \prod_{i \in S} x_i$ are the Fourier characters, and $\widehat{f}(S) = \mathbf{E}_{\boldsymbol{x} \sim \{-1,1\}^n}[f(\boldsymbol{x}) \chi_S(x)]$. For $\rho > 0$ and $x \in \{-1,1\}^n$, we let $N_\sigma(x)$ be the distribution supported on $\{-1,1\}^n$ given sampling $\boldsymbol{y} \sim N_\sigma(x)$, where for each $i \in [n]$, we set $\boldsymbol{y}_i = \boldsymbol{x}_i$ with probability $1-\sigma$, and a uniform random bit with probability σ. We denote $T_\rho f(x) = \mathbf{E}_{\boldsymbol{y} \sim N_\rho(x)}[f(\boldsymbol{y})]$, and we have for every $S \subset [n]$, $\widehat{T_\rho f}(S) = \rho^{|S|} \widehat{f}(S)$. In a slight abuse of notation, we consider for any $x, y \in \{-1,1\}^n$ and $t \in [0,1]$, the distribution $N_{t,1-t}(x,y)$, supported on $\{-1,1\}^n$, to be the distribution given by letting $\boldsymbol{z} \sim N_{t,1-t}(x,y)$ have each $i \in [n]$ set to $\boldsymbol{z}_i = x_i$ with probability t and $\boldsymbol{z}_i = y_i$ otherwise. For a function $g\colon \{-1,1\}^n \to \mathbb{R}$ and $t \in [0,1]$, the function $g\colon \{-1,1\}^n \times \{-1,1\}^n \to \mathbb{R}$ is given by letting $g_{t,1-t}(x,y) = \mathbf{E}_{\boldsymbol{z} \sim N_{t,1-t}(x,y)}[g(\boldsymbol{z})] = \sum_{S \subset [n]} \widehat{g}(S) \prod_{i \in S}(tx_i + (1-t)y_i)$. Lastly, for any $\gamma > 0$, we let $\Delta^\gamma f$ be the linear operator given by $\Delta^\gamma f(x) = \sum_{S \subset [n]} \widehat{f}(S) |S|^\gamma \chi_S(x)$.

3.2 Plan of the Proof of Lemma 3.1 Let $t \in [n-1]$ be the parameter in the statement of Lemma 3.1. For clarity, the rest of this section will always use T to denote a size-t subset of $[n]$ and S to denote a size-$(t+1)$ subset of $[n]$. For each size-t subset T of $[n]$, we write $\alpha(T) \stackrel{\text{def}}{=} d_{\mathrm{TV}}(p_{\overline{T}}, \mathcal{U})$. Then $\alpha = \mathbf{E}_{\mathbf{T} \sim \mathcal{S}(t)}[\alpha(\mathbf{T})]$. We also write $\mathcal{H}(T)$ to denote the undirected graph over the hypercube $\{-1,1\}^{\overline{T}}$ that consists of undirected edges $\{x, x^{(i)}\}$ with $x \in \{-1,1\}^{\overline{T}}$ and $i \in \overline{T}$. Again we will abuse the notation in the rest of the section to refer to $\mathcal{H}(T)$ as its edge set as well.

The proof of Lemma 3.1 consists of two steps. For each t-subset T, we first classify undirected edges $\{x, x^{(i)}\}$ in $\mathcal{H}(T)$ into different types according its *weight* defined as

$$w\big(\{x, x^{(i)}\}\big) \stackrel{\text{def}}{=} \frac{\big|p_{\overline{T}}(x) - p_{\overline{T}}(x^{(i)})\big|}{\max\big\{p_{\overline{T}}(x), p_{\overline{T}}(x^{(i)})\big\}}.$$

For each type of edges in $\mathcal{H}(T)$, we describe a method in Section 3.3 to assign each edge a direction. This then leads to a sequence of directed graphs over $\{-1,1\}^{\overline{T}}$, one for each type of edges in $\mathcal{H}(T)$, and their union is an orientation G of $\mathcal{H}(T)$ over $\{-1,1\}^{\overline{T}}$. At the end of Section 3.3 we apply the robust Pisier inequality (Theorem 3.1) on a shifted and scaled version of the probability mass function of $p_{\overline{T}}$ with the orientation G and $s = 1$. The result will be Lemma 3.5, which says there is a type of edges in $\mathcal{H}(T)$ such that its corresponding directed graph has

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim p_{\overline{T}}}\left[\sqrt{\text{out-degree of } \boldsymbol{x}}\right] \tag{3.19}$$

bounded from below by a quantity that is linear in $\alpha(T)$; see Lemma 3.5 for details.

In the second step, we use this family of directed graphs promised by Lemma 3.5, one for each t-subset T, with the desired bound on (3.19) to finish the proof of Lemma 3.1. To this end we first apply standard bucketing arguments in Section 3.4 to simplify the situation, by focusing on one specific type of directed edges that makes the most significant contribution in the family. The final connection from these directed graphs to the mean vectors of randomly restricted distributions is made in Sections 3.5.There will be two cases, depending on whether the type of edges we consider has large or small weights.

3.3 From Total Variation to Directed Graphs

Let ℓ be a probability distribution over $\{-1,1\}^m$ with $m = n - t$. (Later we will identify ℓ as $p_{\overline{T}}$ for some t-subset T of $[n]$, and $\{-1,1\}^m$ as $\{-1,1\}^{\overline{T}}$.) Let $\mathcal{H}$ denote the undirected graph over $\{-1,1\}^m$ that consists of undirected edges $\{x, x^{(i)}\}$ for all $x \in \{-1,1\}^m$ and $i \in [m]$. Looking ahead, the purpose of this section is to construct an orientation G of $\mathcal{H}$ for our application of Theorem 3.1 later with $s = 1$ and the function $f\colon \{-1,1\}^m \to [-1,\infty)$ given by:

$$f(y) = 2^m \cdot \ell(y) - 1. \tag{3.20}$$

Note that $\mathbf{E}_{\boldsymbol{y}}[f(\boldsymbol{y})] = 0$ and the left hand side of the robust Pisier inequality is $2d_{\mathrm{TV}}(\ell, \mathcal{U})$.

For the construction of G, we start with a classification of undirected edges in $\mathcal{H}$.

Definition 3.2. *An undirected edge $\{x, x^{(i)}\} \in \mathcal{H}$ is said to be a* zero edge *if $\ell(x) = \ell(x^{(i)})$ (they are called zero edges because the difference $\ell(x) - \ell(x^{(i)}) = 0$).*

For each nonzero edge $\{x, x^{(i)}\} \in \mathcal{H}$, we define its weight *as*

$$w(\{x, x^{(i)}\}) \stackrel{\text{def}}{=} \frac{\big|\ell(x) - \ell(x^{(i)})\big|}{\max\big\{\ell(x), \ell(x^{(i)})\big\}}.$$

Note that the weight of an undirected edge is always in $(0,1]$. *We say a nonzero edge is* uneven *if its weight is at least $2/3$; otherwise, we call it an* even edge *(i.e., any nonzero edge with weight smaller than $2/3$). We say an even edge is at* scale κ *for some $\kappa \geq 1$ if*

$$2^{-\kappa} < w\big(\{x, x^{(i)}\}\big) \leq 2^{-\kappa+1}.$$

We write $\mathcal{H}^{[z]}$ to denote the set of all zero edges, $\mathcal{H}^{[u]}$ to denote the set of all uneven edges, and $\mathcal{H}^{[\kappa]}$ for each $\kappa \geq 1$ to denote the set of even edges at scale κ. Hence $\mathcal{H}^{[z]}$, $\mathcal{H}^{[u]}$ and $\mathcal{H}^{[\kappa]}$ with $\kappa \geq 1$ together form a partition of $\mathcal{H}$; we also view them as undirected graphs over $\{-1,1\}^m$.

Next we construct a sequence of directed graphs $G^{[z]}$, $G^{[u]}$ and $G^{[\kappa]}$, $\kappa \geq 1$, as orientations of $\mathcal{H}^{[z]}$, $\mathcal{H}^{[u]}$ and $\mathcal{H}^{[\kappa]}$, respectively. We start with $G^{[z]}$ and $G^{[u]}$. For each zero edge $\{x, x^{(i)}\} \in \mathcal{H}^{[z]}$, we orient it arbitrarily in $G^{[z]}$. Next for each uneven edge $\{x, x^{(i)}\} \in \mathcal{H}^{[u]}$, we orient it from x to $x^{(i)}$ if $\ell(x) > \ell(x^{(i)})$ (note that if $\ell(x) = \ell(x^{(i)})$ then it is a zero edge).

Orientations of even edges at scale κ in $G^{[\kappa]}$ are more involved. For a fixed $\kappa \geq 1$, we consider $\mathcal{H}^{[\kappa]}$ as an undirected graph over $\{-1,1\}^m$. We will consider a bijection $\varrho_\kappa\colon \{-1,1\}^m \to [2^m]$ as an ordering of vertices in $\{-1,1\}^m$ (so x is the $\varrho_\kappa(x)$-th vertex in the ordering) such that the following property holds: For every $i \in [2^m - 1]$, the degree of $\varrho_\kappa^{-1}(i)$ has the largest degree among all vertices in the subgraph of $\mathcal{H}^{[\kappa]}$ induced by $\{\varrho_\kappa^{-1}(j) : j \geq i\}$. Such a bijection exists, e.g., by keeping deleting vertices one by one and each time deleting the one with the largest degree in the remaining graph (with tie breaking done arbitrarily). We fix such a bijection ϱ_κ and use it to orient edges in $G^{[\kappa]}$ as follows: For each undirected $\{x, x^{(i)}\} \in \mathcal{H}^{[\kappa]}$, we orient it from x to $x^{(i)}$ if $\varrho_\kappa(x) < \varrho_\kappa(x^{(i)})$ and orient it from $x^{(i)}$ to x otherwise. As a result, every $(x, x^{(i)}) \in G^{[\kappa]}$ has $\varrho_\kappa(x) < \varrho_\kappa(x^{(i)})$.

The above orientation will effectively streamline an argument from [KMS18, Section 6]. We record the property needed later for the orientation $G^{[\kappa]}$ of $\mathcal{H}^{[\kappa]}$:

Lemma 3.3. *Let U be a set of vertices in $\{-1,1\}^m$ and let $v \in \{-1,1\}^m \setminus U$. If the out-degree of every vertex $u \in U$ in $G^{[\kappa]}$ is bounded from above by a positive integer g, then the number of directed edges (u, v) from a vertex $u \in U$ to v in $G^{[\kappa]}$ is also at most g.*

Proof: Consider the vertex with the smallest $\varrho_\kappa(\cdot)$ value in $U \cup \{v\}$. If it is v, then every undirected $\{u, v\}$ with $u \in U$, if any, in $\mathcal{H}^{[\kappa]}$ is oriented as (v, u) in $G^{[\kappa]}$. So the number we care about is 0.

Otherwise, let $u \in U$ be the vertex with the smallest $\varrho_\kappa(\cdot)$ value among $U \cup \{v\}$. Then at the time when u is picked, all vertices $U \cup \{v\}$ remain in current undirected subgraph of $\mathcal{H}^{[\kappa]}$, denoted by H. At this moment, the degree of u in H is exactly its out-degree in $G^{[\kappa]}$, which by assumption is at most g. On the other hand, by the choice of u, v has degree at most g in H. Since the whole set U remains in H, the number of undirected edges $\{u, v\}$, $u \in U$, in $\mathcal{H}^{[\kappa]}$ is at most g. Even if all of them are oriented towards v in $G^{[\kappa]}$, the number we care about in the lemma is at most g. ∎

With $G^{[z]}$, $G^{[u]}$ and $G^{[\kappa]}$ ready, we finally define G to be the union of these graphs, which is an orientation of $\mathcal{H}$ over $\{0, 1\}^m$. Applying the robust Pisier inequality on f, G and $s = 1$, we have

$$\frac{d_{\mathrm{TV}}(\ell,\mathcal{U})}{\log n} \tag{3.21}$$

$$\lesssim \mathop{\mathbf{E}}_{\boldsymbol{x},\boldsymbol{y}\sim\{-1,1\}^m}\left[\left|\sum_{\substack{i\in[m]\\(\boldsymbol{x},\boldsymbol{x}^{(i)})\in G}} \boldsymbol{y}_i\boldsymbol{x}_i L_i f(\boldsymbol{x})\right|\right] \tag{3.22}$$

$$\leq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^m}\left[\sqrt{\sum_{\substack{i\in[m]\\(\boldsymbol{x},\boldsymbol{x}^{(i)})\in G}} \big(L_i f(\boldsymbol{x})\big)^2}\right]. \tag{3.23}$$

The second inequality is Khintchine's, which implies that for any vector $a \in \mathbb{R}^m$,

$$\mathop{\mathbf{E}}_{\boldsymbol{y}\sim\{-1,1\}^m}\left[\left|\sum_{i\in[m]} \boldsymbol{y}_i a_i\right|\right] \leq \sqrt{\sum_{i\in[m]} a_i^2}.$$

Letting G' be the directed graph that contains the union of edges in $G^{[u]}$ and $G^{[\kappa]}$, $\kappa \in \mathbb{Z}_{\geq 0}$, but not those in $G^{[z]}$, we can continue the inequality above to have

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\{-1,1\}^m}\left[\sqrt{\sum_{\substack{i\in[m]\\(\boldsymbol{x},\boldsymbol{x}^{(i)})\in G}} \big(L_i f(\boldsymbol{x})\big)^2}\right]$$

$$= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\ell}\left[\sqrt{\sum_{\substack{i\in[m]\\(\boldsymbol{x},\boldsymbol{x}^{(i)})\in G'}} \left(\frac{L_i f(\boldsymbol{x})}{1+f(\boldsymbol{x})}\right)^2}\right]$$

$$= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\ell}\left[\sqrt{\sum_{\substack{i\in[m]\\(\boldsymbol{x},\boldsymbol{x}^{(i)})\in G'}} \left(\frac{L_i \ell(\boldsymbol{x})}{\ell(\boldsymbol{x})}\right)^2}\right].$$

For the first equation we note that zero edges do not contribute anything and utilize importance sampling, by noting that $\ell(x) = (1+f(x))/2^m$. Also note we never run into a situation of $0/0$ in the second expectation because if $(x, x^{(i)}) \in G'$ has $\ell(x) = 0$, then either $\ell(x^{(i)}) = 0$ and it is a zero edge that should have been excluded from G', or $\ell(x^{(i)}) > 0$ and $\{x, x^{(i)}\}$ is uneven. Then by the construction of $G^{[u]}$ we have $(x^{(i)}, x) \in G$ instead of $(x, x^{(i)})$.

The next lemma connects the sum for each $x \in \{-1, 1\}^m$ in the last expectation with its out-degrees in the directed graphs $G^{[u]}$ and $G^{[\kappa]}$ constructed.

Lemma 3.4. *For every $x \in \{-1, 1\}^m$,*

$$\sum_{\substack{i\in[m]\\(x,x^{(i)})\in G'}} \left(\frac{\ell(x)-\ell(x^{(i)})}{\ell(x)}\right)^2 \leq \mathrm{outdeg}\big(x, G^{[u]}\big) + \sum_{\kappa\geq 1} 2^{-2\kappa+6}\cdot \mathrm{outdeg}\big(x, G^{[\kappa]}\big).$$

Proof: First note that each edge $(x, x^{(i)}) \in G'$ is nonzero and either lies in $G^{[u]}$ or $G^{[\kappa]}$ for some $\kappa \geq 1$. If $(x, x^{(i)})$ is in $G^{[u]}$, then by the way we orient edges in $G^{[u]}$, we have $\ell(x) > \ell(x^{(i)})$ and this implies that the contribution of each such edge to the sum is at most 1.

Next assume that $(x, x^{(i)}) \in G^{[\kappa]}$ for some $\kappa \geq 1$. Since $\{x, x^{(i)}\}$ is even, we have $\ell(x), \ell(x^{(i)}) > 0$ (otherwise it is either zero or uneven). Using $w(\{x, x^{(i)}\}) < 2/3$ (otherwise it is uneven), we have

$$\frac{\max\{\ell(x),\ell(x^{(i)})\}}{\min\{\ell(x),\ell(x^{(i)})\}} \leq 3.$$

As a result, we have

$$\frac{|\ell(x)-\ell(x^{(i)})|}{\ell(x)} \leq w(\{x,x^{(i)}\})\cdot\frac{\max\{\ell(x),\ell(x^{(i)})\}}{\min\{\ell(x),\ell(x^{(i)})\}}$$

$$\leq 3w(\{x,x^{(i)}\}) \leq 2^{-\kappa+3}.$$

So the contribution of each such edge is at most $2^{-2\kappa+6}$ and we thus obtain the desired bound. ∎

We are now ready to prove the main lemma of this subsection:

Lemma 3.5. *Letting $\beta = d_{\mathrm{TV}}(\ell,\mathcal{U})$, one of the following two conditions must hold:*

- *Either the directed graph $G^{[u]}$ of uneven edges satisfies:*

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\ell}\left[\sqrt{\mathrm{outdeg}(\boldsymbol{x}, G^{[u]})}\right] \gtrsim \frac{\beta}{\log n}$$

- *Or, there exists a $\kappa \in [O(\log(n/\beta))]$ such that the directed graph $G^{(\kappa)}$ satisfies:*

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\ell}\left[\sqrt{\mathrm{outdeg}(\boldsymbol{x}, G^{(\kappa)})}\right] \gtrsim \frac{2^\kappa \cdot \beta}{\log n \cdot \log(n/\beta)}.$$

3.4 Bucketing We now start the proof of Lemma 3.1. For each t-subset T of $[n]$, let $\alpha(T) = d_{\mathrm{TV}}(p_{\overline{T}}, \mathcal{U})$ and thus, $\alpha = \mathbf{E}_{\mathbf{T}\sim\mathcal{S}(t)}[\alpha(T)]$. For each T, we partition undirected edges in $\mathcal{H}(T)$ into $\mathcal{H}^{[z]}(T)$ (zero edges), $\mathcal{H}^{[u]}(T)$ (uneven edges), and $\mathcal{H}^{[\kappa]}(T)$ (even edges at scale $\kappa \geq 1$). We orient these edges to obtain directed graphs $G^{[u]}(T)$ and $G^{[\kappa]}(T)$. We apply Lemma 3.5 on $p_{\overline{T}}$ to conclude that one of the two conditions holds for either $G^{[u]}(T)$ or one of the graphs $G^{[\kappa]}(T)$, $\kappa \in [O(\log(n/\alpha(T)))]$.

Since $\alpha(T) \in [0,1]$, there exists a $\zeta > 0$ such that with probability at least ζ over $\mathbf{T} \sim \mathcal{S}(t)$,

$$\alpha(\mathbf{T}) \gtrsim \frac{\alpha}{\zeta \log(1/\alpha)}.$$

Therefore, via another bucketing argument and Lemma 3.5, there exist two cases:

- **Case 1**: With probability at least $\zeta/2$ over the draw of $\mathbf{T} \sim \mathcal{S}(t)$, we have that the directed graph $G^{[u]}(\mathbf{T})$ of uneven edges of $p_{\overline{\mathbf{T}}}$ over $\{-1,1\}^{\overline{\mathbf{T}}}$ satisfies

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim p_{\overline{\mathbf{T}}}}\left[\sqrt{\mathrm{outdeg}\big(\boldsymbol{x}, G^{[u]}(\mathbf{T})\big)}\right] \gtrsim \frac{\alpha}{\zeta \log n \log(1/\alpha)}.$$

  Since the out-degree is always between 0 and n, there exist two parameters $d \in [n]$ and $\xi > 0$ such that with probability $\zeta/(2\log n)$ over the draw of $\mathbf{T} \sim \mathcal{S}(t)$, we have

$$\Pr_{\boldsymbol{x}\sim p_{\overline{\mathbf{T}}}}\left[d \leq \mathrm{outdeg}(\boldsymbol{x}, G^{[u]}(\mathbf{T})) \leq 2d\right] \geq \xi$$

  and ξ satisfies

$$\sqrt{d} \cdot \xi \gtrsim \frac{\alpha}{\zeta \log^2 n \log(1/\alpha)}. \tag{3.24}$$

- **Case 2**: There exists a parameter $\kappa \in [O(\log(n/\alpha))]$ (using $\zeta \leq 1$) such that with probability at least $\zeta/(2 \cdot O(\log(n/\alpha)))$ over the draw of $\mathbf{T} \sim \mathcal{S}(t)$, the directed graph $G^{[\kappa]}(\mathbf{T})$ of even edges at scale κ of $p_{\overline{\mathbf{T}}}$ over $\{-1,1\}^{\overline{\mathbf{T}}}$ satisfies

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim p_{\overline{\mathbf{T}}}}\left[\sqrt{\mathrm{outdeg}(\boldsymbol{x}, G^{[\kappa]}(\mathbf{T}))}\right] \gtrsim \frac{\alpha \cdot 2^\kappa}{\zeta \log n \log(n/\alpha) \log(1/\alpha)}.$$

  By a bucketing argument again, there exist parameters $d \in [n]$ and $\xi > 0$ such that with probability at least $\zeta/(2 \cdot \log n \cdot O(\log(n/\alpha)))$ over the draw of $\mathbf{T} \sim \mathcal{S}(t)$, we have

$$\Pr_{\boldsymbol{x}\sim p_{\overline{\mathbf{T}}}}\left[d \leq \mathrm{outdeg}(\boldsymbol{x}, G^{[u]}(\mathbf{T})) \leq 2d\right] \geq \xi$$

  and ξ satisfies

$$\sqrt{d} \cdot \xi \gtrsim \frac{\alpha \cdot 2^\kappa}{\zeta \log^2 n \log(n/\alpha) \log(1/\alpha)}. \tag{3.25}$$

3.5 From Directed Graphs to Mean Vectors In this section, we will show the crucial connection between analyzing the family of graphs defined in Section 3.3 and 3.4 and mean vectors of restrictions of the distribution. Consider a fixed distribution p supported on $\{-1,1\}^n$ and let $t \in [n-1]$. We consider the family of directed graphs $G^{[u]}(T)$ and $G^{[\kappa]}(T)$, $\kappa \geq 1$, for each t-subset T of $[n]$.

It will be convenient to represent directed edges of these directed graphs as $(y, i) \in \{-1,1\}^{\overline{T}} \times \overline{T}$: we say (y, i) is in a graph if it is the case for $(y, y^{(i)})$. Let $\pi = (\pi(1), \ldots, \pi(t+1))$ be an (ordered) sequence of $t+1$ distinct indices from $[n]$. We use $S(\pi)$ to denote the corresponding $(t+1)$-subset $\{\pi(1), \ldots, \pi(t+1)\}$. Given π and $y \in \{-1,1\}^n$, we define a restriction $\rho(\pi, y) \in \{-1, 1, *\}^n$ as

$$\rho(\pi, y)_i = \begin{cases} * & i = \pi(j) \text{ for some } j \in [t+1] \\ y_i & \text{otherwise} \end{cases}.$$

We will also consider sequences $\tau = (\tau(1), \ldots, \tau(t))$ of t distinct indices from $[n]$; its corresponding set $S(\tau)$ and the restriction $\rho(\tau, y)$ given $y \in \{-1,1\}^n$ are defined similarly.

We consider a slightly different but equivalent way of drawing $\boldsymbol{\rho}$ from $\mathcal{D}(t+1, p)$ and $\boldsymbol{\rho}'$ from $\mathcal{D}(t, p)$ which we will use to analyze Case 1 and Case 2 as specified in Section 3.4. We consider sampling $\boldsymbol{\rho} \sim \mathcal{D}(t+1, p)$ according to the following procedure:

1. First, sample a sequence of $t+1$ random indices $\boldsymbol{\pi} = (\boldsymbol{\pi}(1), \ldots, \boldsymbol{\pi}(t+1))$ uniformly from $[n]$ without replacements (so the set $S(\boldsymbol{\pi})$ can be viewed equivalently as drawn from $\mathcal{S}(t+1)$).

2. Then, sample $\boldsymbol{y} \sim p$.
3. Finally, set $\boldsymbol{\rho} = \rho(\boldsymbol{\pi}, \boldsymbol{y})$.

Similarly we consider sampling $\boldsymbol{\rho}' \sim \mathcal{D}(t, p)$ according to the following procedure:

1. First, sample a sequence of t random indices $\boldsymbol{\tau} = (\boldsymbol{\tau}(1), \ldots, \boldsymbol{\tau}(t))$ uniformly from $[n]$ without replacements (so the set $S(\boldsymbol{\tau})$ can be viewed equivalently as drawn from $\mathcal{S}(t)$).
2. Then, sample $\boldsymbol{y} \sim p$.
3. Finally, set $\boldsymbol{\rho}' = \rho(\boldsymbol{\tau}, \boldsymbol{y})$.

A useful observation is that for each $i \in [t+1]$, the distribution of $(\boldsymbol{\pi}_{-i}, \boldsymbol{y})$ is the same as that of $(\boldsymbol{\tau}, \boldsymbol{y})$, where $\boldsymbol{\pi}_{-i}$ denotes the t-sequence obtained from $\boldsymbol{\pi}$ after removing its ith entry. As a result, $\rho(\boldsymbol{\pi}_{-i}, \boldsymbol{y})$ is distributed according to $\mathcal{D}(t, p)$.

Next we state the lemma that gives the connection between graphs and mean vectors.

Lemma 3.6. *Let π be a $(t+1)$-sequence of distinct elements in $[n]$ and $y \in \{-1,1\}^n$. Then for every $i \in [t+1]$, we have*

(3.26)
$$\left|\mu\big(p_{|\rho(\pi,y)}\big)_{\pi(i)}\right| \geq \frac{1}{3} \cdot \mathbf{1}\left\{\big(y_{\overline{S(\pi_{-i})}}, \pi(i)\big) \in G^{[u]}(S(\pi_{-i}))\right\}$$

(3.27)
$$+ \sum_{k\geq 1} 2^{-\kappa-1} \cdot \mathbf{1}\left\{\big(y_{\overline{S(\pi_{-i})}}, \pi(i)\big) \in G^{[\kappa]}(S(\pi_{-i}))\right\}.$$

Proof: First recall that $G^{[u]}(S(\pi_{-i}))$ and $G^{[\kappa]}(S(\pi_{-i}))$ are orientations of disjoint undirected edges. So the right-hand side of (3.27) is non-zero for at most one value.

Let $\ell = p_{|\rho(\pi,y)}$, $T = S(\pi_{-i})$, $z = y_{\overline{T}}$ and $z' = z^{(\pi(i))}$. Writing

$$a = \Pr_{\boldsymbol{x}\sim p}\left[\boldsymbol{x}_{\overline{T}} = z\right] \quad \text{and} \quad a' = \Pr_{\boldsymbol{x}\sim p}\left[\boldsymbol{x}_{\overline{T}} = z'\right],$$

we have $|\mu(\ell)_{\pi(i)}| = |a - a'|/(a + a')$. The weight $w(\{z, z'\})$, defined as $|a - a'|/\max\{a, a'\}$, is at most $2 \cdot |\mu(\ell)_{\pi(i)}|$. If $(z, \pi(i)) \in G^{[u]}(T)$ is uneven, then the weight is at least $2/3$ and thus, $|\mu(\ell)_{\pi(i)}| \geq 1/3$. If $(z, \pi(i)) \in G^{[\kappa]}(T)$ for some κ, then the weight is at least $2^{-\kappa}$ and $|\mu(\ell)_{\pi(i)}| \geq 2^{-\kappa-1}$. ■

4 Mean Testing

For any $m \in \mathbb{N}$ and any distribution p supported on $\{-1,1\}^m$, we consider $\mathbf{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(q)}), \mathbf{Y} = (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(q)})$, a set of $2q$ i.i.d. samples from p, and let

$$\bar{\mathbf{X}} \stackrel{\text{def}}{=} \frac{1}{q}\sum_{i=1}^{q} \boldsymbol{x}^{(i)}, \quad \bar{\mathbf{Y}} \stackrel{\text{def}}{=} \frac{1}{q}\sum_{i=1}^{q} \boldsymbol{y}^{(i)}$$

be the empirical means in $\mathbb{R}^m$. Our core test statistic will take $2q$ i.i.d. samples from p, and compute the expression

$$\mathbf{Z} \stackrel{\text{def}}{=} \langle \bar{\mathbf{X}}, \bar{\mathbf{Y}} \rangle. \tag{4.28}$$

We will write $\mu(p) = \mathbf{E}_{\boldsymbol{x}\sim p}[\boldsymbol{x}] \in [-1,1]^m$ as the mean vector and $\Sigma(p) \in \mathbb{R}^{m\times m}$ as the symmetric matrix with $\Sigma(p)_{ij} = \mathbf{E}_{\boldsymbol{x}\sim p}[\boldsymbol{x}_i\boldsymbol{x}_j]$ for all $i, j \in [m]$.[6]

Lemma 4.1. *The random variable $\mathbf{Z}$ obtained from two tuples of q samples from p satisfies*

(4.29)
$$\mathbf{E}[\mathbf{Z}] = \langle \mathbf{E}[\bar{\mathbf{X}}], \mathbf{E}[\bar{\mathbf{Y}}] \rangle = \langle \mu(p), \mu(p) \rangle = \|\mu(p)\|_2^2,$$

(4.30)
$$\mathbf{Var}[\mathbf{Z}] \leq \frac{1}{q^2}\|\Sigma(p)\|_F^2 + \frac{4}{q}\|\mu(p)\|_2^2\|\Sigma(p)\|_F.$$

We define the *blowup distribution* $\odot(p)$ as the distribution on $\{-1,1\}^{m^2}$ such that, ordering $[m] \times [m]$ in the lexicographic way, $\odot(p)$ is the distribution of the vector

$$(\boldsymbol{x}_i\boldsymbol{x}_j)_{(i,j)\in[m]\times[m]} =$$

(4.31)
$$(\boldsymbol{x}_1\boldsymbol{x}_1, \boldsymbol{x}_1\boldsymbol{x}_2, \ldots, \boldsymbol{x}_1\boldsymbol{x}_m, \boldsymbol{x}_2\boldsymbol{x}_1, \boldsymbol{x}_2\boldsymbol{x}_m, \ldots, \boldsymbol{x}_m\boldsymbol{x}_m)$$

when $\boldsymbol{x} \sim p$. For $k \in \mathbb{N}$, we let $\odot^{(k)}(p)$ denote the distribution over $\{-1,1\}^{m^{2^k}}$ obtained by iterating this process k times, so that in particular $\odot^{(0)}(p) = p$. Note that, for any $k \geq 0$, a sample from $\odot^{(k)}(p)$ can be obtained from a sample from p in time $O(m^{2^k})$.

Fact 4.1. *For any distribution p over $\{-1,1\}^m$ and $k \in \mathbb{N}$, we have $\|\mu(\odot^{(k+1)}(p))\|_2^2 = \|\Sigma(\odot^{(k)}(p))\|_F^2$.*

Consider the following threshold test:

Algorithm 2 ThresholdZTest(τ, S)

Require: A threshold $\tau > 0$ and a (multi)set $S = \{x_1, \ldots, x_q, y_1, \ldots, y_q\}$.
1: Compute Z from the samples in S according to (4.28).
2: **if** $Z > \tau$ **then return** reject
3: **return** accept

[6] In particular, note that due to the diagonal terms we have $\|\Sigma(p)\|_F \geq \sqrt{m}$ for all p.

Hereafter, for $\varepsilon \in (0,1]$, we consider the sequence $(\tau_k)_{k\in\mathbb{Z}^{\geq 0}}$ of real numbers, defined recursively as

$$\tau_k \stackrel{\text{def}}{=} \begin{cases} \frac{\varepsilon^2 n}{2} & \text{if } k = 0 \\ \frac{1}{5000} \cdot q^2 \tau_{k-1}^2 & \text{if } k \geq 1 \end{cases} \tag{4.32}$$

In particular, writing $a \stackrel{\text{def}}{=} 1/5000$, we have

$$\tau_k = \frac{1}{aq^2} \cdot \left(\frac{aq^2\varepsilon^2 n}{2} \right)^{2^k} \tag{4.33}$$

for $k \geq 0$. We are now ready to state the main testing algorithm. The algorithm takes as input a distribution p which is supported on $\{-1,1\}^n$, and is written with two unspecified parameters, k_0 and q. The parameter k_0 denotes the number of rounds and q denotes the sample complexity.

Algorithm 3 MeanTester(p, ε)

Require: A distribution p supported on $\{-1,1\}^n$.
1: Draw a set $\mathbf{S}$ of $2q$ i.i.d. random samples from p.
2: **for all** $0 \leq k \leq k_0$ **do**
3: Set τ_k as in (4.32).
4: Convert the $2q$ samples from $\mathbf{S}$ to a (multi)set $\mathbf{S}^{(k)}$ of samples from $\odot^{(k)}(p)$ as in (4.31).
5: **if** ThresholdZTest$(\tau_k, \mathbf{S}^{(k)})$ returns reject **then return** reject
6: **return** accept ▷ All k_0+1 tests were successful

Theorem 4.1. *Fix any $k_0 \in \mathbb{N}$. There exists an algorithm (Algorithm 3) which, given sample access to an arbitrary distribution p on $\{-1,1\}^n$ and a parameter $\varepsilon \in (0,1]$, has the following behavior:*

- *If p is the uniform distribution, the algorithm outputs accept with probability at least $2/3$;*
- *If p satisfies $\|\mu(p)\|_2 \geq \varepsilon\sqrt{n}$, the algorithm outputs reject with probability at least $2/3$.*

These guarantees hold as long as

$$q \gtrsim \max\left\{ \frac{1}{\varepsilon^2\sqrt{n}}, \left(\frac{1}{\varepsilon^2}\right)^{\frac{2^{k_0+1}}{2^{k_0+2}-2}} \right\}.$$

The algorithm runs in time $O\left(q \cdot n^{2^{k_0}}\right)$.

In particular, by setting $k_0 = \log\log n$, we obtain an algorithm for distinguishing the uniform distribution from a distribution p on $\{-1,1\}^n$ with $\|\mu(p)\|_2 \geq \varepsilon\sqrt{n}$ which runs in time $n^{\Theta(\log n)}$ and has sample complexity

$$O\left(\max\left\{\frac{1}{\varepsilon^2\sqrt{n}}, \frac{1}{\varepsilon}\right\}\right).$$

Remark. As stated the algorithm is not computationally efficient, as for $k_0 = \log\log n$ it runs in super-polynomial time $n^{O(\log\log n)}$. This follows from using the obvious but naive approach to computing the statistic Z in (4.28) for the various blowup distributions $\odot^{(k)}(p)$; however, this can be greatly improved by computing this statistic in a more careful way, rephrasing it as a sum of inner products of tensor products of the original samples and relying on the mixed-product property of tensor products. Doing so results in a running time polynomial in both q and n; we refer the reader to the proof of [CJLW20, Theorem 6] for details.

Proof of Theorem 4.1: The proof will proceed as follows: we first show that, when p is the uniform distribution $\mathcal{U}$ (the completeness case), then all k_0+1 tests, when run on Line 5 of Algorithm 3, return accept with high probability. To do so, notice that Equation 5 of Algorithm 3 considers samples from $\odot^{(k)}(\mathcal{U})$. Hence, we analyze the mean and variance of the statistic for each $\odot^{(k)}(\mathcal{U})$, and apply Chebyshev's inequality to show that, for any given k, each call to Equation 5 then returns accept with probability at least $1 - 2^{-k}/6$. By a union bound over all k, we get that overall all calls will return accept with probability at least $1 - \sum_{k=0}^{\infty} 2^{-k}/6 = 2/3$.

Lemma 4.2. *For the uniform distribution $\mathcal{U}$ over $\{-1,1\}^n$ and $k \in \mathbb{N}$, we have*

$$\|\Sigma(\odot^{(k)}(\mathcal{U}))\|_F^2 \leq (n2^k)^{2^k}.$$

Combining Fact 4.1 and Lemma 4.2, this implies

$$\|\mu(\odot^{(k)}(\mathcal{U}))\|_2^2 \leq (n2^{k-1})^{2^{k-1}} = \sqrt{(n2^k/2)^{2^k}}. \tag{4.34}$$

Lemma 4.3. (Completeness) *There exists a large enough universal constant $C > 0$, such that for any $k \in \mathbb{N}$ where $2^k \leq \log^2 n$. If $q \geq C/\varepsilon^2\sqrt{n}$, then letting $\mathbf{S} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_q, \boldsymbol{y}_1, \dots, \boldsymbol{y}_q\}$ be $2q$ i.i.d. samples from $\odot^{(k)}(\mathcal{U})$. Then,*

$$\Pr_{\mathbf{S}}[\text{ThresholdZTest}(\tau_k, \mathbf{S}) \text{ outputs reject}] \leq \frac{2^{-k}}{6}. \tag{4.35}$$

By a union bound over all k, we thus get that the algorithm, when run on the uniform distribution $\mathcal{U}$, outputs reject with probability at most $\sum_{k=0}^{\infty} \frac{2^{-k}}{6} = 1/3$. For the soundness case, the following lemma will be useful.

Lemma 4.4. *Let p be a distribution supported on $\{-1,1\}^m$, satisfying*

1. $\|\mu(p)\|_2^2 > 2\tau$, and

2. *When* $\mathbf{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_q\}$ *is set to* $2q$ *i.i.d. samples from* p,

$$\Pr_{\mathbf{S}}[\textsc{ThresholdZTest}(\tau, \mathbf{S}) \text{ outputs } \mathsf{accept}] \geq \frac{1}{3}.$$

Then $\|\mu(\odot(p))\|_2^2 \geq \frac{1}{48^2} \cdot \tau^2 q^2$.

Lemma 4.5. (Soundness) *There exists a large enough* $C > 0$ *such that setting*

$$q \geq \left(\frac{C}{\varepsilon^2}\right)^{\frac{2^{k_0+1}}{2^{k_0+2}-2}},$$

the following holds. For any distribution p *on* $\{-1,1\}^n$ *with* $\|\mu(p)\|_2 > \varepsilon\sqrt{n}$, *there is some* $k \in \{0, \ldots, k_0\}$ *such that letting* $\mathbf{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_q\}$ *be* $2q$ *i.i.d. samples from* $\odot^{(k)}(p)$,

$$\Pr_{\mathbf{S}}[\textsc{ThresholdZTest}(\tau_k, \mathbf{S}) \text{ outputs } \mathsf{reject}] \geq \frac{2}{3}.$$

As per the foregoing discussion and Lemma 4.5, there exists a parameter $0 \leq k \leq k_0$ such that $\textsc{ThresholdZTest}(\tau_k, \mathbf{S})$ returns reject with probability at least $2/3$ when $\mathbf{S}$ is drawn from $\odot^{(k)}(p)$. For this setting of k, the algorithm will return reject in Equation 5 with probability at least $2/3$. Finally, from the above analysis, the sample complexity q is set high enough to satisfy the constraints of Lemma 4.3 and Lemma 4.5. ■

4.1 Application to Gaussian Mean Testing

Theorem 4.2. *There exists an algorithm which, given* q *i.i.d. samples from an arbitrary Gaussian distribution* p *on* $\mathbb{R}^n$ *and a distance parameter* $\varepsilon \in (0,1]$, *has the following behavior:*

- *If* p *is the standard Gaussian* $\mathcal{G}(0_n, \mathrm{I}_n)$, *then it outputs* accept *with probability at least* $2/3$;
- *If* p *is some* $\mathcal{G}(\mu, \Sigma)$ *with* $\|\mu\|_2 > \varepsilon$ *(and any* Σ*), then it outputs* reject *with probability at least* $2/3$.

These guarantees hold as long as

$$q \geq C \cdot \frac{\sqrt{n}}{\varepsilon^2},$$

where $C > 0$ *is an absolute constant, and the algorithm runs in time* $\mathrm{poly}(q, n^{\log n})$. *Moreover, any algorithm for this task must have sample complexity* $\Omega\left(n^{1/2}/\varepsilon^2\right)$.

Acknowledgments. The authors would like to thank Rajesh Jayaram, whose suggestions (in particular, Theorem 4, and a tighter union bound for the completeness case of Theorem 4.1) helped improve an earlier version of the paper.

References

[ABDK18] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In *Advances in Neural Information Processing Systems 31*, NeurIPS '18. Curran Associates, Inc., 2018.

[ACK15a] Jayadev Acharya, Clément L. Canonne, and Gautam Kamath. Adaptive estimation in weighted group testing. In *Proceedings of the 2015 IEEE International Symposium on Information Theory*, ISIT '15, pages 2116–2120, Washington, DC, USA, 2015. IEEE Computer Society.

[ACK15b] Jayadev Acharya, Clément L. Canonne, and Gautam Kamath. A chasm between identity and equivalence testing with conditional queries. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques.*, RANDOM '15, pages 449–466, Dagstuhl, Germany, 2015. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[ADJ+11] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *Proceedings of the 24th Annual Conference on Learning Theory*, COLT '11, pages 47–68, 2011.

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 3577–3598. Curran Associates, Inc., 2015.

[BBC+19] Ivona Bezakova, Antonio Blanca, Zongchen Chen, Daniel Štefankovič, and Eric Vigoda. Lower bounds for testing graphical models: Colorings and antiferromagnetic Ising models. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 283–298, 2019.

[BC17] Tuğkan Batu and Clément L. Canonne. Generalized uniformity testing. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pages 880–889, Washington, DC, USA, 2017. IEEE Computer Society.

[BC18] Rishiraj Bhattacharyya and Sourav Chakraborty. Property testing of joint distributions using conditional samples. *Transactions on Computation Theory*, 10(4):16:1–16:20, 2018.

[BCG17] Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. In *Proceedings of the 32nd Computational Complexity Conference*, CCC '17, pages 28:1–28:40, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[BDKR05] Tuğkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

[BFF+01] Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In

Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science, FOCS '01, pages 442–451, Washington, DC, USA, 2001. IEEE Computer Society.

[BFR+00] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '00, pages 259–269, Washington, DC, USA, 2000. IEEE Computer Society.

[BFRV11] Arnab Bhattacharyya, Eldar Fischer, Ronitt Rubinfeld, and Paul Valiant. Testing monotonicity of distributions over general partial orders. In *Proceedings of the 2nd Conference on Innovations in Computer Science*, ICS '11, pages 239–252, Beijing, China, 2011. Tsinghua University Press.

[BKR04] Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the 36th Annual ACM Symposium on the Theory of Computing*, STOC '04, New York, NY, USA, 2004. ACM.

[BV15] Bhaswar Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2611–2619. Curran Associates, Inc., 2015.

[BW18] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727–749, 2018.

[Can15a] Clément L. Canonne. Big data on the rise? - testing monotonicity of distributions. In *Proceedings of the 42nd International Colloquium on Automata, Languages, and Programming*, ICALP '15, pages 294–305, 2015.

[Can15b] Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22(63), 2015.

[Can16] Clément L. Canonne. Are few bins enough: Testing histogram distributions. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '16, pages 455–463, New York, NY, USA, 2016. ACM.

[CDKS17] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing Bayesian networks. In *Proceedings of the 30th Annual Conference on Learning Theory*, COLT '17, pages 370–448, 2017.

[CDVV14] Siu On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1193–1203, Philadelphia, PA, USA, 2014. SIAM.

[CFGM13] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 561–580, New York, NY, USA, 2013. ACM.

[CFGM16] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. *SIAM Journal on Computing*, 45(4):1261–1296, 2016.

[CJLW20] Xi Chen, Rajesh Jayaram, Amit Levi, and Erik Waingarten. Learning and testing junta distributions with subcube conditioning. *arXiv preprint arXiv:2004.12496*, 2020.

[CR14] Clément L. Canonne and Ronitt Rubinfeld. Testing probability distributions underlying aggregated data. In *Proceedings of the 41st International Colloquium on Automata, Languages, and Programming*, ICALP '14, pages 283–295, 2014.

[CRS14] Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1174–1192, Philadelphia, PA, USA, 2014. SIAM.

[CRS15] Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 44(3):540–616, 2015.

[DDK18] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising models. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 1989–2007, Philadelphia, PA, USA, 2018. SIAM.

[DDS+13] Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, pages 1833–1852, Philadelphia, PA, USA, 2013. SIAM.

[DGPP18] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *Proceedings of the 45th International Colloquium on Automata, Languages, and Programming*, ICALP '18, pages 41:1–41:14, 2018.

[DK16] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 685–694, Washington, DC, USA, 2016. IEEE Computer Society.

[DKN15] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1841–1854, Philadelphia, PA, USA, 2015. SIAM.

[DKP19] Ilias Diakonikolas, Daniel M. Kane, and John Peebles. Testing identity of multidimensional histograms. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 1107–1131, 2019.

[DKW18] Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? In *Proceedings of the 29th Annual*

ACM-SIAM Symposium on Discrete Algorithms, SODA '18, pages 2747–2764, Philadelphia, PA, USA, 2018. SIAM.

[DP17] Constantinos Daskalakis and Qinxuan Pan. Square Hellinger subadditivity for Bayesian networks and its applications to identity testing. In *Proceedings of the 30th Annual Conference on Learning Theory*, COLT '17, pages 697–703, 2017.

[FJO+15] Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Faster algorithms for testing under conditional sampling. In *Proceedings of the 28th Annual Conference on Learning Theory*, COLT '15, pages 607–636, 2015.

[FLV17] Eldar Fischer, Oded Lachish, and Yadu Vasudev. Improving and extending the testing of distributions for shape-restricted properties. In *Proceedings of the 34th Symposium on Theoretical Aspects of Computer Science*, STACS '17, pages 31:1–31:14, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[GGR96] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. In *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '96, pages 339–348, Washington, DC, USA, 1996. IEEE Computer Society.

[GLP18] Reza Gheissari, Eyal Lubetzky, and Yuval Peres. Concentration inequalities for polynomials of contracting Ising models. *Electronic Communications in Probability*, 23(76):1–12, 2018.

[GMV06] Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '06, pages 733–742, Philadelphia, PA, USA, 2006. SIAM.

[Gol17] Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017.

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000.

[GTZ17] Themistoklis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Faster sublinear algorithms using conditional sampling. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1743–1757, Philadelphia, PA, USA, 2017. SIAM.

[GTZ18] Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Certified computation from unreliable datasets. In *Proceedings of the 31st Annual Conference on Learning Theory*, COLT '18, pages 3271–3294, 2018.

[ILR12] Piotr Indyk, Reut Levi, and Ronitt Rubinfeld. Approximating and testing k-histogram distributions in sub-linear time. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '12, pages 15–22, New York, NY, USA, 2012. ACM.

[Kam18] Gautam Kamath. *Modern Challenges in Distribution Testing*. PhD thesis, Massachusetts Institute of Technology, September 2018.

[KMS18] Subhash Khot, Dor Minzer, and Muli Safra. On monotonicity testing and Boolean isoperimetric-type theorems. *SIAM Journal on Computing*, 47(6):2238–2276, 2018.

[KT19] Gautam Kamath and Christos Tzamos. Anaconda: A non-adaptive conditional sampling algorithm for distribution testing. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, Philadelphia, PA, USA, 2019. SIAM.

[NS02] Assaf Naor and Gideon Schechtman. Remarks on non-linear type and Pisier's inequality. *Journal fur die Reine und Angewandte Mathematik*, 552:213–236, 2002.

[OS18] Krzysztof Onak and Xiaorui Sun. Probability-revealing samples. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, AISTATS '18, pages 84:2018–2026. JMLR, Inc., 2018.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[Pis86] Gilles Pisier. Probabilistic methods in the geometry of Banach spaces. In Giorgio Letta and Maurizio Pratelli, editors, *Probability and Analysis*, pages 167–241. Springer, 1986.

[RS09] Ronitt Rubinfeld and Rocco A. Servedio. Testing monotone high-dimensional distributions. *Random Structures and Algorithms*, 34(1):24–44, 2009.

[Rub12] Ronitt Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

[SSJ17] Imdad S. B. Sardharwalla, Sergii Strelchuk, and Richard Jozsa. Quantum conditional query complexity. *Quantum Information & Computation*, 17(7& 8):541–566, 2017.

[Tal93] Michel Talagrand. Isoperimetry, logarithmic Sobolev inequalities on the discrete cube, and Margulis' graph connectivity theorem. *Geometric & Functional Analysis*, 3(3):295–314, 1993.

[Val11] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 51–60, Washington, DC, USA, 2014. IEEE Computer Society.

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[Wag15] Bo Waggoner. l_p testing and learning of discrete distributions. In *Proceedings of the 6th Conference on Innovations in Theoretical Computer Science*, ITCS '15, pages 347–356, New York, NY, USA, 2015. ACM.